Statistical Inference for Sequential Designs of Randomized Clinical Trials with Binary Responses

by

© Apsara Pathum Jayasooriya

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science.

Department of Mathematics and Statistics

Memorial University of Newfoundland

April 2023

St. John's, Newfoundland and Labrador, Canada

Abstract

Sequential designs of Randomized Clinical Trials (RCT) allow repeated significance testing based on cumulative data over time. The sequential testing method enables early termination of the study using a pre-defined stopping rule when preliminary results show a clear superiority of one treatment over the other. Over the decades, researchers have presented several techniques for determining the stopping rule, mainly for continuous data. However, clinical trial data are not necessarily continuous. In certain cases, data can be dichotomous, containing only two distinct values. Some researchers have proposed special sequential testing procedures to analyze binary data considering individual data points at each stage. With the influence of those approaches, we are more focused on a method which can be used to analyse groups of binary data.

The thesis considers the implementation of three main approaches, namely, Pocock [32, 34], O'Brien and Fleming [29] and Haybittle-Peto [31, 15] methods for computing the critical values required for controlling the size and power of tests at various stages of sequential analysis. Critical values are obtained using an iterative Markov chain approach to satisfy the alpha spending at each stage. Considering the discrete nature of the data, a likelihood ratio test statistic is used for testing the proportions. Examples of two-stage and three-stage analysis were used to illustrate the computation of the critical values, size and power of tests of proportions, and then the outcomes based on Pocock, O'Brien & Fleming and Haybittle-Peto methods are compared.

This work is dedicated to my friends in St. John's
who have become my family away from home.

Lay summary

Clinical research includes searching for possible medical treatments, comparing the benefits of different treatments, and finding out which combinations of treatments work best. Sequential designs of Randomized Clinical Trials (RCT) are a useful strategy to find a way to stop the study early if the initial results show that one treatment is clearly better than the other; so that researchers can save time as well as the costs of the study. In this method, data is evaluated as collected, and sampling is discontinued using a predefined stopping rule as soon as the study finds it statistically significant.

The existing methods that developed various stopping rules are based on the assumption that the data are continuous. However, clinical data may be categorical, where the counts within each category need to be considered. Sometimes, data might be binary, such as medical test results (negative or positive), survival status (alive or dead), and family medical history (yes or no). Thus, only a particular sequential testing procedure can be compatible with binary data. Therefore, this study proposes an appropriate test statistic for testing proportions considering the discrete nature of binary data. The study also proposes techniques to calculate the decision boundary, the corresponding error rate and power to determine whether the test leads to the correct decision in a very large proportion of the time. Various approaches are utilized to define the decision rule, and these approaches are compared. Furthermore, generalized equations are provided, allowing the suggested technique to be used for any number of stages.

Acknowledgements

I would like to express my gratitude to my supervisors, Dr. Asokan Variyath and Dr. Yanqing Yi, for their patience, enthusiasm, motivation and guidance. I gratefully acknowledge the financial support from Dr. Variyath's research grant, Dr. Yi's *Discovery Grant* from the *Natural Sciences and Engineering Research Council (NSERC)* of Canada and the graduate fellowship from the school of Graduate studies as well as the teaching assistantship provided by the Department of Mathematics and Statistics of Memorial University of Newfoundland.

Table of contents

List of tables

List of figures

Chapter 1

Introduction

1.1 Randomized clinical trials

Clinical research involves examining potential medical treatments, comparing the benefits of competing treatments, and determining the best treatment combinations. Randomized clinical trials (RCT) are often regarded as the most reliable clinical research strategy (Hardon et al. 1996 [13], Altman et al. 2001 [3], Sverdlov et al. 2013 [39]). In RCTs, the participants are randomly assigned to one or more treatments to measure and compare the effect and value of the treatment against a control. The experimental approach of 'randomization' has widespread application in biological experiments. Randomization is the foundation of clinical trial designs and is utilised to ensure that statistical inference at the end of the study is legitimate (Friedman et al. 2004 [41]).

The outcomes of clinical trials will be biased if the treatment groups are systematically different. This systematic bias can be eliminated by generating random groups that differ in both known and unknown prognostic factors. Randomization can also assist in decreasing certain experimental biases such as ascertainment bias, selection bias, and accidental bias (Sverdlov et al. 2013 [39]). Subjects or participants in clinical trials as well as investigators and others should have no prior knowledge of the treatments that have been allocated. The knowledge of treatment assignment can introduce ascertainment bias to the experiment. For example, if a patient knows they are assigned to a placebo, they are more likely to drop out of the trial or not comply with the treatment plan; and researchers or analysts may be biased in their evaluations of patient outcomes based on their expectations of treatment effects. Proper randomization can eliminate this ascertainment bias. According to Schul et al. (2002) [36], the studies that do not employ appropriate or clear randomization tend to overstate treatment effects by up to 40% when compared to studies that use adequate randomization.

Furthermore, a balance of important known and unknown covariates across treatments is required for meaningful treatment comparison. In the analysis of clinical trials, it is common practice to make corrections for covariate imbalance by employing statistical methods such as analysis of covariance (ANCOVA). However, interpreting this post-adjustment technique can be problematic (Frane 1998 [11]) because covariate imbalance commonly leads to unforeseen interaction effects, such as uneven slopes among subgroups of variables. Therefore, the best strategy to balance covariates among groups is to use proper randomization during the design stage of a clinical study rather than after data collection. In clinical trials, there are various ways to assign individuals to treatment groups randomly. The most often used randomization methods are simple randomization, block randomization, stratified randomization, and adaptive randomization.

Simple randomization relies solely on a single set of random assignments. In the case of two treatments, the simplest technique of random allocation provides each patient with an equal chance of receiving either treatment, which is performed by flipping a coin (heads - control, tails - treatment) or rolling a dice (even-control, odd-treatment). However, computer-based random number generators and random number tables in statistical books are the most common methods used in practice. Simple randomization is straightforward and easy to implement. However, it may result in an uneven number of participants between groups, particularly in clinical trials with small sample sizes (Altman et al. 1999 [2]). Therefore, block randomization is used as an alternative.

Block randomization involves small and well-balanced blocks with specified group assignments that are used to ensure the same number of participants are included in each group (Frane 1998 [11], Altman et al. 1999 [2]). The block size is pre-determined and is a multiple of the number of treatment groups. Following the determination of block size, all balanced possibilities of assignments inside the block are computed.

Randomization is then carried out by assigning random permutations of treatments within each block. Despite the fact that block randomization allows for balanced sample sizes, groups that are rarely comparable in terms of certain covariates may be formed. These unbalanced covariates can significantly impact the interpretation of the results if they are not controlled (Pocock et al. 1975 [34]). Furthermore, the statistical analysis might be biased by the imbalance, which would lower the statistical power of the test. Therefore, stratified randomization is used to adjust and balance the influence of covariates.

In stratified randomization, a separate block is formed for each combination of covariates that impact the dependent variable, and subjects can be allocated to the appropriate block. The block size is recommended to be relatively small to keep the equilibrium in smaller strata. Increasing the number of stratification factors or the number of levels within strata results in a fewer number of patients per stratum. In stratified randomization, the baseline measurements are taken before randomization. However, in clinical trials, patients are often enrolled one at a time on a continuous basis, making this strategy ineffective. Therefore, stratified randomization is problematic if all participants' baseline attributes are unknown (Lachin et al. 1988 [25]) and an alternative method is necessary.

Covariate adaptive randomization is a reliable replacement for conventional randomization approaches in clinical research (Zalene 1990) [51]. Covariate adaptive randomization assigns participants to treatments by examining the allocation of comparable patients who have previously been randomised and then allocating them to achieve the best possible balance among the treatment groups with regard to all stratification variables (Kalish et al. 1985) [20].

The best technique to ensure that the findings of a clinical trial are not influenced by how participants are assigned to each treatment is to use the appropriate method of randomization. The objective of RCT is to obtain an accurate comparison of the effects of an experimental treatment in the target patient population. A common approach for comparing treatments is to fix the sample size/duration in advance and, upon completion of the study, conduct the analysis using a formal test of significance based on a calculated test statistic. However, in clinical trials with sequential patient entrance, fixed sample size or fixed duration designs are unethical (Pocock 1977,1982) [32, 33]. Even if the trial was initially established as a fixed sample size/duration design, the study must be terminated for ethical grounds when intermediate data assessment suggests that continuation is unsuitable. One real-world example is the Beta-Blocker Heart Attack Trial (BHAT), reported in 1981. The duration of the trial, which was originally intended to be fixed, was shortened after it was shown that the group of patients who were given propranolol had a much lower mortality rate from all causes than the control group (Lai 1984) [26]. Moreover, medical trials are concerned about the possibility of early termination of the study if preliminary results show a clear superiority of one treatment over the other. One solution to this concern is group sequential designs. This makes it possible to decide whether or not to terminate the trial based on the results of repeated significance tests performed on the accumulated data after each treatment has been evaluated.

1.2 Group sequential designs of randomized clinical trials

Group sequential designs of RCTs perform multiple tests based on cumulative data over time. This approach is well known as sequential analysis or sequential hypothesis testing. The key feature of this approach is that the sample size is not specified in advance. Instead, data are assessed as obtained, and sampling is terminated with a pre-defined stopping rule as soon as statistically significant findings are revealed, eliminating the need for additional sampling.

The core concept of group sequential designs is repeated significance testing, with one test after each set of collected observations. Therefore, the plan of group sequential sampling systems generally includes collecting data as groups of observations and determining the maximum number of stages. Given the maximum number of inspections, K, a group sequential design will have a $(K-1)$ number of 'interim analyses/stages' and one final stage (Armitage 1991) [5]. Interim analysis refers to an assessment of the present data from a trial in progress, which addresses the core research issue, and which has the potential to change the way the study is conducted (Whitehead et al. 2001) [47]. The idea of group sequential testing is illustrated in Figure 1.1 comparing a non-sequential design with a group sequential design with two stages and three stages (Weigl et al. 2020) [44] (Retrieved from https://meth.psychopen.eu/index.php/meth/article/view/2811/2811.pdf).

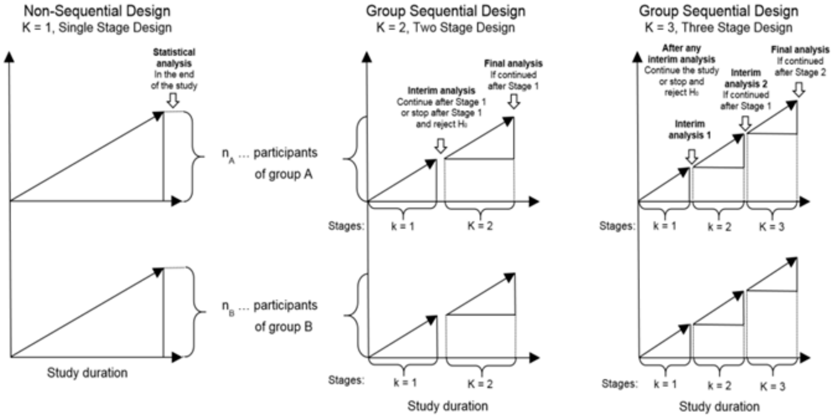

Figure 1.1: Non-sequential design and group sequential designs with two and three stages

The conventional fixed sample design trial has a predetermined sample size. Regardless of whether the true treatment impact is considerably positive, marginal, or truly adverse compared to the control, this design employs the same (fixed) number of participants. The statistical analysis is performed at the end of the study (See the left panel of Figure 1.1). In contrast, the number of participants is not specified in advance in the group sequential design. The interim analyses in sequential designs provide early termination when preliminary results reveal efficacy or ineffectiveness. The middle set of graphs in Figure 1.1 shows a two-stage group sequential design, which consists of one interim analysis and a final analysis. Similarly, the right panel of Figure 1.1 depicts a three-staged group sequential design with two interim analyses and one final analysis. A decision must be made at each interim analysis to continue the study or to stop and reject the null hypothesis. These graphs for group sequential designs clearly demonstrate that sample sizes are growing with time. When compared to typical fixed-sample designs, group sequential approaches allow more flexibility by incorporating interim analyses.

Nevertheless, this interim analysis may have hidden consequences. If the results of the analysis indicate that the trial can be terminated early, then the trial will be stopped, and there won't be any further analysis. However, if the analysis does not reveal that the trial can be stopped early, the fact that the analysis was carried out could potentially undermine the power of the test and increase the type I error rate. That is, every time we have interim looks at the data considering the termination of the trial, we increase or compound the probability of falsely rejecting the null hypothesis (Kumar et al. 2016) [24].

Consider a statistical test with a significance level of 0.05. For a fixed sample design, the data is analysed at the end of the trial; therefore there's no interim analysis. Then there's a 5% chance of incorrectly rejecting the null hypothesis (See Figure 1.2: Retrieved from https://www.quantics.co.uk/blog/hidden-consequences-of-interim-analyses-and-adaptive-trial-options/).

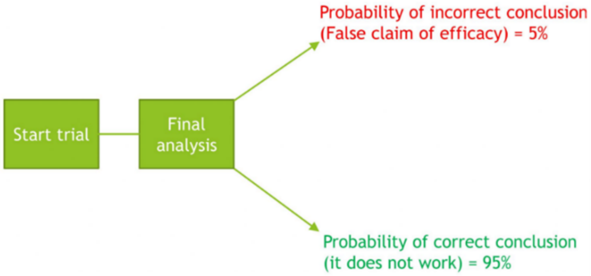
Probability of incorrect conclusion
(False claim of efficacy) = 5%
Start trial
Final analysis
Probability of correct conclusion
(it does not work) = 95%

Figure 1.2: Probability of correct and incorrect decisions for a fixed sample size design.

In contrast, if we examine the data multiple times with several interim looks and consider the significant level to be 0.05, we have a 5% chance of stopping each time which increase the probability of incorrectly rejecting the null hypothesis at every interim analysis (Figure 1.3 : retrieved from https://www.quantics.co.uk/blog/hidden-consequences-of-interim-analyses-and-adaptive-trial-options/).

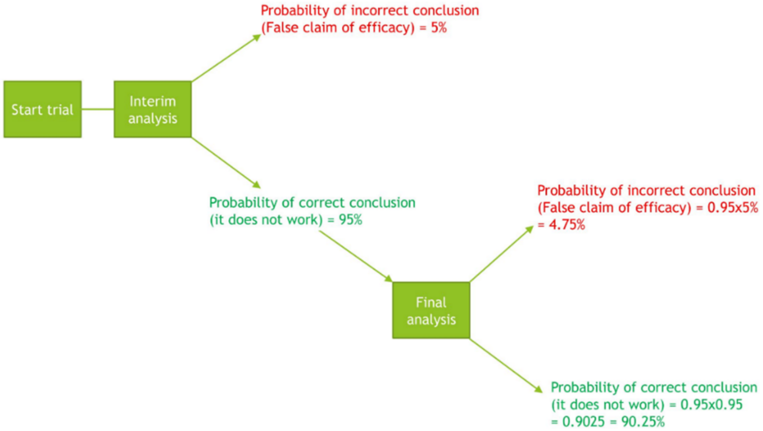
Probability of incorrect conclusion
(False claim of efficacy) = 5%
Start trial
Interim analysis
Probability of correct conclusion
(it does not work) = 95%
Probability of incorrect conclusion
(False claim of efficacy) = 0.95x5%
= 4.75%
Final analysis
Probability of correct conclusion
(it does not work) = 0.95x0.95
= 0.9025 = 90.25%

Figure 1.3: Probabilities of correct and incorrect decisions for a design with one interim analysis.

The overall type I error rate of the fixed sample size design in Figure 1.2 is 5%, and the overall type I error rate of the design with an interim analysis (Figure 1.3) is 5% + 4.75% = 9.75%. Moreover, the probability of correct decision for the fixed sample design is 95%, and it has also decreased to 90.25% in the design with an interim analysis.

However, sequential design has resolved the issue of accommodating the interim analysis. The overall study must have a maximum of 5% likelihood of rejecting null-hypothesis incorrectly, however this 5% can be split between the intermediate analysis (or analyses) and the final analysis. There are different ways to split the type I error rate between the stages and researchers have proposed different functions. This approach is known as 'Alpha spending' and it allows for sequential testing while while maintaining the overall type I error rate. More details on alpha spending functions are explained in Section 1.2.3.

When it comes to the benefits, a group sequential design (two or three stages) with a slight increase in sample size relative to the fixed sample test can lower the expected sample size by approximately 30% (Harington 2001, [14]), while maintaining the statistical power and controlling the type I error rate of the study. Therefore, the main advantage of the group sequential designs over traditional fixed-sample testing is that it enables early conclusions while maintaining accurate decisions. This quicker decision-making is beneficial in terms of conserving cost, resources, and time.

The statistical concept underlying sequential testing is based on a specific numerical recursive integration formula developed by Armitage et al.(1969) [6] and McPherson et al.(1971) [28], which addresses the independent increment structure of the underlying data accumulation process. Pocock (1977, 1982) [32, 33] and O'Brien & Fleming (1979) [29] provided the primary motivation for the creation of group sequential test techniques, which are currently frequently utilised in clinical research. Their application in quality control may be traced all the way back to the research conducted by Dodge & Romig (1929) [9] and Shewhart (1931) [37]. Jennison & Turnbull (2000) [19] and Todd (2007) [40] provided historical reviews of the early work on group sequential designs.

The choice of the probability model for each specific problem is a key point when designing these sequential clinical trials. Most of the models use Normal probability distribution (Pocock 1977) [32] while some studies have used Bernoulli (Kulldrof et al.,2011) [23], Binomial (Hoel et al.,1976) [17], and Poisson data (Abt,1998) [1]. Prior to the start of the trial, the “stopping rule” is specified and stated. The stopping rule specifies when and why the trial should be stopped. For example, a favourable trial outcomes indicate that the study should be continued. Negative trial outcomes (any unfavourable consequences) indicate that the experiment should be stopped.

Sequential testing often involves comparing the value of a test statistic with a critical value at each stage. This technique is used by Wald (1945) [42] in his classical sequential probability ratio test (SPRT), and other researchers in their frequentist methods. Unfortunately, when applied to clinical studies, Wald’s test had two major flaws. First, the test is for two simple hypotheses. In clinical studies, the null hypothesis of no treatment difference is typically accompanied with a two-sided alternative hypothesis of treatment difference. Second, there is no upper limit set on the total number of participants in the trial which makes the SPRT an open design. ‘The boundaries approach’ is the result of a series of improvements that were made to fix either one of these issues or both. The ‘triangular test’ is an adjustment, first suggested by Anderson (1960) [4], then examined further by Lai (1973) [27], and finally detailed by Whitehead et al. (1983) [46]. In ‘Triangular plans’, the stopping zone is defined by two straight lines that intersect at the conclusion of the trial. This can be recognized as the most common design for the boundaries approach. Whitehead (1997) [45] provided a detailed description on the triangular test and other methods based on straight line stopping boundaries.

For a test that is continually monitored, the total type I error rate can be maintained using the critical values calculated using the boundaries technique. However, the type I error rate calculated is lower than the intended level, α (Todd 2007, [40]). In order to accommodate the discretely monitored sample route, Whitehead (1997) [45] proposed a modification to the continuous boundaries. The modification, known as the ’Christmas tree correction’, pulls critical values in by an amount proportional to the predicted overshoot of the discrete sample path.

Another common approach to construct the decision rule is to use an alpha spending function. The idea behind this approach is to control the type I error rate at each stage of the multiple tests using a non-decreasing function. (See Section 1.2.3) This technique provides more flexibility in the form of the stopping boundaries. The approach allows for the development of a test that maintains the type I error rate when inspections depart from their planned route (Todd 2007) [40]. Alpha spending functions proposed by Lan and DeMets (1983) [12] lead to designs that are comparable to those proposed by Pocock, and O'Brien & Fleming. Kim and DeMets (1987) [22] proposed a family of alpha spending functions, that are compatible with the designs created using the boundaries technique. However, the above-mentioned frequentist methods, such as Pocock's method and O'Brien & Fleming's method, have intrinsic alpha spending functions behind them (Silva et al. 2020) [38].

1.2.1 Basic concepts

In group sequential designs, a statistical significance test generates decision regions for rejecting and not rejecting the null hypothesis H_0. A test statistic is calculated after collecting observations and is compared with a critical value. The boundary of the rejection region is defined by the critical value and H_0 is rejected if the test statistic falls in the rejection region. The probability of rejecting H_0 when it is true (Type I error rate) is bounded by the significance level(α) of the test. In order to provide a clear idea of the methodology used in this study, all of the fundamental concepts on group sequential designs, as well as statistical terms, are explained below.

a) Test statistic

A hypothesis test is often stated in terms of a test statistic $T(X_1, X_2, ..., X_n) = T(X)$, which is a function of the sample [7]. For instance, a test might state that the null hypothesis (H_0) must be rejected if the sample mean ($\bar{x}$) is larger than 5. Here, $T(X) = \bar{x}$ is the test statistic and therefore the rejection region can be defined as, $\{(x_1, x_2, ..., x_n) : \bar{x} > 5\}$. Test statistic provides you with information on how likely it is that the results obtained from the sample data are under the assumption that the null hypothesis is true. The null hypothesis is rejected in favour of an alternate hypothesis as the results become less likely under this assumption. The more likely your results are, the more difficult it is to reject the null hypothesis.

In the group sequential set-up, testing is done after groups of observations have been collected or measured. Let the maximum number of stages be K, and then the sizes of the samples introduced to the existing data at each stage are given by

$n_1, \ldots, n_K$; making the maximum sample size of the test procedure $N = \sum_{k=1}^{K} n_k$. At stage k, an appropriate test statistic(T_k^*) for testing the null hypothesis (H_0) is determined. The test statistic summarizes the information up to stage k. Then, H_0 is rejected at stage k if the test statistic T_k^* falls in the rejection region (R^*) of k^{th} stage;

$$T_k^* \in \mathrm{R}_k^*$$

This indicates that the test statistic of stage one to $(k-1)^{th}$ stage has fallen in the continuation region. $T_i^* \in \mathrm{C}_i^*$ for $i = 1, 2, .., (k-1)$

b) Critical values

A critical value is a point on the distribution of the test statistic under the null hypothesis that defines a boundary between non-significant and significant results in a hypothesis test. Armitage et al.(1969)[6] first proposed altering the critical values in order to control the type I error rate at a specific level using an inverse interpolation method. Critical values create the boundaries which separate the acceptance, continuation and rejection regions. However, in this study, the acceptance region is not considered; therefore, the critical values calculated in this study define the boundary between the rejection and the continuation regions.

Let cv_k be the critical value for the k^{th} stage. Then, the stopping rule is (considering two sided alternative hypothesis), reject the null hypothesis (H_0) if $|T_k| \geq \mathrm{cv}_k$ or else continue the test to next stage; where T_k is the test statistic for the k^{th} stage.

c) Log likelihood ratio test

Sequential testing is traditionally done by comparing a test statistic against a critical value. Wald (1945) [42], Pocock (1977) [32], O'Brien & Fleming(1979)[29] employed this approach in their methods for testing group sequential data. Wald (1945) [42] introduced the first classical likelihood method called the 'sequential probability ratio test' (SPRT) for continuous testing. This is a very general method and can be used with many different probability distributions. The critical value of this SPRT is given in the scale of likelihood ratio statistic.

A likelihood function $L(\theta|X)$ is created using a probability distribution that relates to the outcome measurements of the study. To decide whether θ_0 or θ_1 is the more acceptable value of θ, the likelihood ratio (LR) can be used to examine the evidence. The likelihood ratio for comparing θ_0 and θ_1 can be denoted as,

$$LR = \frac{L(\theta_0|X)}{L(\theta_1|X)}. \tag{1.1}$$

According to Wald's SPRT, reject H_0 if $LR \geq \frac{1-\beta}{\alpha}$ and do not reject H_0 if $LR \leq \frac{\beta}{1-\alpha}$ where α is the type I error rate and β is the type II error rate of the test. However, the outcome of Wald's classical SPRT is highly dependent on the relative risk used in specifying the alternative hypothesis. To overcome this, Kudroff et al. (2011)[23] proposed a modification to Wald's SPRT method, which is called the 'maximized sequential probability ratio test' (MaxSPRT). The key feature of MaxSPRT is that it allows for a composite, one-sided alternative hypothesis, and it introduces an upper stopping boundary. This study follows the idea of MaxSPRT, and the log-likelihood ratio has been chosen as the test statistic.

d) Type I error rate

In hypothesis testing, type I error rate (α) is the probability of rejecting a null hypothesis when it is true; that means,

$$\alpha = Pr(Reject\ H_0|H_0\ is\ true). \tag{1.2}$$

In group sequential testing, the overall type I error rate is computed by adding the error rates of each stage. However, controlling the study-wide overall type I error rate is a major concern in group sequential analysis as it should not exceed the overall significance level of the test. It is necessary to make statistical adjustments to control the total type I error rate. Type I error rate control approaches introduced by Pocock (1977) [32], O'Brien & Fleming (1979) [29] and Haybittle (1971) [15] & Peto (1976) [31] are examples of such adjustments. Moreover, alpha spending functions(see Subsection 1.2.3), provide a more flexible way to control the overall type I error rate.

e) Power

It is important to conduct high-quality hypothesis tests in order to be confident in results. An essential aspect of determining the overall quality of a hypothesis test is to ensure it is "powerful".

The power of a hypothesis test is the probability of rejecting the null hypothesis H_0 when it is false. That is the probability of making the correct decision when the alternative hypothesis H_a is true, and it is denoted as,

$$(1 - \beta) = Pr(Reject\ H_0 | H_a\ is\ true). \tag{1.3}$$

Here, β is the probability of failing to reject the null hypothesis when it is false. In group sequential testing, the overall power is computed by adding the partial probabilities of each stage.

f) The recursive integration

Recursive integration is widely applied to evaluate high-dimensional integral expressions. This is used in many areas of statistical inference where high-dimensional integral evaluations are required for probability calculations and critical point assessments (Hayter 2006) [16]. Recursive integration enables the evaluation of an integral expression in one dimension by a series of calculations in a smaller dimension. As a result, the computation time is much decreased. Recursive approaches for sequential analysis were demonstrated by Armitage et al.(1969) [6] and Jennison & Turnbull (1991) [18]. The recursive integration approach is commonly used to calculate the critical values of Pocock's method and O'Brien & Fleming's method.

1.2.2 Frequentist methods

Assume a group sequential plan with K stages. Let $T_1, T_2, \ldots, T_K$ be the test statistics and $cv_1, cv_2, \ldots, cv_K$ be the critical values for each stage. Considering a two sided alternative hypothesis, the sequential test is terminated at the k^{th} interim analysis due to the rejection of the null hypothesis, if

$$|T_k| \geq cv_k \qquad \text{for } k = 1, 2, \ldots, K$$

The critical values are selected in such a way that the total significance level does not exceed the intended α level. Different approaches have been developed for determining these stopping boundaries. This study focuses mainly on three main designs.

a) Pocock method

Pocock (1977) [32] proposed a stopping boundary for testing two-sided alternative hypotheses in group sequential analysis. This approach has same critical value (cv^*) at each interim analysis and at the final analysis ensuring,

$$P_{H_0}(|T_1^*| \geq cv^* \quad or \quad ... \quad or \quad |T_K^*| \geq cv^*) = \alpha, \tag{1.4}$$

where the critical values(cv^*) depend on K and α, and can be denoted by $c_P(K, \alpha)$. The constant $c_P(K, \alpha)$ is calculated numerically using the recursive integration. Table 1.1 provides the critical values $cv^* = c_P(K, \alpha)$ together with the expected alpha spending that were computed for $\alpha = 0.05$ and $K = 2, 3, 4$, and 5 respectively.

Though the exact boundaries were obtained for a trial with two treatments and a response from a Normally distributed population with known variance, Pocock justified that group sequential methods could also be adapted to many other types of response variables, such as responses from a Normally distributed population with unknown variance, Binomial, or Exponential. The Pocock boundary is simple, as the critical values are the same for each interim analysis. This approach is more likely to result in the trial being terminated early, resulting in a smaller expected sample size. Early termination, however, depends on several other factors, such as the sample size and the power. In Pocock's method, the number of interim analyses must be fixed at the beginning, and this creates a disadvantage as the method is not able to incorporate additional analysis after a trial has already begun.

b) O'Brien & Fleming method

O'Brien and Fleming (1979) [29] proposed monotonically decreasing critical values fulfilling the condition,

$$P_{H_0}(|T_1^*| \geq cv_1^* \quad or \ \ ... \ \ or \quad |T_K^*| \geq cv_K^*) = \alpha, \tag{1.5}$$

where $cv_1^* > cv_2^* > \ldots > cv_K^*$.

The critical values defined by cv_k^* depend on α, K and the current stage k. This can be expressed as $cv_k^* = c_{OBF}(K, \alpha)/\sqrt{k}$ where $c_{OBF}(K, \alpha)$ is a constant and it is computed numerically by recursive integration. Critical values (cv_k^*) for $\alpha = 0.05$ and $K = 2, 3, 4,$ and 5 are provided in Table 1.1.

These boundary values are inversely proportional to the square root of information levels on the standardized Z scale. The boundary is conservative, with large critical values in the early stages; therefore, the O'Brien-Fleming (OBF) design is less likely to reject the null hypothesis in the early stages compared to the Pocock design. However, the final critical value is close to the critical value for the fixed-sample design.

c) Haybittle-Peto method

Haybittle (1971) [15], and Peto (1976) [31] suggested a simple approach of using a very large critical value for the interim analysis and adjusting the final analyses to achieve the desired type-I error rate. The same critical value is used in every interim analysis with a threshold p-value of 0.001. However, the final analysis is still assessed at the expected level of significance, making the final critical value close to the critical value for the fixed-sample test.

The condition for the Haybittle-Peto method is,

$$P_{H_0}(|T_1^*| \geq cv_1 \quad or \ \ ... \ \ or \quad |T_K^*| \geq cv_K) = \alpha, \tag{1.6}$$

where $cv_1 = cv_2 = ... = cv_{K-1} \approx 3$ and $cv_K = Z_{\alpha/2}$.

However, the main drawback of this method is that it is very conservative, and the chance of terminating the test in the early stages is less likely. The critical values of all three methods discussed are in Table 1.1 to facilitate straightforward implementation without requiring further computation. The critical values (cv) and the expected alpha spending (α') have been computed for the significance level $\alpha = 0.05$ and $K = 2, 3, 4$, and 5, respectively.

Table 1.1: Critical values (cv) and expected alpha spending (α') for each stage of a group sequential analysis for Pocock, OBF and Haybittle-Peto methods.(Retrieved from https://online.stat.psu.edu/stat509/lesson/9/9.5)

K	Stage	Pocock		O'Brien - Fleming		Haybittle-Peto	
		cv	α'	cv	α'	cv	α'
2	1	2.178	0.0294	2.782	0.0054	3	0.002
	2	2.178	0.0294	1.967	0.0492	1.96	0.05
3	1	2.289	0.0221	3.438	0.0006	3.291	0.001
	2	2.289	0.0221	2.431	0.0151	3.291	0.001
	3	2.289	0.0221	1.985	0.0471	1.96	0.05
4	1	2.361	0.0182	4.084	0.00005	3.291	0.001
	2	2.361	0.0182	2.888	0.0039	3.291	0.001
	3	2.361	0.0182	2.358	0.0184	3.291	0.001
	4	2.361	0.0182	2.042	0.0412	1.96	0.05
5	1	2.413	0.0158	4.555	0.000005	3.291	0.001
	2	2.413	0.0158	3.221	0.0013	3.291	0.001
	3	2.413	0.0158	2.63	0.0085	3.291	0.001
	4	2.413	0.0158	2.277	0.0228	3.291	0.001
	5	2.413	0.0158	2.037	0.417	1.96	0.05

From Table 1.1, we can see that the critical value and alpha spending are the same for each stage for a chosen K value in Pocock's method. It is clearly noticeable that the O'Brien & Fleming method has monotonically decreasing critical values. In contrast, the expected alpha spending is monotonically increasing, resulting in very low values for the first stage. In all interim analyses, the Haybittle-Peto approach has higher critical values. The final stage, on the other hand, has a critical value of 1.96. As shown, the expected alpha spending at each interim analysis of the Haybittle-Peto method is very small (0.001), resulting in 0.05 for the final stage.

1.2.3 Alpha spending function approach

Repeated statistical data analysis needs modifications to maintain the type I error rate at a specified level. Three strategies for constructing discrete sequential boundaries for clinical trials were discussed in Subsection 1.2.2. The total number of interim looks must be stated in advance for these strategies to work. To overcome this drawback, Lan and DeMets (1983) [12] proposed a more flexible method for constructing discrete sequential boundaries named as alpha spending function approach. This method is based on the selection of α^*(t), a function that describes the rate at which the significance level α is spent. In other words, every time an interim analysis is performed, a portion of the overall alpha is "spent". The decision boundary is determined by past and current decision times, but it does not depend on the number of observations at the k^{th} analysis or the maximum number of analyses, K, as well as the future decision times. The function α^*(t) is monotonically increasing in t, with $\alpha^*(0) = 0$ and $\alpha^*(1) = \alpha$, where α is the expected overall type I error rate.

Let T_1^* be the test statistic for the first analysis. Then the critical value, cv_1 for the two-sided case is defined by,

$$P_{H_0}(|T_1^*| \geq cv_1) = \alpha^*(t_1), \tag{1.7}$$

where $\alpha^*(t_1)$ is the type I error rate spent at the first analysis.

The condition for the critical value for the second stage(cv_2) is,

$$P_{H_0}(|T_1^*| < cv_1, |T_2^*| \geq cv_2) = \alpha^*(t_2) - \alpha^*(t_1). \tag{1.8}$$

Here, $\alpha^*(t_2)$ is the type I error rate spent up to the second stage and $(\alpha^*(t_2) - \alpha^*(t_1)$) represents the type I error rate spent only at the second stage of the analysis. A similar criterion can be used to calculate the critical values of the remaining stages. The general condition to find the critical value for k^{th} stage is,

$$P_{H_0}(\bigcap_{i=1}^{k-1}\{|T_i^*| < cv_i\}, |T_k^*| \geq cv_k) = \alpha^*(t_k) - \alpha^*(t_{k-1}), \tag{1.9}$$

which can be solved using a recursive integration formula.

Over the years, researchers have proposed different forms of alpha spending functions. Some of the functions approximate the group sequential boundaries discussed in Subsection 1.2.2. This research makes use of three alternative alpha spending functions that are based on the ideas represented by Pocock, O'Brien & Fleming, and Haybittle-Peto.

1.2.4 Markov chain methods for sequential designs of clinical trials

The stochastic model known as Markov chains, named after Andrey Markov, is a series of potential events where predictions or probabilities for the subsequent state are only based on the prior event state, not the states before. In other words, the probability of $(n+1)^{th}$ step will depend solely on the n^{th} step and not the entire series of steps before n^{th} step. This quality is often referred to as 'memorylessness' or the Markov property.

Markov chains have numerous applications as statistical models for analysing sequential data. Kemperman (1961) [21] describes sequential testing using an analytic approach to Markov processes. Choi (1968) [8] developed a novel sort of closed sequential design for clinical trials based on a fundamental equation of Markov chains. Choi's designs are simple to implement and can be used for a variety of sequential clinical studies.

Woodall and Reynolds (1983) [49] introduced the Markov chain approach to approximate the Sequential Probability Ratio Test (SPRT) using test statistic values of SPRT. Tests that can precisely be represented by discrete Markov chains with absorbing barriers have been used to simulate the properties of the SPRT. The method is applicable to both discrete and continuous test statistics. Examples of the grouped rank test for sequential data of two-sample design proposed by Wilcoxon et al. (1963) [48] are used to demonstrate that the Markov chain approach produces extremely good approximations.

Douke (1994) [10] presented a sequential design based on Markov chains for identifying the better of two medical treatments in clinical trials by using delayed observations with a reaction time lag. The goal of his study was to find the best sequential design based on the minimum expected loss, for which an iterative Markov chain formula based on the trinomial random walk on a lattice diagram was used.

Pulford (2003) [35] presented a technique for analysing the sequential probability ratio test for automatic track maintenance. The approach uses finite state Markov chain analysis to calculate the probabilities of track confirmation events and their anticipated timings. Yi (2013) [50] used the Markov chain method to compute the statistical power for response adaptive designs. Paxinou et al.(2021) [30] recently conducted research on using Markov chain models to analyse sequence data in scientific experiments. The ability of students to conduct a science experiment is estimated by sequential analysis using a Markov chain model.

1.3 Motivation of the study

Group sequential analysis plays an important role in clinical trials because of its advantages over fixed sample designs. The advantages include the fact that data can be assessed as they are obtained. It also promotes faster decisions. Over the years, many researchers worked on different concepts to analyse data sequentially. Most of the well-known methods are proposed for numerical data, especially normally distributed.

However, the data collected from clinical trials are not always continuous. Sometimes, the data may be qualitative/categorical, necessitating consideration of the counts within each category. In certain cases, data can be dichotomous, such as medical test findings (negative or positive), survival status (alive or dead), readmission (yes or no), family medical history (yes or no), smoking status (yes or no), and so on. Therefore, analysts have to chose specific sequential testing methods that can work with this dichotomous data.

Researchers have been interested in finding a method that works well with binary data for several decades. Kulldroff et al. (2011)[23] proposed a Maximized Sequential Probability Ratio Test(MaxSPRT) for binary data comparing individuals exposed to a drug or vaccine with matched unexposed individuals. A list of critical values was produced for various sample sizes and matching ratios (unexposed individuals per exposed).

Silva et al. (2020) [38] provided an excellent explanation of the theory underlying the computations of MaxSPRT and a proposal for optimal alpha spending for sequential analysis with binary data. However, the approaches employ Bernoulli trials, in which individuals are added one at a time in each stage of the analysis. Therefore, we are more focused on a method which can be used to analyse groups of binary data.

With the influence of the research works mentioned above, this study finds critical values considering groups of binary data at each stage using three different alpha spending functions based on Pocock, O'Brien & Fleming(OBF) and Haybittle-Peto methods. That is, the alpha spending functions used in this study are modified with the idea of the Pocock, OBF, Haybittle-Peto methods. For example, the alpha spending function influenced by Pocock's method does not use the same alpha spending values proposed by Pocock, but it follows the idea of spending the same alpha at each stage. Therefore, the modified alpha spending function spends the same portion of alpha (α/K, where K is the maximum number of stages) at each stage in order to obtain the specified significance level (α). Similarly, the alpha spending function influenced by OBF method does not use the same alpha spending values proposed by O'Brien and Fleming, but it follows the idea of monotonically increasing alpha spending at each stage. Moreover, the alpha spending function impacted by Peto's method has extremely tiny alpha spending at each interim analysis and the remainder goes to the final stage to determine the overall significance level.

1.3.1 Outline of the thesis

This thesis focuses on developing a group sequential analysis with binary responses and calculating the critical values, type I error rate and power. The critical values are obtained using an iterative Markov chain approach to satisfy the alpha spending at each stage. Chapter 2 provides a brief overview of the Markov chain technique, as well as an explanation of how we used the Markov chain approach to binomial data in a single sample scenario. The construction of Log Likelihood Ratio (LLR), type I error rate and power computation are explained stage-wise, particularly for the first stage, which is unconditional, and then the second and k^{th} stages, which are conditional on the preceding stage. The critical values, type I error rate and power for the three-stage design have been calculated, and results are presented for different sample sizes. The convergence of critical values with increasing sample size was depicted using graphs.

In Chapter 3, the proposed group sequential analysis is extended to compare two binomial proportions. Similar to the single sample scenario, the critical value, type I error rate, and power calculations are described stage-wise: first interim analysis as unconditional case and the second interim analysis for the conditional case. Then the k^{th} analysis/stage is explained with a general technique that may be applied to any conditional stage ($k > 1$). The alpha spending functions are constructed based on Pocock, O'Brien & Fleming, Haybittle-Peto methods, and the results obtained from the three methods are compared. In Chapter 4, the findings of the study are examined, and a summary is provided along with some suggestions for future research work.

Chapter 2

Markov chain method to determine critical values

2.1 Iterative Markov chain method for one sample case

A Markov chain is a stochastic model describing a sequence of events where the probability of an event depends only on the previous event. In other words, given the present, the future is independent of the past.

Let $X_1, X_2, \dots, X_n$ be a sequence of random variables with the Markov property. Then the probability of the $(n+1)^{th}$ event is,

$$Pr(X_{n+1} = x | X_1 = x_1, X_2 = x_2, \dots X_n = x_n) = Pr(X_{n+1} = x | X_n = x_n), \qquad (2.1)$$

Markov chains have a wide range of applications as statistical models for real-life processes in different fields. Kulldorff et al. (2011) [23] have used an iterative Markov chain approach to calculate critical values for Bernoulli data in drug and vaccine safety surveillance. The methodology that they employed, had an impact on this study. The idea of utilizing the log-likelihood ratio as the test statistic and the approach of employing an iterative Markov chain to determine critical values are influenced by Kulldorff's research. Therefore, before commencing the methodology of my study, let's examine Kulldorff's approach.

Kulldorff et al. (2011) [23] suggested a maximized sequential probability ratio test (MaxSPRT) based on a composite alternative hypothesis as it is effective across a wide variety of relative risk levels in the context of drug and vaccine safety surveillance. Since accurate estimates of the anticipated number of incidents are not available before starting the drug and vaccine safety surveillance, they have considered different designs to collect more information. Collecting data on potential detrimental effects from the times when the subject was exposed to the drug/vaccine and when they were not exposed is one possible design. In a self-controlled design, for instance, an exposed time period after vaccination can be compared to an unexposed time period before vaccination for the same individual or to an unexposed time period long after vaccination. Alternately, the individuals exposed to the drug/vaccine can be compared with matching individuals who are not exposed. As stated in their study, the Poisson model can only be utilized if the unexposed time period is significantly longer than the exposed period. Otherwise, a binomial probability model should be used to derive the log-likelihood function and the critical values. The concept of using a binomial model in this study is similar to a series of coin flips (adverse events) that may result in heads or tails (exposed or unexposed). When the exposed and unexposed time periods have the same duration, it is referred as 1:1 matching ratio or z=1 where z is the ratio of the length of the matched unexposed time period and the length of the exposed time period. Then the probability of a head under the null hypothesis is p = 0.5. Similarly, if the matching ratio is 1:3 then p=0.25. In their study, the "increased relative risk due to the drug/vaccine" (RR) [23] is the unknown parameter. If RR=1, then the parameters of binomial distribution is n and p=[1/(1+1)]=0.5 when matching ratio is 1:1 (z=1). On the other hand, if z=2, then the parameters of the binomial distribution are n and p=[1/(1+2)]=0.67 under RR=1.

Therefore, this probability can be computed as,

$$p = \frac{1}{1 + z/RR}. \tag{2.2}$$

Even though their derivations are applicable for any pair of hypotheses, Kulldorff et al. (2011) [23] mainly focused on the standard form of non-excess versus elevated risk given as,

$$H_0 : RR = 1 \qquad vs \qquad H_1 : RR > 1. \tag{2.3}$$

Let, n represents the number of adverse events recorded so far during the sequential data collection, and c_n represents the number of adverse events that have occurred within the exposed time periods among those n events ($c_n < n$), and z is the ratio of the length of the matched unexposed time period and the length of the exposed time period. Then, the log-likelihood ratio is the test statistic for Binomial distribution and can be written as,

$$LLR_n = c_n\, ln\Big(\frac{c_n}{n}\Big) + (n - c_n)\, ln\Big(\frac{n - c_n}{n}\Big) - c_n\, ln\Big(\frac{1}{z+1}\Big) - (n - c_n)\, ln\Big(\frac{z}{z+1}\Big). \tag{2.4}$$

when $\frac{zc_n}{n-c_n} > 1$ and 0 otherwise.

2.1.1 Modelling Binomial data

Let's consider a group sequential analysis for a single sample with binomial data. A null hypothesis with the probability of success equal to a given value, p_0, is tested against a two-sided alternative hypothesis as,

$$H_0 : p = p_0 \qquad vs \qquad H_a : p \neq p_0. \tag{2.5}$$

Kulldorff et al. (2011) [23] considered the sequential entries of individuals in their study. That is, the analysis is carried out by considering a single entry at a time until N individuals have been analysed. On the other hand, our research is based on the sequential entries of a group of participants at each interim analysis. At each stage of sequential testing, a test statistic is compared to a critical value which is an upper signalling threshold. The test statistic for this study is the log-likelihood ratio (LLR), defined as,

$$LLR = 2ln\left[\frac{L(\hat{p}|H_a)}{(L(p_0|H_0)}\right] = 2[ln(L(\hat{p}|H_a)) - ln(L(p_0|H_0))], \tag{2.6}$$

where $\hat{p}$ is the maximum likelihood estimator (MLE) for the parameter p.

Prior to starting the analysis, the maximum number of stages, 'K', must be determined. The trial may be terminated due to the rejection of the null hypothesis at any interim analysis, or it may proceed to the next stage. As there is no further continuation, the final stage (K^{th} stage) can result in acceptance or rejection of the null hypothesis.

The first stage of group sequential analysis is always unconditional. If the trial is to be continued after the first interim analysis, the second stage is conditional on the first stage. Similarly, all subsequent stages are dependent on the preceding stage. Consequently, the methodology is broken down into stages.

First stage (Unconditional stage)

At the beginning of the procedure, n_1 participants comprise a single sample, and s_1 is the number of successes. We denote the number of failures, which is equal to $(n_1 - s_1)$, as f_1. Therefore, the log-likelihood ratio for the first stage (LLR_1) is defined as follows.

Assuming a binomial distribution, the log-likelihood under the null hypothesis, H_0 (given in 2.5) is,

$$[l(p_0|H_0)] = ln\binom{n_1}{s_1} + s_1 ln(p_0) + (n_1 - s_1)ln(1 - p_0). \tag{2.7}$$

The log-likelihood under the alternative hypothesis is,

$$[l(\hat{p}|H_a)] = ln\binom{n_1}{s_1} + s_1 ln(\hat{p}) + (n_1 - s_1)ln(1 - \hat{p}), \tag{2.8}$$

where

$$\hat{p} = \frac{s_1}{n_1}. \tag{2.9}$$

Therefore, the test statistic for the first stage(LLR_1) is,

$$LLR_1 = 2\{[l(\hat{p}|H_a)] - [l(p_0|H_0)]\}, \tag{2.10}$$

Considering that s_1 can have any value between zero and n_1, inclusive, the calculation of LLR_1 can be simplified as,

$$LLR_1 = 2[s_1 ln(s_1) + f_1 ln(f_1) - n_1 ln(n_1) - s_1 ln(p_0) - f_1 ln(1 - p_0)], \tag{2.11}$$

where

$$f_1 = n_1 - s_1. \tag{2.12}$$

For a given sample size(n_1) and the probability of success under the null hypothesis (p_0), the test statistic LLR_1 only depends on the number of successes at the given stage. As a result of the discrete nature of data, the test statistic can have only a finite number of values.

Given that s_1 may take on any value between zero and n_1, there are $(n_1 + 1)$ distinct values for s_1. Consequently, there are $(n_1 + 1)$ possible test statistic values at the first stage (T^{1st}). That is,.

$$T^{1st} \in (T_1^{1st}, T_2^{1st}, \ldots, T_{n_1+1}^{1st}).$$

Therefore, the j^{th} test statistic value of the first stage, T_j^{1st} is,

$$T_j^{1st} = LLR_1(S_1 = j - 1); \qquad j = 1, 2, ..., (n_1 + 1) \tag{2.13}$$

In this study, the log-likelihood ratio (LLR) is not a monotonically increasing or decreasing function with the number of successes. Since it is a non-linear function with a minimum, the calculated LLR values are ordered in ascending order and denoted as,

$$T_{(1)}^{1st}, T_{(2)}^{1st}, \ldots, T_{(n_1+1)}^{1st}.$$

Since sample size, n_1 and the probability of success under null hypothesis (p_0) is fixed (or given), the log-likelihood ratio for stage 1 (LLR_1) depends only on s_1 (see equation (2.11)). Therefore, the probability of LLR_1 depends only on the probability of successes at first stage, s_1 (given n_1 and p_0),

$$Pr[LLR(S_1 = s_1|n_1, p_0)] = Pr(s_1|n_1, p_0) = \binom{n_1}{s_1}(p_0)^{s_1}(1 - p_0)^{(n_1 - s_1)}, \tag{2.14}$$

where $s_1 = 0, 1, \ldots, n_1$.

After computing all possible LLR values and their probabilities, we can determine the critical value of the first stage (cv_1). Therefore, let us define,

$$c_1 = min\{j \in \mathbb{N} : Pr(T^{1st} \geq T^{1st}_{(j)} | p = p_0) \leq \alpha_1\}, \tag{2.15}$$

where α_1 is the amount of alpha spending for the first stage.

Note that, for a group sequential design with a given maximum number of stages K, a sequence of signalling thresholds $cv_1, cv_2, ..., cv_K$ is obtained to satisfy the overall level of significance $\alpha \in (0, 1)$.

Using a searching algorithm, we can determine the critical value. When all possible LLR values are ordered along with their probabilities, we choose the middle LLR value and determine the probability of type I error rate associated with that. If the type I error rate is greater than the expected alpha spending of the stage, we limit the next search to LLR values greater than the selected(middle) one. If the computed probability is less than the expected alpha spending of the stage, then in the next search, we only evaluate LLR values less than the selected one. In a similar manner, we only use half of the data in the subsequent search. From that half, chose the middle value and conduct the searching process. This iteration continues until we find two consecutive LLR values, one with type I error rate higher than the expected alpha spending and the other one with type I error rate lower than the expected alpha spending. Then we determined the critical value of the stage to be the LLR with a type I error rate that is nearly equal to the expected alpha spending but does not exceed it.

According to equation (2.15), $T^{1st}_{(c_1)}$ becomes the critical value of the first stage.

$$cv_1 = T^{1st}_{(c_1)}, \tag{2.16}$$

such that, $\left\{Pr\left(T^{1st}_{(c_1)}\right) + Pr\left(T^{1st}_{(c_1+1)}\right) + ... + Pr\left(T^{1st}_{(n_1+1)}\right)\right\} \leq \alpha_1$.

Type I error occurs when we reject the null hypothesis and incorrectly assert that the study discovered significant differences when there were, in fact, none. This can be identified as a false positive and should be controlled in each stage of the sequential analysis. Hypothesis tests at each step guarantee that the chance of a false positive test result does not exceed a certain level α, resulting in just a small fraction of truly unsuccessful treatments being used in the long run (Wassmer & Brannath 2016) [43].

After determining all possible test statistic values, we can calculate the probability of error associated with each test statistic value. For a given value of T^*, this error rate can be explained as the probability (under the null hypothesis) of all the test statistics which is greater than or equal to this given value, that is,

$$error(T^*) = Pr(T \geq T^*). \tag{2.17}$$

For a particular test statistic value, if this error rate is almost equivalent to the amount of alpha that is supposed to be spent during the specified stage, we select that test statistic as the critical value.

For example, consider a trial with sample size 19 and $p_0 = 0.4$. There are 20 different LLR values for $s_1 = 0, 1, ..., 19$. We may arrange all LLR values in ascending order and calculate the type I error rate associated with each LLR value.

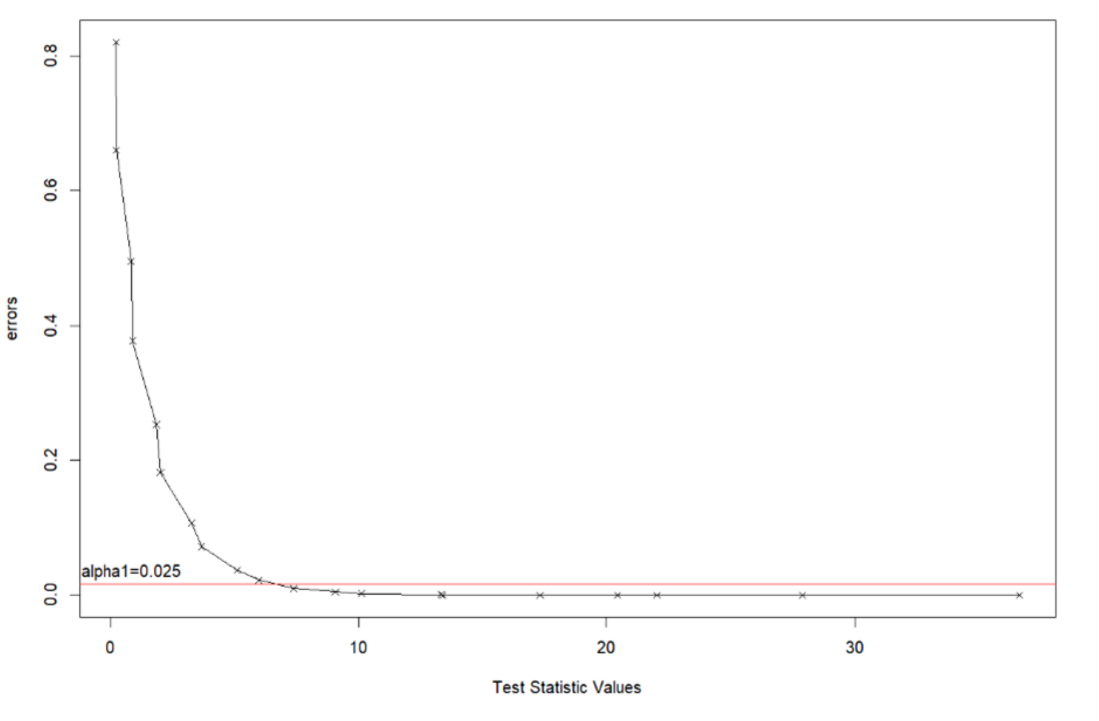

Figure 2.1: Type I error rate of 20 distinct test statistic values.

Figure 2.1 presents twenty distinct test statistic values, along with their respective error rates. The alpha spending value is equal to 0.025, which is represented by the red line. At the intersection of the graph and the red line, we can find the critical value. Once the critical value for the first stage (cv_1) is identified, the type I error rate is computed using the probabilities associated with each test statistic value. Define e_1^* as the type I error rate of the first stage; then,

$$e_1^* = Pr(T^{1st} \geq cv_1) = Pr(T^{1st} \geq T_{(c_1)}^{1st}) = \sum_{i=c_1}^{n_1} Pr(T_{(i)}^{1st}). \tag{2.18}$$

Since cv_1 defines the boundary of the rejection region of the first stage; $Pr(T^{1st} \geq cv_1)$ refers to the probability of all test statistic values in the rejection region. As the sample size (n_1) and the probability of success under the null hypothesis (p_0) are pre-determined, the probability of the test statistic values that fall in the rejection region is computed using the s_1 values that were rejected in the first stage. Therefore, the type I error rate for the first stage is expressed as a sum of partial probabilities,

$$e_1^* = \sum_{s_1 \in s_1^R} \binom{n_1}{s_1} (p_0)^{s_1} (1-p_0)^{n_1-s_1}. \tag{2.19}$$

where S_1^R refers to the set of s_1 values in the rejection region / absorbing state for Stage 1.

Second stage

The trial will continue to the second stage only if the null hypothesis is not rejected in the first stage. When the critical value of stage one is defined, it creates a boundary for the rejection region. Some values in the S_1 vector ($S_1 = S_1^R \bigcup S_1^C$) fall in the rejection region while the rest fall in the continuation region. All the s_1 values which are rejected in the first stage are referred to as the "absorbing state". The rest of the s_1 values in the "non-absorbing state" will be considered in the second stage. We denote by S_1^C, the set of s_1 values in continuation region / non- absorbing state for Stage 1.

After the first interim analysis, the data in the continuation region is combined with a sample of size n_2 with s_2 number of successes. Therefore, all possible values of s_2 are added to each s_1 value in the continuation region, resulting in a set containing the possible total number of successes at the second stage (S_2^T) as,

$$S_2^T = s_1 + s_2 \mid s_1 \in S_1^C, \forall s_2. \tag{2.20}$$

The statistical test is conducted by considering the combined sample; therefore, the combined sample size at stage two will be $N_2 = n_1 + n_2$. Then, the test statistic for the second stage is,

$$LLR_2 = 2\ \ [S_2^T ln(S_2^T) + F_2^T ln(F_2^T) - N_2 ln(N_2) - S_2^T ln(p_0) - F_2^T ln(1-p_0)], \quad (2.21)$$

where F_2^T is the total number of failures in the combined sample at the second stage, which can be calculated as,

$$F_2^T = N_2 - S_2^T. \quad (2.22)$$

The second stage consists of performing the same procedures as the first stage; all possible test statistic (LLR) values are calculated and ordered in ascending order with their corresponding probabilities. The probabilities of LLR for the second stage are conditional on the first stage.Note that, there are several possible combinations of $s_1(\in\ S_1^C)$ and s_2 that result in a given value in S_2^T. Therefore, the conditional probabilities are calculated as,

$$\begin{aligned} &Pr[LLR(S_2^T = s_1 + s_2 | s_1 \in S_1^C, \forall s_2)] \\ &\qquad = \sum_{s_1 \in S_1^C, \forall s_2} \left\{ \binom{n_1}{s_1} (p_0)^{s_1} (1-p_0)^{(n_1 - s_1)} * \binom{N_2}{s_2} (p_0)^{s_2} (1-p_0)^{(N_2 - s_2)} \right\}. \end{aligned} \quad (2.23)$$

Let,

$$c_2 = min\{j \in \mathbb{N} : Pr(T^{2nd} \geq T_{(j)}^{2nd} | p = p_0) \leq \alpha_2\}, \quad (2.24)$$

where α_2 is the amount of alpha spending for the second stage. Therefore, $T_{(c_2)}^{2nd}$ becomes the critical value of the second stage.

$$cv_2 = T_{(c_2)}^{2nd}. \quad (2.25)$$

Once the critical value (cv_2) is identified, the type I error rate of the second stage e_2^* is computed using the conditional probabilities associated with each test statistic value as,

$$e_2^* = Pr(T^{2nd} \geq cv_2) = Pr(T^{2nd} \geq T_{c_2}^{2nd}) = \sum_{\forall i \geq c_2} Pr(T_{(i)}^{2nd}). \quad (2.26)$$

Since cv_2 defines the rejection region of second stage, $Pr(T^{2nd} \geq cv_2)$ is the probability of all test statistic values in the rejection region of the second stage. As the combined sample size at the second stage ($N_2 = n_1 + n_2$) and probability of success under the null hypothesis (p_0) are known, the probability of test statistic values falling in the rejection region can be derived using S_2^T values rejected at the end of the second stage.

Thus, the type I error rate for the second stage is computed as,

$$\begin{aligned} e_2^* &= \sum_{s_2^T \in S_2^R} Pr[LLR(S_2^T = s_1 + s_2 | s_1 \in S_1^C, \forall s_2)] \\ &= \sum_{s_1 \in S_1^C, (s_1+s_2) \in S_2^R} \Big\{ \binom{n_1}{s_1} (p_0)^{s_1} (1-p_0)^{(n_1-s_1)} * \binom{N_2}{s_2} (p_0)^{s_2} (1-p_0)^{(N_2-s_2)} \Big\}. \end{aligned} \tag{2.27}$$

k^{th} stage (any $k > 1$)

The above-discussed procedure can be generalized to use for any stage after the first interim analysis. If we have reached the k^{th} stage of the sequential design, it signifies that there was insufficient evidence to reject the null hypothesis at the $(k-1)^{th}$ stage. Thus the k^{th} stage is conditional on $(k-1)^{th}$ stage.

Let's define the following sets in order to construct the log-likelihood ratio and the associated probabilities of k^{th} stage.

S_{k-1}^R - set of possible number of success values in rejection region of stage (k-1) / absorbing state for stage (k-1).

S_{k-1}^C - set of possible number of success values in continuation region of stage(k-1)/ non- absorbing state for stage (k-1).

At the beginning of k^{th} stage (any $k > 1$) of the sequential test design, a new sample of size n_k is combined with the data that is in the continuation region of $(k-1)^{th}$ stage. Assume there's s_k number of successes out of these n_k. Therefore, the possible total number of successes at the k^{th} stage (S_k^T) can be defined as follows. The values in S_k^T, is formed by combining each of the S_{k-1}^T values in the continuation region of the previous stage ($S_{(k-1)}^C$) with all the s_k values of stage k; so that,

$$S_k^T = S_{k-1}^T + s_k | \ S_{k-1}^T \in S_{k-1}^C, \forall s_k. \tag{2.28}$$

where s_k represents all possible number of success values at stage k. Thus s_k can take any value between zero and the size of the sample introduced at stage k (n_k).

$$s_k = (0, 1, 2, ..., n_k). \tag{2.29}$$

As the test is conducted for the cumulative data, the combined sample size at stage k can be defined as N_k where,

$$N_k = n_1 + n_2 + ... + n_k. \tag{2.30}$$

Therefore, the LLR_k at the k^{th} stage is,

$$LLR_k = 2\Big[S_k^T ln(S_k^T) + F_k^T ln(F_k^T) - N_k ln(N_k) - S_k^T ln(p_0) - F_k^T ln(1 - p_0)\Big], \tag{2.31}$$

where F_k^T is the total number of failures in the combined sample at the k^{th} stage which can be calculated as,

$$F_k^T = N_k - S_k^T. \tag{2.32}$$

After calculating all possible LLR values, corresponding probabilities must be calculated. There can be several possible combinations of $s_{(k-1)}(\in S_{k-1}^C)$ and all $s_k \in S_k$ that result in a given value in S_k^T. Therefore, for given N_{k-1}, N_k, p_0 values and any $k > 1$,

$$\begin{aligned} Pr[LLR(S_k^T)] = \sum_{s_{k-1}^T \in S_{k-1}^C, \forall s_k} & \Big\{ \binom{N_{k-1}}{s_{k-1}^T} (p_0)^{s_{k-1}^T} (1 - p_0)^{(N_{k-1} - s_{k-1}^T)} \\ & * \binom{N_k}{s_k} (p_0)^{s_k} (1 - p_0)^{(N_k - s_k)} \Big\}. \end{aligned} \tag{2.33}$$

When all possible test statistic values (LLR) for the k^{th} stage have been obtained, they are arranged in ascending order along with their probability. Let's assume there are 'M_k' number of possible S_k^T values at k^{th} stage, then 'M_k' number of possible test statistic values (LLR) can be obtained. Therefore, all possible LLR values for k^{th} stage are, $T_1^{kth}, T_2^{kth}, ..., T_{M_k}^{kth}$ and its ordered version is, $T_{(1)}^{kth}, T_{(2)}^{kth}, ..., T_{(M_k)}^{kth}$.

Define,

$$c_k = min\{j \in \mathbb{N} : Pr(T^{kth} \geq T^{kth}_{(j)} | p = p_0) \leq \alpha_k\}, \tag{2.34}$$

where α_k is the amount of alpha sending for the k^{th} stage.

Therefore, $T^{kth}_{(c_k)}$ becomes the critical value of the k^{th} stage.

$$cv_k = T^{kth}_{(c_k)}, \tag{2.35}$$

which makes $\left\{Pr\left(T^{kth}_{(c_k)}\right) + Pr\left(T^{kth}_{(c_k+1)}\right) + ... + Pr\left(T^{kth}_{(M_k)}\right)\right\} \leq \alpha_k$.

Then, the type I error rate of the $k^{th} stage$ e^*_k can be defined as,

$$e^*_k = Pr(T^{kth} \geq cv_k) = Pr(T^{kth} \geq T^{kth}_{c_k}) = \sum_{\forall i \geq c_k} Pr(T^{kth}_{(i)}). \tag{2.36}$$

This refers to the probability of all test statistic values of k^{th} stage that falls in the rejection region. For a known sample size (N_k) and probability of success under the null hypothesis (p_0), the above-mentioned probability only depends on the S^T_k values, which are rejected at the end of the k^{th} stage. Thus, the type I error rate for the k^{th} stage can be computed as,

$$\begin{aligned} e^*_k = \sum_{s^T_{k-1} \in S^C_{k-1}, (s^T_{k-1} + s_k) \in S^R_k} & \Bigg\{ \binom{N_{k-1}}{s^T_{k-1}} (p_0)^{s^T_{k-1}} (1-p_0)^{(N_{k-1} - s^T_{k-1})} \\ & * \binom{N_k}{s_k} (p_0)^{s_k} (1-p_0)^{(N_k - s_k)} \Bigg\}. \end{aligned} \tag{2.37}$$

Overall type I error rate calculation

If the trial consists of a total of K number of stages, the overall error rate (e_T) is the summation of error rates at each stage.

$$\begin{aligned} e_T = P_{H_0}(T^{1st} \in R_1) + P_{H_0}(T^{1st} \in C_1, T^{2nd} \in R_2) + ...+ \\ P_{H_0}(T^{1st} \in C_1, ..., T^{(K-1)th} \in C_{K-1}, T^{Kth} \in R_K), \end{aligned} \tag{2.38}$$

where R_k, C_k denotes the rejection region and continuation region of k^{th} stage, respectively.

Thus, the overall type I error rate is computed as follows,

$$\begin{aligned} e_T &= e_1^* + e_2^* + ... + e_K^*, \\ &= Pr\Big(T^{1st} \geq cv_1\Big) + Pr\Big(T^{2nd} \geq cv_2\Big) + ... + Pr\Big(T^{Kth} \geq cv_K\Big), \\ &= Pr\Big(T^{1st} \geq T^{1st}_{(c_1)}\Big) + Pr\Big(T^{2nd} \geq T^{2nd}_{(c_2)}\Big) + ... + Pr\Big(T^{Kth} \geq T^{Kth}_{(c_K)}\Big), \\ &= \sum_{i=c_1}^{n_1} Pr\Big(T^{1st}_{(i)}\Big) + \sum_{\forall i \geq c_2} Pr\Big(T^{2nd}_{(i)}\Big) + ... + \sum_{\forall i \geq c_K} Pr\Big(T^{Kth}_{(i)}\Big). \end{aligned} \tag{2.39}$$

The sequence of critical values $cv_1, cv_2, ..., cv_K$ (where K is the maximum number of stages) is obtained to satisfy the overall level of significance $\alpha \in (0, 1)$. Each critical value is obtained to satisfy the alpha spending at the corresponding stage. Different methods are used to construct the alpha spending functions that attain the given overall significance level.

In Section 2.2 we discuss three different alpha spending functions based on the methods introduced by Pocock, O'Brien & Fleming and Haybittle-Peto. The first set of analyses was done using the same alpha spending at each stage, as introduced in the Pocock method. If α denotes the overall significance, then the portion of type I error rate spent in each stage is α/K for a sequential test which has a total of K number of analyses. The second set of analyses was done by calculating monotonically decreasing critical values using the idea of the O'Brien & Fleming method. The percentage of alpha spending at each stage is calculated to approximate the values of O'Brien & Fleming method. The third set of analyses followed the Haybittle-Peto's method, which uses the same threshold p-value of 0.001 for all interim analyses and the expected significance level (α) for the final analysis. All the tests were constructed for different sample sizes, considering equally sized samples. Critical values, type I error rate and power have been calculated, and the results were discussed.

2.1.2 Power calculation

Power $(1-\beta)$ of a test is defined as the probability of rejecting null hypothesis when the alternative hypothesis is true. For a given specific alternative hypothesis $(H_a : p = p_1)$, the power is given by,

$$\begin{aligned}(1-\beta) = P_{H_1}(T^{1st} \in R_1) + P_{H_1}(T^{1st} \in C_1, T^{2nd} \in R_2) + ...+ \\ P_{H_1}(T^{1st} \in C_1, ..., T^{(K-1)th} \in C_{K-1}, T^{Kth} \in R_K).\end{aligned} \tag{2.40}$$

For a given critical value (cv_k), the probability of all LLR values which are in rejection region (of k^{th} stage) when H_a is true, can be computed as,

$$(1-\beta)_k = Pr(LLR \geq cv_k | p = p_1) = \sum_{i=c_k}^{M_k} Pr(T_{(i)}^{kth} | p = p_1). \tag{2.41}$$

where M_k is the number of possible test statistic (LLR) values at the k^{th} stage.

$$\begin{aligned}(1-\beta)_k = \sum_{s_{k-1}^T \in S_{k-1}^C, s_k^T \in S_k^R} \Big\{ \binom{N_{k-1}}{s_{k-1}^T} (p_1)^{s_{k-1}^T} (1-p_1)^{(N_{k-1}-s_{k-1}^T)} \\ * \binom{N_k}{s_k} (p_1)^{s_k} (1-p_1)^{(N_k - s_k)} \Big\}.\end{aligned} \tag{2.42}$$

The power of the test can be expressed as a sum of partial probabilities.

$$(1-\beta) = \sum_{k=1}^{K} (1-\beta)_k = \sum_{k=1}^{K} \Big\{ \sum_{i=c_k}^{M_K} Pr(T_{(i)}^{kth} | p = p_1) \Big\}. \tag{2.43}$$

2.2 Critical value calculation for two-stage and three-stage designs

2.2.1 Two-stage design

Consider a two-stage RCT with alpha spending functions based on Pocock, O'Brien & Fleming, and Haybittle-Peto methods. To obtain a 0.05 confidence level ($\alpha = 0.05$), the following alpha spending values were used as the predicted type I error rate of each stage.

Table 2.1: Expected alpha values for each stage of two-stage design

	Stage1	Stage2
Pocock	0.025	0.025
OBF	0.010	0.040
Peto	0.001	0.049

The alpha spending under the Pocock design is calculated by dividing the overall significance level by the number of stages to get the same amount of alpha spending at each stage. OBF alpha spending values are monotonically increasing, spending a very small amount of alpha at the first stage. In Haybittle-Peto's method, the alpha spent at each stage prior to the final stage is 0.001, and the remainder is spent on the final stage.

Using the alpha spending values above, critical values and type I error rates have been calculated for each stage of a two-stage RCT along with the overall type I error rate for various sample sizes considering the hypotheses as,

$$H_0 : p = 0.4 \qquad vs \qquad H_a : p \neq 0.4 \tag{2.44}$$

Tables 2.2, 2.3 and 2.4 shows the critical values (cv_1, cv_2) and the associated type I error rate at each stage (e_1, e_2) along with the overall type I error rate for various sample sizes with the same sample size(n) increment at each stage.

Table 2.2: Critical values and type I error rates for a two-stage analysis based on Pocock's method.

n	cv1	cv2	e1	e2	Total error rate
5	9.1629	6.6959	0.010240000	0.015090586	0.025330586
10	6.6959	5.9575	0.018341171	0.017708755	0.036049926
15	6.8475	5.4910	0.014519696	0.020906184	0.035425879
20	5.9575	5.5113	0.022427038	0.017319979	0.039747017
25	5.8453	5.1956	0.022639904	0.021857018	0.044496923
30	6.6277	5.4831	0.01396038	0.01939634	0.03335672
35	5.6765	5.0688	0.02342428	0.02105284	0.04447712
40	5.5113	5.0877	0.02391439	0.02196849	0.04588288
45	5.7686	4.8251	0.02136447	0.02431841	0.04568288
50	5.7057	4.9331	0.02086795	0.02435597	0.04522392
55	5.9768	5.3377	0.01844729	0.01884275	0.03729004
60	5.4831	4.8986	0.02423177	0.02353796	0.04776972
65	5.4874	4.7690	0.02269001	0.02398646	0.04667648
70	5.8052	4.9327	0.01977414	0.02368098	0.04345512
75	5.4226	4.8389	0.02445622	0.02249362	0.04694984
80	5.4656	5.0097	0.02247818	0.02177770	0.04425587
85	5.7915	4.9427	0.01952433	0.02191344	0.04143777
90	5.4737	5.1152	0.02330531	0.01985612	0.04316143
95	5.5436	4.8494	0.02121975	0.02414796	0.04536771
100	5.2518	4.8083	0.02478046	0.02238170	0.04716216
110	5.3377	4.7701	0.02474090	0.02427956	0.04902045
120	5.7429	4.9379	0.01945160	0.02272354	0.04217514
130	5.3069	4.7327	0.02483451	0.02329804	0.04813255
140	5.1984	4.7506	0.02494327	0.02423258	0.04917585
150	5.3390	4.9411	0.02410969	0.02136066	0.04547035
160	5.2631	4.7923	0.02388265	0.02235908	0.04624172
170	5.4120	4.835	0.02290973	0.02301809	0.04592781
180	5.3610	4.7133	0.02247962	0.02375154	0.04623116
190	5.5129	4.7727	0.02145596	0.02417348	0.04562944
200	5.4820	4.6716	0.02091034	0.02468673	0.04559707

Table 2.3: Critical values and type I error rates for a two-stage analysis based on O'Brien & Fleming's method.

n	cv1	cv2	e1	e2	Total error rate
5	9.1629	6.6959	0.010240000	0.015090586	0.025330586
10	10.2165	5.0773	0.007724339	0.033619681	0.04134402
15	8.7878	4.8656	0.007099804	0.035034729	0.042134533
20	9.0516	4.9691	0.005222997	0.031881205	0.037104202
25	7.9836	5.1956	0.006693157	0.026167547	0.032860704
30	7.6705	4.6841	0.008512679	0.031286717	0.039799396
35	7.4184	4.7081	0.008662804	0.033029320	0.041692125
40	7.3213	4.3998	0.009415491	0.035659930	0.045075421
45	7.2984	4.5294	0.008922285	0.035793119	0.044715403
50	7.3120	4.3202	0.009048068	0.036929099	0.045977167
55	7.3738	4.4904	0.008322680	0.036184056	0.044506736
60	7.4594	4.3400	0.008130040	0.036430161	0.044560201
65	7.5528	4.5267	0.007372329	0.035221423	0.042593752
70	7.0175	4.4158	0.009901840	0.033922136	0.043823976
75	7.1364	4.6081	0.009310046	0.032420737	0.041730782
80	7.3108	4.3238	0.008268719	0.038913724	0.047182443
85	7.4710	4.2514	0.007691930	0.037919775	0.045611706
90	7.0262	4.4599	0.009650250	0.034947334	0.044597584
95	7.2326	4.4067	0.008501543	0.034091535	0.042593077
100	7.4085	4.1802	0.007798368	0.039276610	0.047074978
110	7.2506	4.3593	0.008307757	0.034940107	0.043247865
120	7.1120	4.3773	0.008955023	0.036351489	0.045306511
130	7.0518	4.1861	0.009172055	0.038245144	0.047417199
140	6.9661	4.2403	0.009550067	0.039008936	0.048559003
150	6.9609	4.4375	0.009554073	0.034280292	0.043834365
160	6.9123	4.2961	0.009727508	0.035249204	0.044976711
170	6.9397	4.366	0.009585790	0.035604074	0.045189864
180	6.9190	4.2494	0.009611121	0.036201922	0.045813043
190	6.9663	4.3313	0.009375077	0.036267736	0.045642813
200	6.9677	4.2334	0.009297009	0.036581260	0.045878269

Table 2.4: Critical values and type I error rates for a two-stage analysis based on Haybittle-Peto's method.

n	cv1	cv2	e1	e2	Total error rate
5	9.1629	6.6959	0.010240000	0.015090586	0.025330586
10	18.3258	5.0773	0.000104858	0.036902774	0.037007632
15	14.0866	4.8656	0.000749089	0.037914604	0.038663694
20	13.3033	4.9691	0.000841080	0.033921365	0.034762445
25	13.2247	4.3190	0.000710013	0.043491765	0.044201777
30	13.5773	4.3380	0.000535595	0.046696026	0.047231621
35	11.646	4.7081	0.000807917	0.036688629	0.037496546
40	11.9149	4.3998	0.000997217	0.039632535	0.040629752
45	12.7505	4.5294	0.000641500	0.039819478	0.040460978
50	11.6905	4.3202	0.000749824	0.040993380	0.041743204
55	12.1448	4.4904	0.000800537	0.039879055	0.040679593
60	11.4228	4.3400	0.000852171	0.040024099	0.040876271
65	11.8971	3.9940	0.000861862	0.048597607	0.049459469
70	11.353	4.2075	0.000872350	0.046168945	0.047041295
75	11.8478	4.1115	0.000852625	0.045110637	0.045963262
80	11.4049	4.3238	0.000837578	0.042675063	0.043512641
85	11.9204	4.2514	0.000799906	0.041417010	0.042216915
90	11.5371	4.0084	0.000771083	0.047586420	0.048357503
95	12.0737	4.2281	0.000724311	0.044719869	0.04544418
100	11.4115	4.1802	0.000999172	0.042741113	0.043740286
110	11.6738	4.1958	0.000861009	0.045064448	0.045925457
120	11.9656	3.9873	0.000732360	0.047712003	0.048444363
130	11.2817	4.0444	0.000865837	0.048835726	0.049701563
140	11.3378	4.2403	0.000982373	0.043193565	0.044175939
150	11.7103	4.0881	0.000804851	0.044634815	0.045439666
160	11.2116	4.1632	0.000889424	0.044970044	0.045859468
170	11.3342	4.0385	0.000958722	0.045689950	0.046648672
180	11.7405	4.126	0.000773244	0.045795357	0.046568601
190	11.3368	4.0217	0.000823008	0.046112724	0.046935732
200	11.5077	4.1169	0.000859932	0.045758776	0.046618708

From Tables 2.2, 2.3 and 2.4 we see that as the sample size increases, the critical value of each stage converges. Note that due to the discrete nature of data, only a few specific values can serve as critical values. Consequently, the critical value for small sample sizes (especially $n \leq 20$) is slightly greater than the critical value for large sample sizes.

Similarly, when we examine the values of type I error rate and total error rate, we can observe that these values converge as the sample size grows. When the sample size increases, it is evident that e1 and e2 values rise to achieve the expected alpha spending at each stage. Accordingly, the total error rate may increase to achieve the expected level of statistical significance (α).

Once the critical value and type I error rate for a particular value of p_0 have been identified, the power can be calculated for various alternative hypotheses (p_1). Tables 2.5, 2.6 and 2.7 present the statistical power for two different alternative hypotheses; $p_1 = 0.2$ and $p_1 = 0.7$ for a two-stage design with $p_0 = 0.4$.

From Tables 2.5, 2.6 and 2.7, we see that the overall power increases as the sample size grows. Additionally, when the value of p_1 is far from p_0, the overall power for a given sample size is greater than when p_1 is close to p_0. For example, consider a RCT with a sample size of 30 (total sample size =60) with $p_0 = 0.4$ based on Pocock's method (Table 2.5). The overall power is 0.8753 when $p_1 = 0.2$, and it is 0.9956 when $p_1 = 0.7$. Examining the pattern from Tables 2.5, 2.6 and 2.7, we see that the overall power converges as the sample size increases. It converges rapidly when p_1 is considerably away from p_0. In addition, it is evident that when the power $> 90\%$, a greater proportion of total power is achieved in the first stage, leaving just a small portion for the second stage.

Table 2.5: Power for a two-stage design based on Pocock's method where $H_a : p_1 = 0.2$ and $H_a : p_1 = 0.7$ when $H_0 : p_0 = 0.4$.

	$p_1 = 0.2$			$p_1 = 0.7$		
n	Power1	Power2	Overall Power	Power1	Power2	Overall Power
5	0.00032	0.107433574	0.107753574	0.16807	0.242127547	0.410197547
10	0.107452109	0.317054642	0.424506751	0.382788691	0.266342628	0.649131319
15	0.167138229	0.448534966	0.615673195	0.515491576	0.339068283	0.854559859
20	0.411450707	0.343137776	0.754588483	0.608010355	0.333068674	0.941079029
25	0.420676373	0.404784936	0.825461309	0.810564011	0.178348672	0.988912682
30	0.427512722	0.447814966	0.875327688	0.840678208	0.154904199	0.995582407
35	0.599332979	0.349613422	0.9489464	0.926931047	0.071448731	0.998379778
40	0.73177715	0.233927178	0.965704328	0.936687125	0.063006989	0.999694114
45	0.720470736	0.264891633	0.985362369	0.971654895	0.028230361	0.999885256
50	0.813943011	0.175963005	0.989906015	0.974912958	0.025066432	0.999979391
55	0.803195324	0.189289588	0.992484913	0.9889189	0.011073208	0.999992108
60	0.869379323	0.127680822	0.997060144	0.995227305	0.004771404	0.999998709
65	0.914954768	0.083918606	0.998873374	0.995637343	0.004362122	0.999999465
70	0.907461037	0.091686865	0.999147902	0.998127592	0.001872322	0.999999914
75	0.939664563	0.060010871	0.999675434	0.999209415	0.000790552	0.999999966
80	0.961180651	0.038587701	0.999768352	0.999260176	0.000739818	0.999999994
85	0.956897384	0.043009146	0.999906529	0.99968765	0.000312348	0.999999998
90	0.972192596	0.027739785	0.999932381	0.999869666	0.000130334	1
95	0.982238438	0.017736041	0.99997448	0.999875876	0.000124124	1
100	0.988751021	0.011239433	0.999990454	0.999948141	5.1859e-05	1
110	0.991846371	0.008150858	0.999997228	0.999991178	8.822e-06	1
120	0.994071557	0.005927639	0.999999196	0.999996459	3.541e-06	1
130	0.997634567	0.002365322	0.99999989	0.99999941	5.9e-07	1
140	0.99907541	0.00092456	0.99999997	0.999999761	2.39e-07	1
150	0.999317022	0.000682969	0.999999991	0.999999961	3.9e-08	1
160	0.999735479	0.00026452	0.999999999	0.999999984	1.6e-08	1
170	0.999803189	0.00019681	1	0.999999997	3e-09	1
180	0.999924278	7.5722e-05	1	0.999999999	1e-09	1
190	0.999943318	5.6682e-05	1	1	0	1
200	0.9999783	2.17e-05	1	1	0	1

Table 2.6: Power for a two-stage design based on O'Brien & Fleming's method where $H_a : p_1 = 0.2$ and $H_a : p_1 = 0.7$ when $H_0 : p_0 = 0.4$.

	$p_1 = 0.2$			$p_1 = 0.7$		
n	Power1	Power2	Overall Power	Power1	Power2	Overall Power
5	0.00032000	0.107433574	0.107753574	0.16807000	0.242127547	0.410197547
10	0.107378381	0.317068049	0.424446429	0.149314251	0.624291142	0.773605393
15	0.167126779	0.448536525	0.615663304	0.296868444	0.619347496	0.91621594
20	0.206084899	0.530698552	0.736783452	0.416370867	0.551943931	0.968314798
25	0.233993526	0.582734124	0.81672765	0.676928128	0.311183511	0.988111639
30	0.427512476	0.497375795	0.924888271	0.730370389	0.264977648	0.995348037
35	0.432841714	0.513743365	0.946585079	0.864953155	0.134238745	0.999191900
40	0.593127136	0.385859518	0.978986654	0.88485335	0.11482653	0.99967988
45	0.587955517	0.396817537	0.984773054	0.94505064	0.054895957	0.999946597
50	0.710667606	0.283454731	0.994122336	0.952236165	0.047742401	0.999978565
55	0.702062628	0.293617678	0.995680306	0.977829638	0.022166861	0.999996499
60	0.793458179	0.204891702	0.998349881	0.980425714	0.019572865	0.99999858
65	0.784586458	0.214187887	0.998774345	0.991073801	0.00892597	0.999999771
70	0.851917329	0.147617461	0.999534791	0.996031943	0.003967966	0.999999909
75	0.900277319	0.099382858	0.999660177	0.996405	0.003594985	0.999999985
80	0.893395586	0.106472747	0.999868333	0.998413834	0.001586164	0.999999998
85	0.928408122	0.071542656	0.999950779	0.998550271	0.001449728	0.999999999
90	0.952624034	0.047339432	0.999963466	0.99936391	0.00063609	1
95	0.948376494	0.051609546	0.999986039	0.999725017	0.000274983	1
100	0.96584837	0.034146458	0.999994828	0.999744167	0.000255833	1
110	0.975286484	0.024712044	0.999998528	0.999952953	4.7047e-05	1
120	0.989500075	0.010499513	0.999999588	0.99998108	1.892e-05	1
130	0.992360406	0.007639537	0.999999943	0.999996653	3.347e-06	1
140	0.996841955	0.003158029	0.999999984	0.999998649	1.351e-06	1
150	0.99768782	0.002312175	0.999999995	0.999999767	2.33e-07	1
160	0.999062038	0.000937961	0.999999999	0.999999906	9.4e-08	1
170	0.999309079	0.000690921	1	0.999999984	1.6e-08	1
180	0.999723515	0.000276484	1	0.999999993	7e-09	1
190	0.999795195	0.000204805	1	0.999999999	1e-09	1
200	0.99991888	8.112e-05	1	1	0	1

Table 2.7: Power for a two-stage design based on Haybittle-Peto's method where $H_a : p_1 = 0.2$ and $H_a : p_1 = 0.7$ when $H_0 : p_0 = 0.4$.

	$p_1 = 0.2$			$p_1 = 0.7$		
n	Power1	Power2	Overall Power	Power1	Power2	Overall Power
5	0.00032000	0.107433574	0.107753574	0.16807000	0.242127547	0.410197547
10	1.02e-07	0.411463992	0.411464094	0.028247525	0.74406974	0.772317265
15	0.035184429	0.572422768	0.607607197	0.126827729	0.788766635	0.915594364
20	0.069175304	0.663207353	0.732382657	0.237507781	0.730578237	0.968086018
25	0.098225225	0.791306383	0.889531609	0.340654904	0.647085516	0.987740421
30	0.122710807	0.800129684	0.922840491	0.431517906	0.566331222	0.997849128
35	0.14349171	0.802022869	0.945514579	0.651555411	0.347608833	0.999164244
40	0.28589137	0.692487390	0.978378760	0.703249067	0.296422381	0.999671448
45	0.297456736	0.686992689	0.984449425	0.746215267	0.253729381	0.999944648
50	0.307331628	0.686613415	0.993945043	0.859440124	0.140537891	0.999978014
55	0.446344441	0.549246716	0.995591157	0.879215047	0.120781332	0.999996379
60	0.448617474	0.54968581	0.998303284	0.93676187	0.063236675	0.999998545
65	0.573501222	0.425857848	0.999359071	0.945199298	0.054800466	0.999999764
70	0.570876965	0.428645826	0.999522791	0.972377193	0.027622769	0.999999962
75	0.675854105	0.323966356	0.999820462	0.975846329	0.024153656	0.999999985
80	0.670750726	0.329114548	0.999865274	0.988150034	0.011849963	0.999999998
85	0.755674148	0.244275304	0.999949452	0.989549243	0.010450756	0.999999999
90	0.749712206	0.250268968	0.999981174	0.994974921	0.005025079	1
95	0.816796804	0.183188903	0.999985707	0.995533554	0.004466446	1
100	0.868646783	0.131347907	0.99999469	0.997885383	0.002114617	1
110	0.902637096	0.097361398	0.999998494	0.999114797	0.000885203	1
120	0.927877901	0.072121894	0.999999795	0.999630795	0.000369205	1
130	0.946588062	0.053411879	0.999999942	0.999924233	7.5767e-05	1
140	0.974601457	0.025398526	0.999999983	0.999968841	3.1159e-05	1
150	0.981303166	0.018696832	0.999999998	0.999987187	1.2813e-05	1
160	0.986222402	0.013777598	0.999999999	0.999997526	2.474e-06	1
170	0.993747312	0.006252688	1	0.999998988	1.012e-06	1
180	0.995397389	0.004602611	1	0.999999585	4.15e-07	1
190	0.996608113	0.003391887	1	0.999999923	7.7e-08	1
200	0.998506447	0.001493553	1	0.999999968	3.2e-08	1

2.2.2 Three-stage design

Consider three-stage RCT with alpha spending functions based on Pocock, O'Brien & Fleming, and Haybittle-Peto methods. To obtain a 0.05 confidence level ($\alpha = 0.05$), the following alpha spending values are used as the predicted type I error rate of each stage.

Table 2.8: Expected alpha spending values for each stage of Pocock, O'Brien & Fleming, and Haybittle-Peto designs.

$\alpha = 0.05$			
	Stage 1	Stage 2	Stage 3
Pocock	$\alpha_1 = 0.0167$	$\alpha_2 = 0.0167$	$\alpha_3 = 0.0167$
OBF	$\alpha_1 = 0.004$	$\alpha_2 = 0.014$	$\alpha_3 = 0.032$
Peto	$\alpha_1 = 0.001$	$\alpha_2 = 0.001$	$\alpha_3 = 0.048$

Using the alpha spending values shown in Table 2.8, critical values and type I error rates have been calculated for each stage of a three-stage RCT along with the overall type I error rate for various sample sizes considering the following hypotheses.

$$H_0 : p = 0.4 \qquad vs \qquad H_a : p \neq 0.4 \tag{2.45}$$

A R software programme for a three-stage group sequential design was developed to calculate the critical values (cv_1, cv_2, cv_3) and the associated type I error rate at each stage (e_1, e_2, e_3) along with the overall type I error rate. The results are given in Tables 2.9 , 2.10 and 2.11 for various sample sizes with same sample size (n) increment at each stage with $p_0 = 0.4$ and $\alpha = 0.05$.

Table 2.9: Critical values and type I error rates for 3-stage design with equally spent alpha (Pocock method).

n	cv1	cv2	cv3	e1	e2	e3	Total Error rate
5	9.1629	6.6959	6.8475	0.0102400	0.01509059	0.00763724	0.03296782
10	10.2165	7.3515	6.6277	0.0077243	0.00842396	0.00920424	0.02535254
15	6.8475	6.6277	5.7686	0.0145197	0.01035335	0.01374960	0.03862264
20	7.3515	6.4874	5.4831	0.0100774	0.01150513	0.01624512	0.0378276
25	6.8015	5.7057	5.4226	0.0137972	0.01645368	0.01476799	0.04501889
30	6.6277	5.9828	5.4737	0.0139604	0.01374479	0.01448306	0.04218823
35	6.4061	5.8052	5.5884	0.0154286	0.01533145	0.01243456	0.04319457
40	6.4874	6.1490	5.1842	0.0144064	0.01211054	0.01661025	0.04312720
45	6.4101	5.8703	5.3992	0.0146728	0.01365699	0.01355314	0.04188297
50	6.5703	5.8641	5.3390	0.0133051	0.01441341	0.01483982	0.04255834
55	6.5792	5.6794	5.0978	0.0130092	0.01538607	0.01616845	0.04456376
60	6.7629	5.7429	5.3610	0.0116486	0.01565124	0.01355657	0.04085639
65	6.2470	5.6162	5.3734	0.0154850	0.01536876	0.01372802	0.04458181
70	6.3057	5.7151	5.2128	0.0145402	0.01519428	0.01465372	0.04438816
75	6.5542	5.6282	5.2628	0.0127904	0.01561678	0.01494942	0.04335657
80	6.1490	5.7456	5.1381	0.0160085	0.01438035	0.01541771	0.04580657
85	6.2384	5.6882	5.2105	0.0147480	0.01446906	0.01550095	0.04471799
90	6.5071	5.8150	5.8150	0.0128980	0.01420011	0.01602344	0.04312154
95	6.1687	5.5129	5.1980	0.0155631	0.01658967	0.01475747	0.04691019
100	6.2805	5.4820	5.1195	0.0141932	0.01639444	0.01499385	0.04558147
110	6.2578	5.6197	5.1515	0.0145972	0.01495085	0.01496034	0.04450842
120	6.1078	5.5319	5.0349	0.0153359	0.01621753	0.01655778	0.04811116
130	6.1486	5.6914	5.1026	0.0153223	0.01451355	0.01615155	0.04598744
140	6.0482	5.4948	5.1804	0.0156579	0.01569825	0.01488036	0.04623653
150	6.1233	5.4716	4.9786	0.0153762	0.01657262	0.01620150	0.04815032
160	6.0583	5.6565	5.0801	0.0154338	0.01467831	0.01552507	0.04563716
170	6.0583	5.6565	5.0801	0.0154338	0.01467831	0.01467831	0.04563716
180	6.1141	5.5229	5.1557	0.0148622	0.01600635	0.01491946	0.04578803
190	6.2185	5.7193	5.0129	0.0143286	0.01413455	0.01650553	0.04496869
200	6.2017	5.6141	5.1345	0.0140818	0.01465133	0.01495221	0.04368536

Table 2.10: Critical values and type I error rates for 3-stage design based on O'Brien & Fleming' method.

n	cv1	cv2	cv3	e1	e2	e3	Total Error rate
5	NA	10.2165	5.1664	NA	0.006833050	0.030110763	0.036943813
10	11.0132	7.3515	4.8656	0.001677722	0.009435496	0.031423609	0.042536827
15	10.0439	6.6277	4.8411	0.002397954	0.012956156	0.02559735	0.04095146
20	10.1036	6.4874	4.6841	0.002135574	0.013464699	0.026920946	0.042521219
25	9.6278	6.5703	4.7104	0.003572209	0.012030635	0.025981442	0.041584287
30	10.3328	6.7629	4.8251	0.002366466	0.010755227	0.024469587	0.03759128
35	9.4053	6.3057	4.7013	0.002821010	0.013425370	0.027028106	0.043274487
40	9.4444	6.6567	4.8986	0.003274851	0.010691123	0.024273157	0.038239132
45	9.0134	6.5071	4.5856	0.003386938	0.011608433	0.027025193	0.042020563
50	9.1516	6.2805	4.6081	0.003570906	0.012798045	0.027899293	0.044268244
55	8.9202	6.2578	4.3844	0.003478354	0.013199718	0.029711826	0.046389897
60	9.1224	6.1078	4.4599	0.003483537	0.013911137	0.029927489	0.047322163
65	8.9841	6.1486	4.2904	0.003292699	0.013948038	0.031065185	0.048305922
70	9.2343	6.5546	4.5728	0.003196654	0.010910469	0.026891983	0.040999105
75	9.1409	6.4803	4.6654	0.002971623	0.011211225	0.026752271	0.040935119
80	9.4308	6.5315	4.5439	0.002826403	0.011047608	0.027269333	0.041143344
85	8.8262	6.1523	4.2702	0.003812472	0.013494571	0.031168909	0.048475952
90	8.842	6.1141	4.3955	0.003453891	0.013485678	0.030546050	0.047485619
95	9.1297	6.2185	4.3081	0.003209505	0.013036195	0.030609972	0.046855672
100	9.1385	6.2017	4.4375	0.002899928	0.012900127	0.029874981	0.045675036
110	8.9969	6.3122	4.4969	0.003369227	0.011884929	0.028853997	0.044108154
120	8.6761	6.1594	4.2494	0.00370211	0.013140161	0.031417519	0.048259789
130	9.0432	6.3022	4.344	0.002998137	0.012200854	0.030000466	0.045199458
140	9.0097	6.097	4.3227	0.00325276	0.01332333	0.031423792	0.047999881
150	8.7962	6.2703	4.4289	0.003408039	0.011981373	0.029655675	0.045045088
160	8.7996	6.2046	4.2624	0.003580038	0.01262802	0.031142052	0.04735011
170	8.6557	6.0595	4.3835	0.003656739	0.013389787	0.028250724	0.04529725
180	8.6865	6.0332	4.3964	0.00375733	0.013861392	0.029094227	0.046712949
190	8.5882	6.2244	4.2661	0.003772122	0.012306499	0.031036117	0.047114739
200	8.6403	6.1186	4.2959	0.003815046	0.012791879	0.031563348	0.048170273

Table 2.11: Critical values and type I error rates of 3-stage design based on Haybittle-Peto method.

n	cv1	cv2	cv3	e1	e2	e3	Total Error rate
5	NA	11.0132	5.1664	NA	0.000786432	0.034237942	0.035024374
10	18.3258	13.3033	4.8656	0.000104858	0.000823649	0.037633234	0.038561741
15	14.0866	13.5773	4.4210	0.000749089	0.000463364	0.045810833	0.047023287
20	13.3033	11.9149	4.3380	0.000841080	0.000880006	0.045886345	0.047607431
25	13.2247	11.6905	4.3981	0.000710013	0.000663250	0.043356220	0.044729482
30	13.5773	11.4228	4.5294	0.000535595	0.000772349	0.039311010	0.040618954
35	11.646	11.3530	4.1057	0.000807917	0.000757834	0.045642398	0.047208149
40	11.9149	11.4049	4.3400	0.000997217	0.000718131	0.039501117	0.041216465
45	12.7505	11.5371	4.3612	0.000641500	0.000683658	0.041938339	0.043263497
50	11.6905	11.4115	4.1115	0.000749824	0.000888336	0.044548430	0.046186590
55	12.1448	11.6738	4.1944	0.000800537	0.000750218	0.045509243	0.047059998
60	11.4228	10.9662	4.0084	0.000852171	0.000904485	0.046896347	0.048653004
65	11.8971	11.2817	4.1210	0.000861862	0.000751079	0.046982443	0.048595385
70	11.353	11.3378	3.9759	0.000872350	0.000858681	0.047553604	0.049284635
75	11.8478	10.8451	4.1037	0.000852625	0.000960937	0.046858743	0.048672305
80	11.4049	11.2116	3.9873	0.000837578	0.000770340	0.047121313	0.048729231
85	11.9204	11.3342	4.1228	0.000799906	0.000839402	0.046110629	0.047749937
90	11.5371	10.9474	4.0276	0.000771083	0.000899811	0.045760622	0.047431515
95	12.0737	11.0872	4.1667	0.000724311	0.000953702	0.044558397	0.046236411
100	11.4115	10.7772	4.0881	0.000999172	0.000968746	0.043848450	0.045816368
110	11.6738	11.3588	4.0383	0.000861009	0.000796101	0.047857349	0.049514460
120	11.9656	11.2720	4.1260	0.000732360	0.000838604	0.045217255	0.046788219
130	11.2817	11.2324	4.2213	0.000865837	0.000838858	0.042466789	0.044171485
140	11.3378	11.2298	4.0284	0.000982373	0.000821697	0.045180074	0.046984144
150	11.7103	11.2564	4.0373	0.000804851	0.000819712	0.047200679	0.048825242
160	11.2116	10.7428	4.1548	0.000889424	0.000975090	0.043670212	0.045534726
170	11.3342	10.824	4.0079	0.000958722	0.000923261	0.045470798	0.047352781
180	11.7405	10.9191	4.0393	0.000773244	0.000888723	0.046742942	0.048404909
190	11.3368	11.0258	4.1671	0.000823008	0.000827699	0.043205688	0.044856395
200	11.5077	10.964	4.0498	0.000859932	0.000933609	0.044243317	0.046036857

From Tables 2.9, 2.10 and 2.11 we see that as the sample size increases, the critical value of each stage converges. Similarly, when we examine the values of type I error rate and total error rate, we can observe that these values converge as the sample size grows. When the sample size increases, it is evident that e1, e2, and e3 values rise to achieve the expected alpha spending at each stage. Accordingly, the total error rate may increase to achieve the expected level of statistical significance (α).

Let us consider the output related to three-stage design based on Pocock's method as an example. Considering the convergence of data, we may conclude that $cv_1 \approx 6.1$, $cv_2 \approx 5.7$ and $cv_3 \approx 5.0$ for large samples. These outcomes are evident in convergence plots given in Figures 2.2, 2.3 and 2.4.

All three stages converge to an approximate type I error rate of 0.015. Since these observed error rate values are selected to be less than the alpha spending at each stage (0.05/3 =0.0167), the total error rate is always less than the expected significance level ($\alpha = 0.05$). These outcomes are clearly visible in convergence plots given in Figures 2.5, 2.6, 2.7 and 2.8.

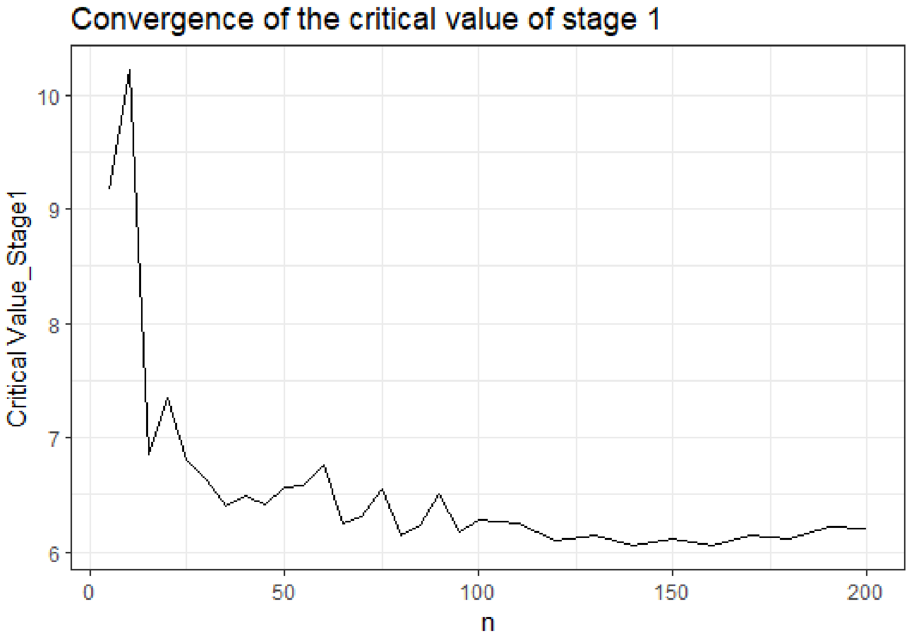

Figure 2.2: Convergence of the critical value of stage 1 for various sample sizes (Pocock method).

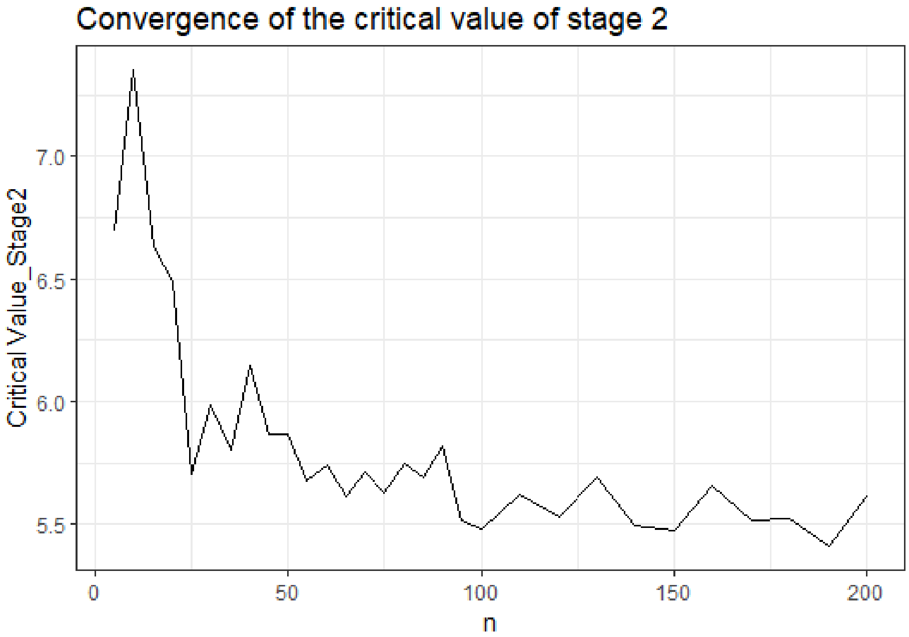

Figure 2.3: Convergence of the critical value of stage 2 for various sample sizes (Pocock method).

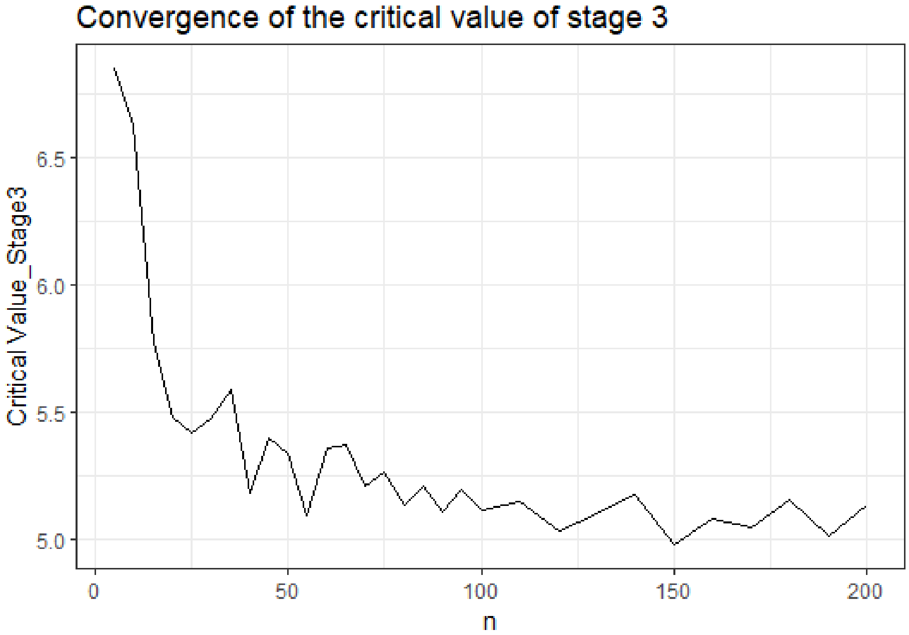

Figure 2.4: Convergence of the critical value of stage 3 for various sample sizes (Pocock method).

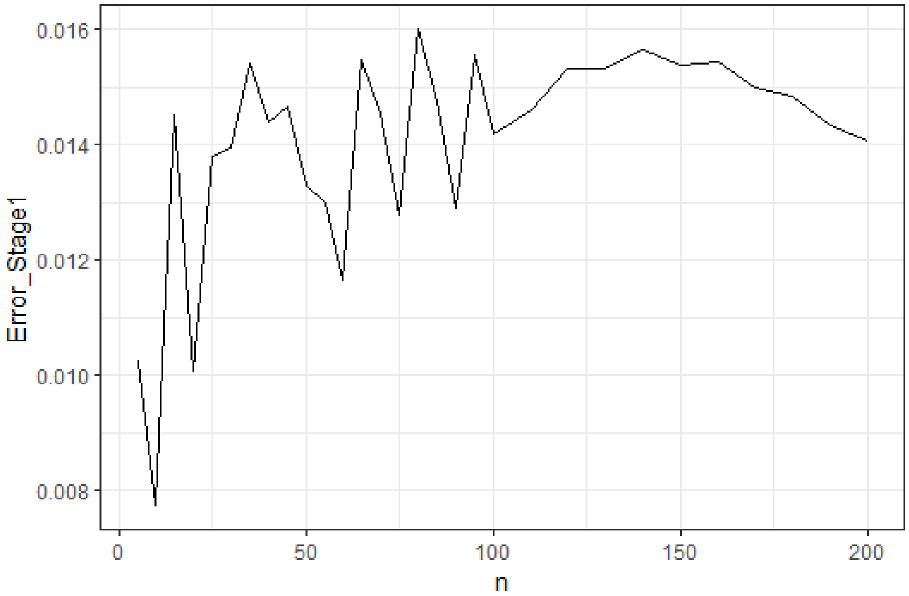

Figure 2.5: Observed type I error rate of stage 1 for various sample sizes(Pocock method).

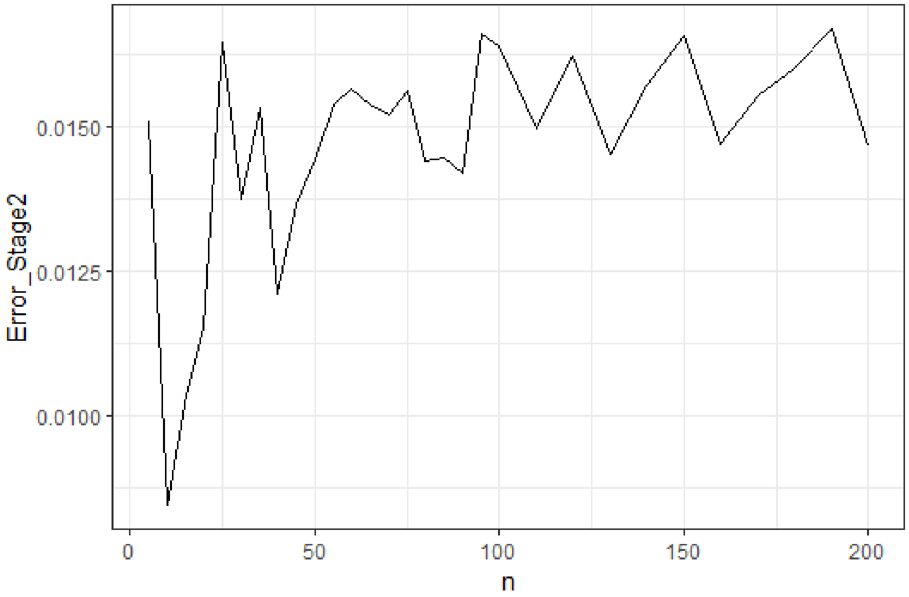

Figure 2.6: Observed type I error rate of stage 2 for various sample sizes(Pocock method).

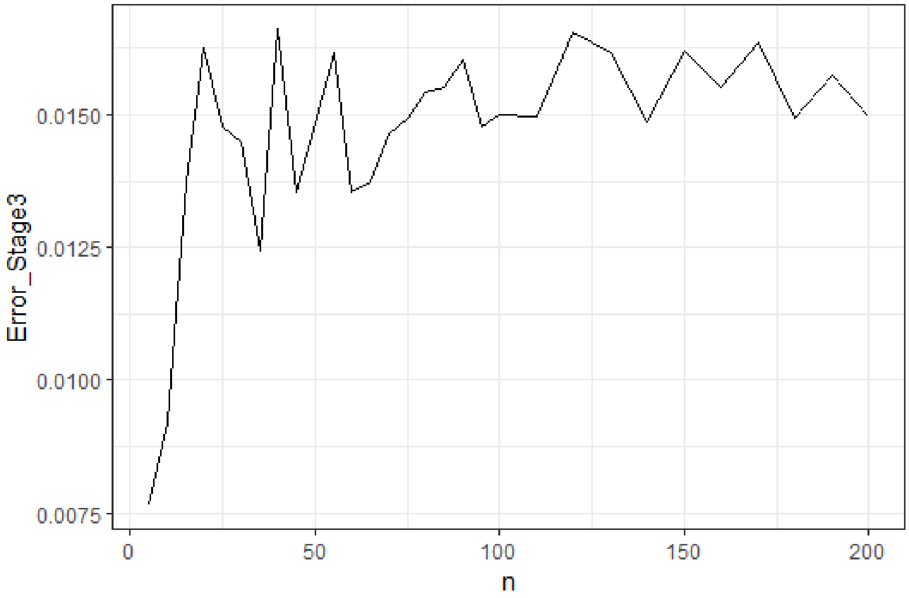

Figure 2.7: Observed type I error rate of stage 3 for various sample sizes(Pocock method).

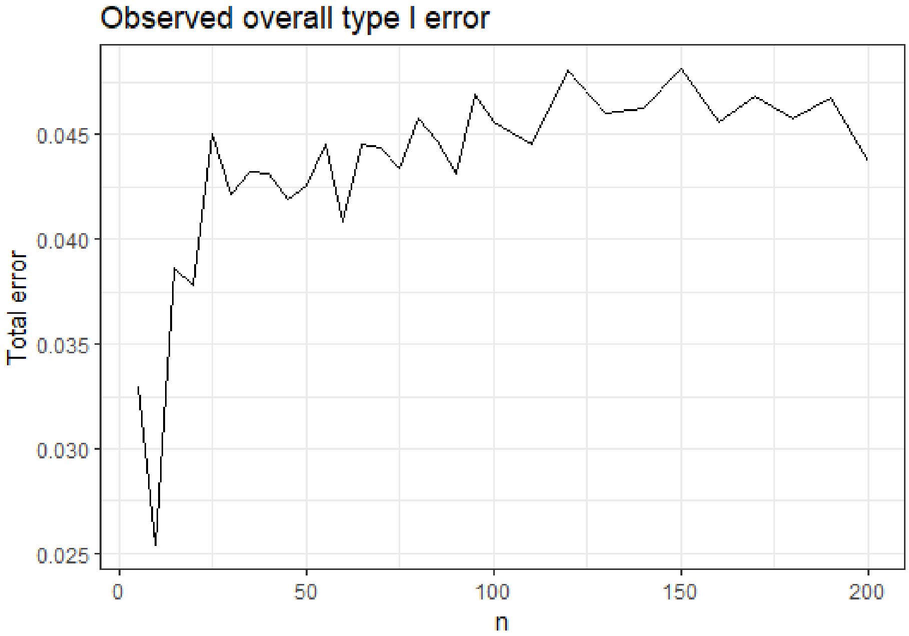

Figure 2.8: Observed Overall type I error rate for various sample sizes(Pocock method).

Once the critical value and type I error rate for a particular value of p_0 have been identified, the power can be calculated for various alternative hypotheses (p_1). Tables 2.12, 2.13 and 2.14 present the statistical power for two different alternative hypotheses; $p_1 = 0.2$ and $p_1 = 0.7$ for a three-stage design with $p_0 = 0.4$.

Table 2.12: Power when $H_a : p_1 = 0.2$ and $H_a : p_1 = 0.7$ where $H_0 : p_0 = 0.4$ based on Pocock's method

	p1=0.2				p1=0.7			
n	Power1	Power2	Power3	Overall Power	Power 1	Power 2	Power 3	Overall_Power
5	0.00032	0.1074336	0.08796749	0.195721090	0.16807	0.2421275	0.1620324	0.5722299
10	0.1073784	0.1333082	0.22142950	0.462116100	0.1493143	0.4647328	0.2390519	0.8530990
15	0.1671382	0.2842042	0.28873810	0.740080500	0.5154916	0.3390683	0.1206092	0.9751690
20	0.2060866	0.4008794	0.27143070	0.878396700	0.6080098	0.3330687	0.05468934	0.9957679
25	0.4206746	0.4047849	0.12428120	0.949740700	0.6769282	0.2993568	0.02300551	0.9992904
30	0.4275127	0.447815	0.10088990	0.976217600	0.8406782	0.1504924	0.008714996	0.9998856
35	0.5993328	0.3169737	0.07289585	0.989202350	0.8649532	0.1332823	0.001746757	0.9999822
40	0.5931272	0.3455385	0.05840378	0.997069480	0.9366871	0.06265049	0.000659469	0.9999971
45	0.7204707	0.2541576	0.02412292	0.998751220	0.9450506	0.05467939	0.00026947	0.9999995
50	0.7106676	0.2706346	0.01810646	0.999408660	0.974913	0.02504006	4.70E-05	1
55	0.8031953	0.1892896	0.007376292	0.999861192	0.9778296	0.02215117	1.92E-05	1
60	0.7934582	0.2009679	0.005508123	0.999934223	0.9900391	0.009957637	0.009957637	1
65	0.8603561	0.1374424	0.002173129	0.999971629	0.9956373	0.004361408	1.25E-06	1
70	0.907461	0.09098629	0.001545675	0.999992965	0.9960319	0.003967836	2.21E-07	1
75	0.9002773	0.0990814	0.000638148	0.999996848	0.9982721	0.001727844	8.63E-08	1
80	0.9339744	0.06556674	0.000458100	0.999999240	0.9992602	0.00073981	1.41E-08	1
85	0.9568974	0.04292534	0.000176936	0.999999676	0.9993122	0.000687843	5.90E-09	1
90	0.952624	0.0472411	0.000134781	0.999999881	0.999706	0.000293994	9.57E-10	1
95	0.969038	0.03091017	5.18E-05	0.999999947	0.9998759	0.000124124	1.53E-10	1
100	0.9799798	0.02000051	1.97E-05	0.999999996	0.9998826	0.000117378	6.46E-11	1
110	0.9855264	0.01446786	5.73E-06	1	0.9999793	2.07E-05	4.12E-12	1
120	0.9940716	0.005926857	1.59E-06	1	0.9999917	8.32E-06	1.09E-13	1
130	0.995678	0.004321512	4.64E-07	1	0.9999986	1.43E-06	6.94E-15	1
140	0.9982647	0.001735187	6.48E-08	1	0.9999994	5.76E-07	4.66E-16	1
150	0.9987251	0.001274883	1.89E-08	1	0.9999999	9.68E-08	1.13E-17	1
160	0.9994951	0.000504875	5.26E-09	1	1	3.93E-08	7.58E-19	1
170	0.9996265	0.000373492	7.56E-10	1	1	6.52E-09	4.85E-20	1
180	0.9998536	0.000146445	2.11E-10	1	1	2.66E-09	1.21E-21	1
190	0.999891	0.000108984	6.20E-11	1	1	4.37E-10	7.78E-23	1
200	0.9999576	4.24E-05	8.38E-12	1	1	1.79E-10	5.23E-24	1

Table 2.13: Power when $H_a : p_1 = 0.2$ and $H_a : p_1 = 0.7$ where $H_0 : p_0 = 0.4$ based on O'Brien & Fleming's method

p1=0.2					p1=0.7			
n	Power1	Power2	Power3	Overall Power	Power 1	Power 2	Power 3	Overall_Power
5	0.00032	0.10737623	0.296878025	0.404574255	0.16807	0.060536315	0.31401713	0.542623445
10	4.198e-06	0.206086417	0.405753635	0.61184425	0.149308346	0.464732811	0.304288686	0.918329843
15	0.035185383	0.394476385	0.400191218	0.829852986	0.296867942	0.546757499	0.130573346	0.974198788
20	0.069175471	0.525919853	0.329816038	0.924911362	0.416370831	0.521422041	0.057864968	0.99565784
25	0.233993288	0.484389488	0.24836118	0.966743957	0.511848547	0.463446873	0.023982489	0.999277909
30	0.255233259	0.542478533	0.187260501	0.984972293	0.588808685	0.401358475	0.009713139	0.999880299
35	0.272091699	0.636214857	0.085138631	0.993445187	0.772925378	0.223241156	0.003824957	0.999991491
40	0.437145899	0.498492543	0.061369528	0.99700797	0.807448245	0.191008047	0.001542316	0.999998609
45	0.440716407	0.512803141	0.045737826	0.999257374	0.901367112	0.098349206	0.000283466	0.999999783
50	0.583559419	0.396947008	0.019166565	0.999672991	0.915197401	0.084688453	0.000114131	0.999999985
55	0.579762625	0.406049959	0.014108569	0.999921153	0.95842343	0.041556548	2.002e-05	0.999999998
60	0.694430604	0.299793135	0.005741206	0.999964945	0.963762481	0.036229426	8.093e-06	1
65	0.68760135	0.308160665	0.004229688	0.999991702	0.982727862	0.017270757	1.38e-06	1
70	0.776461726	0.22049076	0.003043656	0.999996142	0.984764398	0.015235041	5.61e-07	1
75	0.768986505	0.229762133	0.001249608	0.999998246	0.992878313	0.007121464	2.23e-07	1
80	0.836585246	0.162506661	0.000907685	0.999999592	0.993653369	0.006346593	3.8e-08	1
85	0.886804757	0.112835544	0.000359608	0.999999909	0.99707514	0.002924853	6e-09	1
90	0.880492068	0.119364478	0.000143413	0.999999958	0.998679719	0.001320279	3e-09	1
95	0.91767937	0.082215231	0.000105389	0.99999999	0.998801205	0.001198795	0	1
100	0.912524615	0.087433756	4.1625e-05	0.999999996	0.999462985	0.000537015	0	1
110	0.959414593	0.040573533	1.1874e-05	1	0.999781434	0.000218566	0	1
120	0.970424737	0.029571813	3.45e-06	1	0.999958303	4.1697e-05	0	1
130	0.978400758	0.021598239	1.002e-06	1	0.999983112	1.6888e-05	0	1
140	0.99046784	0.009532016	1.44e-07	1	0.999993147	6.853e-06	0	1
150	0.993029126	0.006970832	4.2e-08	1	0.99999875	1.25e-06	0	1
160	0.997016402	0.002983587	1.2e-08	1	0.999999493	5.07e-07	0	1
170	0.997811029	0.00218897	2e-09	1	0.99999991	9e-08	0	1
180	0.999083606	0.000916394	0	1	0.999999963	3.7e-08	0	1
190	0.999325042	0.000674957	0	1	0.999999994	6e-09	0	1
200	0.999722092	0.000277908	0	1	0.999999997	3e-09	0	1

From Tables 2.12, 2.13 and 2.14, we see that the overall power increases as the sample size grows. When the value of p_1 is far from p_0, the overall power for a given sample size is greater than when p_1 is close to p_0. Also, it is clearly noticeable that the overall power converges as the sample size increases. It converges rapidly when p_1 is considerably away from p_0.

Table 2.14: Power when $H_a : p_1 = 0.2$ and $H_a : p_1 = 0.7$ where $H_0 : p_0 = 0.4$ based on Haybittle-Peto's method

	p1=0.2				p1=0.7			
n	Power1	Power2	Power3	Overall Power	Power 1	Power 2	Power 3	Overall_Power
5	0.00032	2.048e-06	0.398033095	0.398355143	0.16807	0.06053041	0.314018093	0.542618503
10	1.02e-07	0.069175303	0.538173595	0.607349001	0.028247525	0.213504964	0.673831905	0.915584394
15	0.035184429	0.09990563	0.691182367	0.826272426	0.126827729	0.31829637	0.54137043	0.986494529
20	0.069175304	0.229002229	0.624836078	0.923013611	0.237507781	0.473376188	0.286967779	0.997851749
25	0.098225225	0.227441009	0.63987044	0.965536675	0.340654904	0.522087206	0.136914393	0.999656504
30	0.122710807	0.338315412	0.523431331	0.98445755	0.431517906	0.506519614	0.061907199	0.999944719
35	0.14349171	0.435242272	0.417524498	0.99625848	0.651555411	0.322195672	0.026239993	0.999991076
40	0.28589137	0.400556429	0.311859562	0.998307361	0.703249067	0.285370003	0.011379477	0.999998547
45	0.297456736	0.461404205	0.240366048	0.999226989	0.746215267	0.248917789	0.004866849	0.999999906
50	0.307331628	0.563719642	0.12876922	0.999820489	0.859440124	0.138570258	0.001989603	0.999999985
55	0.446344441	0.460021072	0.093552024	0.999917536	0.879215047	0.119941215	0.000843737	0.999999999
60	0.448617474	0.481366252	0.069997467	0.999981192	0.93676187	0.063064158	0.000173972	1
65	0.573501222	0.37589867	0.050591398	0.99999129	0.945199298	0.054728275	7.2428e-05	1
70	0.570876965	0.404447717	0.024673357	0.999998038	0.972377193	0.027593831	2.8975e-05	1
75	0.675854105	0.30635805	0.01778693	0.999999085	0.975846329	0.024147866	5.805e-06	1
80	0.670750726	0.315991048	0.013258021	0.999999795	0.988150034	0.011847648	2.318e-06	1
85	0.755674148	0.238348997	0.005976759	0.999999904	0.989549243	0.010449796	9.61e-07	1
90	0.749712206	0.24584357	0.004444203	0.999999979	0.994974921	0.005024901	1.78e-07	1
95	0.816796804	0.181256423	0.001946762	0.99999999	0.995533554	0.004466372	7.4e-08	1
100	0.868646783	0.129948638	0.001404577	0.999999998	0.997885383	0.002114604	1.3e-08	1
110	0.902637096	0.096913178	0.000449725	1	0.999114797	0.000885201	2e-09	1
120	0.927877901	0.071980432	0.000141667	1	0.999630795	0.000369205	0	1
130	0.946588062	0.053367864	4.4074e-05	1	0.999924233	7.5767e-05	0	1
140	0.974601457	0.02538532	1.3223e-05	1	0.999968841	3.1159e-05	0	1
150	0.981303166	0.018692777	4.058e-06	1	0.999987187	1.2813e-05	0	1
160	0.986222402	0.013776361	1.237e-06	1	0.999997526	2.474e-06	0	1
170	0.993747312	0.006252323	3.65e-07	1	0.999998988	1.012e-06	0	1
180	0.995397389	0.004602501	1.11e-07	1	0.999999585	4.15e-07	0	1
190	0.996608113	0.003391853	3.3e-08	1	0.999999923	7.7e-08	0	1
200	0.998506447	0.001493547	5e-09	1	0.999999968	3.2e-08	0	1

In addition, it is evident that when the power > 85%, a greater proportion of total power is achieved in the first stage, and a substantial component of the remainder is achieved in the second stage, leaving just a small portion for the third stage.

Chapter 3

Sequential designs of randomized clinical trials to compare two groups proportions

3.1 Markov chain modelling for two populations

In Chapter 2 we considered the group sequential analysis for testing the proportion for one sample case. In this chapter we focus on sequential designs of RCT to compare proportions for two samples case. The methodology for the Markov chain approach in Chapter 2 is extended for two groups case.

Consider a testing of hypothesis for comparing two proportions. A null hypothesis with equal probabilities of success is tested against a two-sided alternative hypothesis.

$$H_0 : p_A = p_B, \qquad vs \qquad H_a : p_A \neq p_B. \tag{3.1}$$

Let,
n_A be the sample size of group A
n_B be the sample size of group B
s_A be the number of successes in group A and
s_B be the number of successes in group B

For testing the null hypothesis, we construct the log-likelihood ratio as,

$$LLR = 2ln\left[\frac{L(\hat{p}|H_a)}{(L(p_0|H_0)}\right] = 2[ln(L(\hat{p}|H_a)) - ln(L(p_0|H_0))]. \tag{3.2}$$

Under the null hypothesis (H_0), the probability of success of two groups are equal, so an estimate of p_0 is given by,

$$\hat{p_0} = \frac{s_A + s_B}{n_A + n_B}. \tag{3.3}$$

Under the alternative hypothesis $p_A \neq p_B$, the estimated values of p_A and p_B are,

$$\hat{p_A} = \frac{s_A}{n_A}, \qquad and \qquad \hat{p_B} = \frac{s_B}{n_B}. \tag{3.4}$$

Group sequential analysis starts with an unconditional initial stage. The process moves to the second stage if the test is not terminated at the first interim analysis. The second stage is conditional on the first stage. Since the testing procedure at any stage depends on the previous stage, the procedure is explained stage-wise.

First Stage

The first stage of the analysis starts with n_{A1} participants in group A and n_{B1} participants in group B. Assume that there are s_{A1} number of successes out of n_{A1} and s_{B1} number of successes out of n_{B1}.

The data the likelihood under the null hypothesis is,

$$L(p_0|H_0) = \binom{n_{A1}}{s_{A1}}(\hat{p_0})^{s_{A1}}(1-\hat{p_0})^{(n_{A1}-s_{A1})} * \binom{n_{B1}}{s_{B1}}(\hat{p_0})^{s_{B1}}(1-\hat{p_0})^{(n_{B1}-s_{B1})}. \tag{3.5}$$

and the likelihood under the alternative hypothesis is,

$$L(p_A, p_B|H_a) = \binom{n_{A1}}{s_{A1}}(\hat{p_A})^{s_{A1}}(1-\hat{p_A})^{(n_{A1}-s_{A1})} * \binom{n_{B1}}{s_{B1}}(\hat{p_B})^{s_{B1}}(1-\hat{p_B})^{(n_{B1}-s_{B1})}. \tag{3.6}$$

The log-likelihood ratio statistic for testing the null hypothesis is,

$$\begin{aligned} LLR_1 = 2\Bigg\{ & s_{A1}ln(s_{A1}) + s_{B1}ln(s_{B1}) + f_{A1}ln(f_{A1}) + f_{B1}ln(f_{B1}) - n_{A1}ln(n_{A1}) \\ & -n_{B1}ln(n_{B1}) - (s_{A1}+s_{B1})ln(s_{A1}+s_{B1}) - (f_{A1}+f_{B1})ln(f_{A1}+f_{B1}) \\ & +(n_{A1}+n_{B1})ln(n_{A1}+n_{B1})\Bigg\}, \end{aligned} \tag{3.7}$$

where f_{A1} and f_{B1} denotes the number of failures in group A and group B, respectively as,

$$f_{A1} = n_{A1} - s_{A1}, \qquad and \qquad f_{B1} = n_{B1} - s_{B1}. \tag{3.8}$$

For fixed n_{A1} and n_{B1}, the value of log-likelihood ratio depends on s_{A1} and s_{B1}. In order to determine all possible LLR values, we consider all possible combinations of s_{A1} and s_{B1}. For given n_{A1} and n_{B1} values; s_{A1} can have any value from zero to n_{A1} and s_{B1} can have any value from zero to n_{B1}.

$$s_{A1} = 0, 1, 2, \ldots, n_{A1}, \qquad and \qquad s_{B1} = 0, 1, 2, \ldots, n_{B1}. \tag{3.9}$$

Considering all possible combinations of s_{A1} and s_{B1}, we can have $(n_{A1}+1)(n_{B1}+1)$ number of LLR values. Let M_1 be the maximum number of possible LLR values at the first stage. As the first stage is unconditional, we know that

$$M_1 = (n_{A1}+1)(n_{B1}+1). \tag{3.10}$$

Therefore, all possible test statistics values of the first stage are as follows,

$$LLR_1 = T^{1st} = (T_1^{1st}, T_2^{1st}, \ldots, T_{M_1}^{1st}). \tag{3.11}$$

and the ordered LLR values are denoted as,

$$T_{(1)}^{1st}, T_{(2)}^{1st}, \ldots, T_{(M_1)}^{1st}.$$

The probability of getting each LLR value depends on the probability of getting s_{A1} number of successes in group A and s_{B1} number of successes in group B, given the probability of success under H_0 as p_0. Thus, for a specific pair of success probability values (s_{A1}, s_{B1}), the probability of the corresponding LLR value can be calculated as follows.

$$\begin{aligned}Pr[LLR(s_{A1}, s_{B1}|n_A, n_B, p_0)] &= Pr(s_{A1}|n_A, p_0) * Pr(s_{B1}|n_B, p_0)\\ &= \binom{n_{A1}}{s_{A1}}(p_0)^{s_{A1}}(1-p_0)^{(n_{A1}-s_{A1})} * \binom{n_{B1}}{s_{B1}}(p_0)^{s_{B1}}(1-p_0)^{(n_{B1}-s_{B1})}.\end{aligned} \tag{3.12}$$

After computing all possible LLR values and their probabilities, we can determine the critical value of the first stage (cv_1).

Let,

$$c_1 = min\{j \in \mathbb{N} : Pr(T^{1st} \geq T^{1st}_{(j)}|p_A = p_B) \leq \alpha_1\}, \tag{3.13}$$

where α_1 is the amount of alpha spending for the first stage. Then, $T^{1st}_{(c_1)}$ becomes the critical value of the first stage.

$$cv_1 = T^{1st}_{(c_1)}. \tag{3.14}$$

Second Stage

Failure to reject the null hypothesis at the first stage, will tend to drawing a new sample of the same size from each group and the analysis will move on to the second stage of the process.

Let n_{A2} and n_{B2} be the size of the samples added to the existing sample at stage 2 from group A and group B, respectively. Then,

N_{A2}-Total number of data points (size of the combined sample) of group A at stage 2

$$N_{A2} = n_{A1} + n_{A2}. \tag{3.15}$$

N_{B2}-Total number of data points (size of the combined sample) of group B at stage 2

$$N_{B2} = n_{B1} + n_{B2}. \tag{3.16}$$

All possible (s_{A1}, s_{B1}) pairs, which are not in the rejection region of the first stage, are considered in the second stage of the analysis.

In the second stage, s_{A2} and s_{B2} can take the following values.

$$s_{A2} = 0, 1, 2, \ldots, n_{A2}, \qquad \textit{and} \qquad s_{B2} = 0, 1, 2, \ldots, n_{B2}. \tag{3.17}$$

Any given pair of s_{A1} and s_{B1} values which is in the continuation region of the first stage, say (s^*_{A1}, s^*_{B1}), can create $(n_{A2} + 1)(n_{B2} + 1)$ number of combinations uniting with the data in second stage.

Table 3.1: Creating combinations of stage 2 using one pair (s^*_{A1}, s^*_{B1}) from stage 1.

$\mathrm{s}_{A1} + s_{A2}$	$\mathrm{s}_{B1} + s_{B2}$	Combinations of stage2
$\mathrm{s}^*_{A1} + 0$	$\mathrm{s}^*_{B1} + 0$	$(\ \mathrm{s}^*_{A1} + 0, s^*_{B1} + 0)$
	$\mathrm{s}^*_{B1} + 1$	$(\ \mathrm{s}^*_{A1} + 0, s^*_{B1} + 1)$
	.	.
	.	.
	$\mathrm{s}^*_{B1} + n_{B2}$	$(\ \mathrm{s}^*_{A1} + 0, s^*_{B1} + n_{B2})$
$\mathrm{s}^*_{A1} + 1$	$\mathrm{s}^*_{B1} + 0$	$(\ \mathrm{s}^*_{A1} + 1, s^*_{B1} + 0)$
	$\mathrm{s}^*_{B1} + 1$	$(\ \mathrm{s}^*_{A1} + 1, s^*_{B1} + 1)$
	.	.
	.	.
	$\mathrm{s}^*_{B1} + n_{B2}$	$(\ \mathrm{s}^*_{A1} + 1, s^*_{B1} + n_{B2})$
	.	.
	.	.
	.	.
$\mathrm{s}^*_{A1} + n_{A2}$	$\mathrm{s}^*_{B1} + 0$	$(\ \mathrm{s}^*_{A1} + n_{A2}, s^*_{B1} + 0)$
	$\mathrm{s}^*_{B1} + 1$	$(\ \mathrm{s}^*_{A1} + n_{A2}, s^*_{B1} + 1)$
	.	.
	.	.
	$\mathrm{s}^*_{B1} + n_{B2}$	$(\ \mathrm{s}^*_{A1} + n_{A2}, s^*_{B1} + n_{B2})$

Table 3.1 shows the combinations of the second stage using one pair from the continuation region of stage one. The number of pairs for stage two will depend on the number of pairs in the continuation region of stage one. If there's 'm_1' number of pairs in the continuation region, then $m_1 * (n_{A2} + 1)(n_{B2} + 1)$ number of pairs will be created using the above procedure. For each pair in the second stage, a log-likelihood ratio and its probability are computed with the help of the probabilities of

the previous stage using the Markov chain approach.

For a given pair from stage two, say $(\tilde{s}_{A2}, \tilde{s}_{B2})$ which was formed by the pair (s^*_{A1}, s^*_{B1}) from the continuation region of stage one,

$$\begin{aligned} Pr[LLR(\tilde{s}_{A2}, \tilde{s}_{B2}|s^*_{A1}, s^*_{B1}, p_0)] &= Pr[LLR(s^*_{A1}, s^*_{B1})] * Pr(\tilde{s}_{A2}|n_{A2}, p_0) * Pr(\tilde{s}_{B2}|n_{B2}, p_0) \\ &= Pr[LLR(s^*_{A1}, s^*_{B1})] * \binom{n_{A2}}{\tilde{s}_{A2}} (p_0)^{\tilde{s}_{A2}} (1-p_0)^{(n_{A2}-\tilde{s}_{A2})} \\ &\quad * \binom{n_{B2}}{\tilde{s}_{B2}} (p_0)^{\tilde{s}_{B2}} (1-p_0)^{(n_{B2}-\tilde{s}_{B2})}. \end{aligned} \tag{3.18}$$

The identical pair for the second stage is created using various combinations of pairs from the continuation region of the previous stage. For example, consider the resulting pair at the second stage as (1,2). This pair (1,2) can be created using six combinations of first stage and second stage as shown below.

Table 3.2: Examples of different combinations which create the same pair (1,2).

Combinations from stage 1	(s$_{A2}$, s_{B2})	Resulting pair ($\tilde{s}_{A2}$,$\tilde{s}_{B2}$)
(0,0)	(1,2)	(0+1, 0+2) = (1,2)
(0,1)	(1,1)	(0+1, 1+1) = (1,2)
(0,2)	(1,0)	(0+1, 2+0) = (1,2)
(1,0)	(0,2)	(1+0, 0+2) = (1,2)
(1,1)	(0,1)	(1+0, 1+1) = (1,2)
(1,2)	(0,0)	(1+0,2+0) = (1,2)

The probability associated with each unique pair is computed by adding the probabilities of all possible combinations that result in the given unique pair. That is, for example, the probability associated with the pair (1,2) can be computed by adding the probabilities associated with the six different combinations shown in Table 3.2.

Using all possible LLR values and associated probabilities that are known, the critical values for the second stage are computed using the same approach explained for stage one.

k^{th} **Stage** ($k > 1$)

The above-described methodology is generalized for any conditional stage ($k > 1$) as follows. Let n_{Ak} and n_{Bk} be the size of the samples added to the existing sample at stage k from group A and group B, respectively. Then,

N_{Ak}-Total number of data points (size of the combined sample) of group A at stage k

$$N_{Ak} = n_{A1} + n_{A2} + ... + n_{Ak}. \tag{3.19}$$

N_{Bk}-Total number of data points (size of the combined sample) of group B at stage k

$$N_{Bk} = n_{B1} + n_{B2} + ... + n_{Bk}. \tag{3.20}$$

Let s_{Ak} is the number of successes out of n_{Ak} and s_{Bk} is the number of successes out of n_{Bk}. At k^{th} stage, s_{Ak} can have any value from zero to n_{Ak} and s_{Bk} can have any value from zero to n_{Bk}.

$$s_{Ak} = 0, 1, 2, \ldots, n_{Ak}, \qquad and \qquad s_{Bk} = 0, 1, 2, \ldots, n_{Bk}. \tag{3.21}$$

All the $(s_{A(k-1)}, s_{B(k-1)})$ pairs, which are not in the rejection region of $(k-1)^{th}$ stage, are considered in the k^{th} stage of the analysis. For a given pair which is in the continuation region of the $(k-1)^{th}$ stage, say $(s^*_{A(k-1)}, s^*_{B(k-1)})$, can create $(n_{Ak}+1)(n_{Bk}+1)$ number of combinations uniting with the data in k^{th} stage. If there's '$m_{(k-1)}$' number of pairs in continuation region of $(k-1)^{th}$ stage, then $m_{(k-1)} * (n_{Ak}+1)(n_{Bk}+1)$ number of pairs will be created using the above procedure.

For each pair in the k^{th} stage, we compute a log-likelihood ratio using the Markov chain approach, as detailed earlier. For any single pair from stage k $(\tilde{s}_{Ak}, \tilde{s}_{Bk})$ which was formed by the pair $(s^*_{A(k-1)}, s^*_{B(k-1)})$ from the continuation region of the previous stage,

$$
\begin{aligned}
&Pr[LLR(\tilde{s}_{Ak}, \tilde{s}_{Bk}|s^*_{A(k-1)}, s^*_{B(k-1)}, p_0)] \\
&\qquad = Pr[LLR(s^*_{A(k-1)}, s^*_{B(k-1)}|p_0)] * Pr(\tilde{s}_{Ak}|n_{Ak}, p_0) * Pr(\tilde{s}_{Bk}|n_{Bk}, p_0) \\
&\qquad = Pr[LLR(s^*_{A(k-1)}, s^*_{B(k-1)}|p_0)] * \binom{n_{Ak}}{\tilde{s}_{Ak}} (p_0)^{\tilde{s}_{Ak}} (1-p_0)^{(n_{Ak}-\tilde{s}_{Ak})} \\
&\qquad\quad * \binom{n_{Bk}}{\tilde{s}_{Bk}} (p_0)^{\tilde{s}_{Bk}} (1-p_0)^{(n_{Bk}-\tilde{s}_{Bk})}.
\end{aligned}
\tag{3.22}
$$

Let's assume there are 'M_k' number of distinct LLR values at k^{th} stage. Therefore, all possible test statistic values for k^{th} stage are, $T_1^{kth}, T_2^{kth}, \ldots, T_{M_k}^{kth}$ and the ordered values are,

$$T_{(1)}^{kth}, T_{(2)}^{kth}, \ldots, T_{(M_k)}^{kth}$$

An algorithm resembling the bisection method is used to search for a specific LLR value that has an error rate extremely close to the expected alpha spending at a given stage. The search starts from the center of an ordered list of LLR values and proceeds to higher or lower test statistic/LLR values until the desired LLR value is found. Once the LLR value is identified, it is selected as the critical value.

Let,

$$c_k = min\{j \in \mathbb{N} : Pr(T^{kth} \geq T_{(j)}^{kth}|p_A = p_B) \leq \alpha_k\}, \tag{3.23}$$

where α_k is the amount of alpha sending for the k^{th} stage. Thus, $T_{(c_k)}^{kth}$ becomes the critical value of the k^{th} stage.

$$cv_k = T_{(c_k)}^{kth}, \tag{3.24}$$

which makes, $\left\{Pr\left(T_{(c_k)}^{kth}\right) + Pr\left(T_{(c_k+1)}^{kth}\right) + ... + Pr\left(T_{(M_k)}^{kth}\right)\right\} \leq \alpha_k$.

3.2 Type I error rate and power

Type I error rate and power can be calculated using the same methods described in Chapter 2. For a given stage, the probability of rejecting the null hypothesis when it is true can be expressed as a sum of partial probabilities. Let's discuss the unconditional (first) stage.

Once the critical value for the first stage (cv_1) has been identified, the type I error rate can be computed using the probabilities associated with each test statistic value.

Let,

$$e_1^* = Pr(T^{1st} \geq cv_1) = Pr(T^{1st} \geq T^{1st}_{(c_1)}) = \sum_{i=c_1}^{n_1} Pr(T^{1st}_{(i)}). \tag{3.25}$$

The probability of the test statistic value that falls in the rejection region is computed using the pairs of (s_{A1}, s_{B1}) that were rejected in the first stage, as,

$$e_1^* = \sum_{(s_{A1},s_{B1})\in s_1^R} \Big\{ \binom{n_{A1}}{s_{A1}} (p_0)^{s_{A1}} (1-p_0)^{n_{A1}-s_{A1}} * \binom{n_{B1}}{s_{B1}} (p_0)^{s_{B1}} (1-p_0)^{n_{B1}-s_{B1}} \Big\}. \tag{3.26}$$

As the second stage or any subsequent stage, is conditional on the previous stage, the type I error rate calculation is generalized as for any k^{th} stage ($k > 1$) as follows.

Let,

$$e_k^* = Pr(T^{kth} \geq cv_k) = Pr(T^{kth} \geq T^{kth}_{(c_k)}) = \sum_{\forall i \geq c_k} Pr(T^{kth}_{(i)}). \tag{3.27}$$

This refers to the probability of all test statistic values of k^{th} stage that falls in the rejection region. For a known sample sizes (N_{Ak}, N_{Bk}) and probability of success under null hypothesis (p_0), the above-mentioned probability only depends on all pairs of $(\tilde{s}_{Ak}, \tilde{s}_{Bk})$ which were rejected at the k^{th} stage.

Thus, the type I error rate for the k^{th} stage is computed as,

$$\begin{aligned} e_k^* = & \sum_{(s^*_{A(k-1)},s^*_{B(k-1)})\in s^C_{(k-1)},(\tilde{s}_{Ak},\tilde{s}_{Bk})\in s_k^R} \Big\{ Pr[LLR(s^*_{A(k-1)}, s^*_{B(k-1)}|p_0)] \\ & * \binom{n_{Ak}}{\tilde{s}_{Ak}} (p_0)^{\tilde{s}_{Ak}} (1-p_0)^{(n_{Ak}-\tilde{s}_{Ak})} * \binom{n_{Bk}}{\tilde{s}_{Bk}} (p_0)^{\tilde{s}_{Bk}} (1-p_0)^{(n_{Bk}-\tilde{s}_{Bk})} \Big\}. \end{aligned} \tag{3.28}$$

If the analysis consists of a total of K stages, the overall type I error rate (e_T) is the summation of error rates at each stage.

$$\begin{aligned} e_T &= e_1^* + e_2^* + ... + e_K^*, \\ &= Pr\Big(T^{1st} \geq cv_1\Big) + Pr\Big(T^{2nd} \geq cv_2\Big) + ... + Pr\Big(T^{Kth} \geq cv_K\Big), \\ &= Pr\Big(T^{1st} \geq T^{1st}_{(c_1)}\Big) + Pr\Big(T^{2nd} \geq T^{2nd}_{(c_2)}\Big) + ... + Pr\Big(T^{Kth} \geq T^{Kth}_{(c_K)}\Big), \\ &= \sum_{\forall i \geq c_1} Pr\Big(T^{1st}_{(i)}\Big) + \sum_{\forall i \geq c_2} Pr\Big(T^{2nd}_{(i)}\Big) + ... + \sum_{\forall i \geq c_K} Pr\Big(T^{Kth}_{(i)}\Big). \end{aligned} \tag{3.29}$$

Under a specific alternative hypothesis H_a, the power of the test $(1-\beta)$ is the probability of rejecting the null hypothesis when the alternative hypothesis is true. That is the probability of having an LLR that is larger than the specified critical value.

$$\begin{aligned} (1-\beta) = P_{H_1}(T^{1st} \in R_1) + P_{H_1}(T^{1st} \in C_1, T^{2nd} \in R_2) + ...+ \\ P_{H_1}(T^{1st} \in C_1, ..., T^{(K-1)th} \in C_{K-1}, T^{Kth} \in R_K). \end{aligned} \tag{3.30}$$

In order to compute power, the hypotheses for comparing two proportions are,

$$\begin{aligned} H_0 : p_A = p_B = p_0, \qquad vs \qquad & H_a : p_A \neq p_B, \\ & H_a : p_A = p_0, \quad and \quad p_B = p_1. \end{aligned} \tag{3.31}$$

For a given critical value (cv_k), the probability of all LLR values which are in rejection region (of k^{th} stage) when H_a is true, is computed as,

$$(1-\beta)_k = Pr(LLR \geq cv_k | p_A = p_0, p_B = p_1) = \sum_{i=c_k}^{M_k} Pr(T^{kth}_{(i)} | p_A = p_0, p_B = p_1), \tag{3.32}$$

where M_k is the number of possible test statistic (LLR) values at the k^{th} stage.

Therefore, the power of the test at k^{th} stage is,

$$\begin{aligned} (1-\beta)_k = \sum_{(s^*_{A(k-1)}, s^*_{B(k-1)}) \in s^C_{(k-1)}, (\tilde{s}_{Ak}, \tilde{s}_{Bk}) \in s^R_k} & \Big\{ Pr[LLR(s^*_{A(k-1)}, s^*_{B(k-1)} | p_A = p_0, p_B = p_1)] \\ & * \binom{n_{Ak}}{\tilde{s}_{Ak}} (p_0)^{\tilde{s}_{Ak}} (1-p_0)^{(n_{Ak} - \tilde{s}_{Ak})} * \binom{n_{Bk}}{\tilde{s}_{Bk}} (p_1)^{\tilde{s}_{Bk}} (1-p_1)^{(n_{Bk} - \tilde{s}_{Bk})} \Big\}. \end{aligned} \tag{3.33}$$

Then, the overall power of the test can be expressed as the sum of partial probabilities given by,

$$(1-\beta)=\sum_{k=1}^{K}(1-\beta)_k=\sum_{k=1}^{K}\Big\{\sum_{i=c_k}^{M_K}Pr(T_{(i)}^{kth}|p_A=p_0,p_B=p_1)\Big\}. \qquad (3.34)$$

3.3 Comparison of alpha spending functions

3.3.1 Critical values and type-I error rate

Consider a RCT with alpha spending functions based on Pocock, O'Brien & Fleming, and Haybittle-Peto methods. Tables 3.3 -3.8 present the critical values and type I error rates for each stage along with the overall type I error rate for various sample sizes for two-stage and three-stage RCT where $p_0 = 0.4$.

Pocock Design

Table 3.3: Critical values and type I error rate for a two-stage analysis based on Pocock's method.

n	cv1	cv2	error1	error2	Overall error rate
5	8.45621	5.93597	0.018844876	0.020286117	0.039130994
10	6.55586	5.13450	0.015698283	0.024377983	0.040076267
15	5.68292	4.90216	0.021472697	0.021640569	0.043113267
20	5.38341	4.59392	0.023777523	0.024974426	0.048751950
25	5.30886	4.54035	0.023368385	0.024931787	0.048300173
30	5.49067	4.64352	0.024736939	0.024351758	0.049088697
35	5.12608	4.54556	0.024681524	0.024069818	0.048751343
40	5.18861	4.55816	0.023707343	0.024455248	0.048162591

The critical values and error rates were obtained for groups of the same size ($n_{Ak} = n_{Bk}$). Additionally, at each stage, a sample of the same size is combined with the existing sample. Therefore, 'n' in Table 3.3 represents the initial sample size of each group as well as the size of the sample introduced at each stage.

In Table 3.3, cv1 and cv2 refer to the critical values of stage one and stage two, respectively. For example, when $n = 10$, if the calculated test statistic for a given data is greater than 6.55586 at the first stage, the null hypothesis is rejected. The trial will continue to the second stage if the calculated test statistic is less than 6.55586. Then, at the end of the second stage, reject the null hypothesis if the calculated test statistic is greater than 5.13450. Furthermore, the observed type I error rates of stages one and stage two are represented by error1 and error2, accordingly. The observed overall type I error rate is the summation of error1 and error2.

From Table 3.3, we see that as the sample size increases, the critical value of each stage converges. Note that due to the discrete nature of data, only a few specific values can serve as critical values. Consequently, the critical value for small sample sizes (especially $n = 5$) is slightly greater than the critical value for large sample sizes. Similarly, we observe that the observed type I error rates converge as the sample size grows. The error rate values rise to achieve the expected alpha spending at each stage when the sample size increases. Accordingly, the total error rate increases to achieve the expected α. Moreover, we see that all two stages converge to an approximate type I error rate of 0.024. Since these observed error rates are selected to be less than the alpha spending at each stage, the total type I error rate is always less than the expected significance level ($\alpha = 0.05$). Thus, the overall error rate seems to be converged to 0.048.

Table 3.4: Critical values and type I error rate for a three-stage analysis based on Pocock's method.

N	cv1	cv2	cv3	error 1	error2	error3	Overall error rate
5	13.86294	6.55586	5.68292	0.001592525	0.01509269	0.01488195	0.03156716
10	6.55586	5.81223	5.02296	0.01569828	0.01461591	0.01629138	0.04660557
15	6.16301	5.64865	5.05254	0.01620210	0.01591045	0.01525180	0.04736435
20	5.99054	5.36311	5.01293	0.01581980	0.01642942	0.01575049	0.04799970
25	5.89344	5.30258	4.88654	0.01580713	0.01608464	0.01645130	0.04834307
30	5.83114	5.34004	5.02913	0.01490047	0.01591968	0.01664155	0.04746169
35	6.00666	5.27015	4.99932	0.01568328	0.01625600	0.01658618	0.04852545
40	5.83197	5.28233	4.98509	0.01654489	0.01596070	0.01619842	0.04870401

From Table 3.4, we see that as the sample size increases, the critical value and the error rate of each stage converges. However, considering the convergence of data, we see that $cv_1 \approx 5.9, cv_2 \approx 5.3$ and $cv_3 \approx 5.0$ for large samples. Also it is noticeable that all three stages converge to an approximate type I error rate of 0.016 and the overall error rate seems to be converged to 0.049.

O'Brien and Fleming Design

Table 3.5: Critical values and type I error rate for a two-stage analysis based on O'Brien & Fleming's method.

n	cv1	cv2	error1	error2	Overall error rate
5	13.86294	5.30022	0.001592524	0.039314396	0.040906920
10	8.63046	4.40239	0.005973725	0.037163302	0.043137028
15	7.45888	4.40205	0.008813241	0.038158973	0.046972214
20	7.36182	4.22001	0.008275853	0.039915118	0.048190971
25	7.08980	4.17060	0.009104756	0.038479678	0.047584434
30	7.11113	4.16780	0.008889434	0.038612362	0.047501797
35	7.05801	4.19081	0.009892919	0.039939426	0.049832346
40	6.76542	4.00595	0.009395277	0.039724777	0.049120055

The critical values and type I error rate for various sample sizes for a two-stage RCT using O'Brien and Fleming's approach are shown in Table 3.5. It is noticable that as the sample size increases, the critical value and the error rate of each stage converges.

The critical values and type I error rates for various sample sizes for a three-stage RCT using O'Brien and Fleming's approach are shown in Table 3.6. It is clearly noticeable that the critical values are monotonically decreasing. Consider $n = 15$. The critical value is 9.50504 for the first stage, 5.83114 for the second stage and 4.20760 for the final stage. Furthermore, we see that the critical value of each stage converges as the sample size grows. It is noticeable that the critical values for $n = 5$ and $n = 10$ are slightly greater than the critical value for large sample sizes. However, considering the convergence of data, we see that $cv_1 \approx 8.6, cv_2 \approx 5.9$ and $cv_3 \approx 4.1$ for large samples.

Table 3.6: Critical values and type I error rate for a three-stage analysis based on O'Brien & Fleming's method

N	cv1	cv2	cv3	error 1	error2	error3	Overall error rate
5	13.86294	7.70979	5.05812	0.001592525	0.01203605	0.02470096	0.03832954
10	10.97434	6.58263	4.43483	0.003505039	0.0133913	0.02880882	0.04570516
15	9.50504	5.83114	4.20760	0.003369754	0.01362822	0.03160567	0.04860364
20	9.09623	6.23276	4.23581	0.003448242	0.0133202	0.03097545	0.04774389
25	8.97213	5.98982	4.13729	0.003277688	0.01364946	0.03063392	0.04756106
30	8.59352	5.91863	4.10696	0.003564179	0.01389037	0.03149775	0.04895229
35	8.60081	5.90597	4.16238	0.003519533	0.01371822	0.03183764	0.04907539
40	8.65892	5.89788	4.12003	0.00399697	0.01385637	0.03158019	0.04943354

In a similar manner, we notice that the error rates converge as the sample size increases. Thus, the approximate error rates can be noted as 0.004, 0.014 and 0.032 for stages 1, 2 and 3, respectively. It is clear that as the sample size grows, stage-wise error rates increase to obtain the anticipated alpha spending at each stage. To reach the desired level of statistical significance (α), the overall error rate may therefore grow. Therefore, the overall error rate seems to be converged to 0.049.

Haybittle-Peto Design

Table 3.7: Critical values and type I error rate for a two-stage analysis based on Haybittle-Peto's method.

n	cv1	cv2	error1	error2	Overall error rate
5	NA	4.69108	NA	0.042919960	0.042919960
10	13.68058	4.05375	0.000867957	0.047604472	0.048472429
15	12.98614	3.89145	0.000655116	0.048380587	0.049035703
20	11.20025	4.15227	0.000977691	0.046912610	0.04789030
25	10.84868	4.08804	0.000989207	0.048016429	0.049005636
30	11.66421	4.06670	0.000906538	0.048857664	0.049764202
35	11.22909	3.82504	0.000987286	0.047791324	0.048778610
40	11.22740	3.85974	0.000909966	0.048543445	0.049453412

In Table 3.7, the expected alpha spending at stage one is relatively small (0.001) compared to the alpha spending at stage 2 (0.049). When n=5, the computer program has failed to produce a critical value for the first stage. This indicates that the error rate associated with all potential test statistic values exceeds the expected alpha spending for the initial stage. Having a very small alpha spending at the first stage makes the test to be continued to the second stage when n=5. Then the calculated test statistic for the second stage can be compared with 4.69108, and the null hypothesis can be rejected if the test statistic exceeds this cv2.

Table 3.8: Critical values and type I error rates for a three-stage analysis based on Haybittle-Peto's method.

N	cv1	cv2	cv3	error 1	error2	error3	Overall error rate
5	NA	13.68058	4.14392	NA	0.000867957	0.04794403	0.04881198
10	13.68058	11.20025	3.89145	0.000867957	0.000867957	0.04765534	0.04939700
15	12.98614	11.36583	3.91293	0.000655116	0.000927798	0.04741874	0.04900165
20	11.20025	10.79097	4.08543	0.000977691	0.000906737	0.04744912	0.04933354
25	10.84868	10.81085	3.96245	0.000989207	0.000961602	0.04775777	0.04970858
30	11.66421	10.75651	3.92100	0.000906539	0.000979851	0.04666388	0.04855027
35	11.22909	10.75109	3.92176	0.000987286	0.000945166	0.0463897	0.04832216
40	11.22740	10.72623	3.90766	0.000909966	0.000931972	0.04683233	0.04867426

Table 3.8 shows the critical values and type I error rate for various sample sizes for a three-stage RCT using Haybittle-Peto's approach. We can see that having a very small alpha spending at the first stage makes the test to be continued to the second stage when n=5. Then the calculated test statistic for the second stage is compared with 13.68058, and the null hypothesis is rejected if the test statistic exceeds 13.68058. If the trial continues to the third stage, the null hypothesis is rejected if the calculated test statistic exceeds 4.14392.

The critical values and observed type I error rate for each stage converge as the sample size increases. We can observe that $cv_1 \approx 11.2, cv_2 \approx 10.8$ and $cv_3 \approx 3.9$ for large samples. Also, the approximate error rate values can be noted as 0.0009, 0.0009 and 0.047 for stages 1, 2 and 3, respectively. Additionally, the overall error rate seems to be converged to 0.049.

From Tables 3.3 to 3.8, there is one more noticeable outcome that needs to be addressed. In the first stage of any design, we conduct a general likelihood ratio test, so the critical value should follow a chi-squared distribution with the degree of freedom equal to one. When we consider the convergence of the critical values of the above methods, we can see that at each stage, the converged value is approximately equal to the chi-squared value under the expected significance level. However, the critical values of the second stage and subsequent stages deviate from this chi-squared value as those stages are conditional on the previous stage.

The alpha spending functions of the above-discussed approaches are designed to achieve the overall significance level of the test. Let's consider the three-stage RCT. As we can see from Tables 3.4, 3.6 and 3.8, the overall error rate seems to converge to 0.049 for all three designs. Therefore, let's compare the convergence speed of these three approaches.

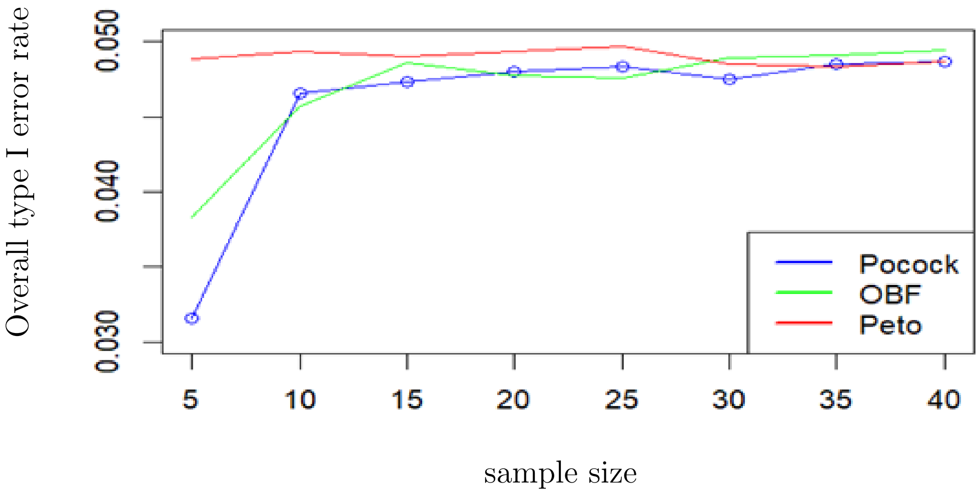

Figure 3.1: Observed Overall type I error rate for various sample sizes (three-stage design based on Pocock, O'Brien & Fleming, Haybittle-Peto methods with confidence level =0.05)

Figure 3.1 compares the observed overall type I error rate for Pocock, OBF and Haybittle-Peto designs. It is noticeable that Haybittle-Peto's design has less deviated values and achieves the expected significance level even for small sample sizes. Pocock and OBF method have more deviated values for sample sizes less than 20. However, all the methods converge to a value approximately equal to 0.049 when the sample size increases.

3.3.2 Power

The following tables show the power $(1 - \beta)$ for each stage along with the overall power for a two-stage design, where

$$\begin{aligned} H_0 : p_A = p_B = 0.4, \qquad & vs \qquad H_a : p_A = 0.4, p_B = 0.2, \qquad and \\ & \qquad\quad H_a : p_A = 0.4, p_B = 0.5, \qquad and \\ & \qquad\quad H_a : p_A = 0.4, p_B = 0.7. \end{aligned} \tag{3.35}$$

Table 3.9: Power for a two-stage design based on Pocock's method

	$p_1 = 0.2$			$p_1 = 0.5$			$p_1 = 0.7$		
n	**power1**	**power2**	**Overall Power**	**power1**	**power2**	**Overall Power**	**power1**	**power2**	**Overall Power**
5	0.03332106	0.0960127	0.1293338	0.027	0.02643861	0.05343861	0.08513994	0.1122081	0.1973481
10	0.08586644	0.142454	0.2283204	0.02335223	0.04934045	0.07269268	0.1265415	0.2799866	0.4065281
15	0.1434708	0.213306	0.3567767	0.03855402	0.05794425	0.09649828	0.2455363	0.3413866	0.5869229
20	0.1926369	0.2819507	0.4745876	0.04879363	0.08015245	0.1289461	0.3474889	0.3951685	0.7426574
25	0.2499045	0.3114267	0.5613312	0.0546873	0.09901775	0.153705	0.4305695	0.4087034	0.8392729
30	0.2746106	0.3597293	0.63434	0.07621129	0.1064223	0.1826336	0.5595594	0.3427688	0.9023282
35	0.3510623	0.357879	0.7089414	0.07695307	0.1219779	0.198931	0.6122212	0.3267708	0.938992
40	0.3875469	0.3804764	0.7680233	0.07953149	0.1351	0.2146315	0.6635039	0.2986361	0.96214

Table 3.9 presents the statistical power of three different alternative hypotheses; $p_B = 0.2, p_B = 0.5$ and $p_B = 0.7$ for a two-stage RCT with $p_A = p_0 = 0.4$ and equal alpha spending at each stage (Pocock's design). The results were obtained for various sample sizes(n) for groups A and B, assuming that equal-sized samples were introduced at each stage. Since this study has two stages, 'power1' and 'power2' refer to the statistical power at stages one and stage two, respectively. 'Overall power' of the analysis has been obtained by adding the power of the two stages.

From Table 3.9, we see that the power of stage one increases when the sample size increases for any given p_B. Similarly, the overall power increases as the sample size grows. Additionally, when the value of p_B is far from p_0 value, the overall power for a given sample size is greater than that of when p_1 is close to p_0. Consider $n = 25$ as an example. As $p_B = 0.5$ is the nearest value to 0.4 and $p_B = 0.7$ is the farthest among these three p_B values given. Note that the overall power is 0.1537 when $p_1 = 0.5$, it is 0.5613 when $p_1 = 0.2$ and 0.8393 when $p_1 = 0.7$. In addition, Table 3.9 shows that when the power is $> 80\%$, a higher proportion of total power is achieved in the first stage, leaving the remainder for the second stage.

Tables 3.10 and 3.11 present the statistical power using O'Brien & Fleming design and Haybittle-Peto design for a two-stage test with $p_A = p_0 = 0.4$ and three different alternative hypotheses; $p_B = 0.2, p_B = 0.5$ and $p_B = 0.7$.

Table 3.10: Power for a two-stage design based on O'Brien & Fleming's method

	$p_1 = 0.2$			$p_1 = 0.5$			$p_1 = 0.7$		
n	**power1**	**power2**	**Overall Power**	**power1**	**power2**	**Overall Power**	**power1**	**power2**	**Overall Power**
5	0.00338	0.128337	0.1317177	0.00275	0.06025394	0.06300393	0.01309401	0.2358144	0.2489084
10	0.043437	0.229691	0.2731284	0.008533438	0.07440134	0.08293477	0.05809485	0.3925739	0.4506688
15	0.080459	0.295713	0.3761722	0.01769037	0.08940288	0.1070932	0.1513178	0.4722911	0.6236089
20	0.110255	0.373963	0.484218	0.02077368	0.1083667	0.1291404	0.220262	0.5337677	0.7540297
25	0.147525	0.437367	0.5848914	0.02624519	0.1239889	0.1502341	0.3041631	0.5387831	0.8429462
30	0.176537	0.476192	0.6527288	0.02998943	0.1431633	0.1731527	0.378025	0.5255427	0.9035676
35	0.215926	0.509293	0.7252191	0.0421287	0.1648648	0.2069934	0.4918917	0.4541692	0.9460608
40	0.272505	0.523851	0.7963559	0.04505358	0.1993241	0.2443777	0.5541985	0.4177051	0.9719036

Table 3.11: Power for a two-stage design based on Haybittle-Peto's method.

	$p_1 = 0.2$			$p_1 = 0.5$			$p_1 = 0.7$		
n	**power1**	**power2**	**Overall Power**	**power1**	**power2**	**Overall Power**	**power1**	**power2**	**Overall Power**
5	NA	0.155527	0.155527	NA	0.063182	0.063182	NA	0.24895	0.2489498
10	0.006389	0.285172	0.291561	0.001887	0.085741	0.087628	0.019437	0.444118	0.4635551
15	0.012487	0.386814	0.399301	0.001764	0.121754	0.123518	0.031243	0.632068	0.6633117
20	0.032235	0.446986	0.479221	0.003249	0.130764	0.134013	0.070857	0.693614	0.7644705
25	0.038275	0.541985	0.58026	0.004697	0.155176	0.159873	0.119695	0.735309	0.855004
30	0.047798	0.611114	0.658911	0.005944	0.191313	0.197257	0.171411	0.748575	0.919986
35	0.070561	0.676868	0.747429	0.007303	0.215696	0.222999	0.228421	0.725154	0.9535751
40	0.083017	0.718746	0.801764	0.008253	0.231267	0.23952	0.281649	0.690004	0.9716529

From Tables 3.10 and 3.11 pwe can make similar conclusions as we did based on Table 3.9. The primary observation is that the overall power increases as the sample size grows. The following graphs can be used to compare these three approaches.

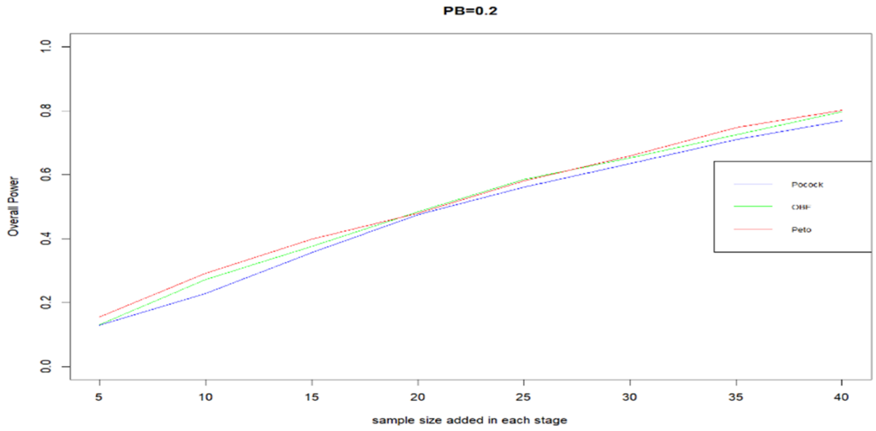

Figure 3.2: Observed Overall Power for various sample sizes when $p_A = 0.4$ and $p_B = 0.2$ (Two stage design based on Pocock, O'Brien & Fleming, Haybittle-Peto methods with confidence level =0.05)

The Figure 3.2 compares the overall power between Pocock, OBF and Haybittle-Peto designs. The results were obtained for various sample sizes and $p_B = 0.2$. The blue line denotes the power values obtained using Pocock's approach, the green line is for OBF, and the red line is for the Haybittle-Peto approach. The three lines do not appear to be significantly different from one another. Therefore, we can conclude that the power of the test for a given sample size is almost the same for all three designs when $p_B = 0.2$.

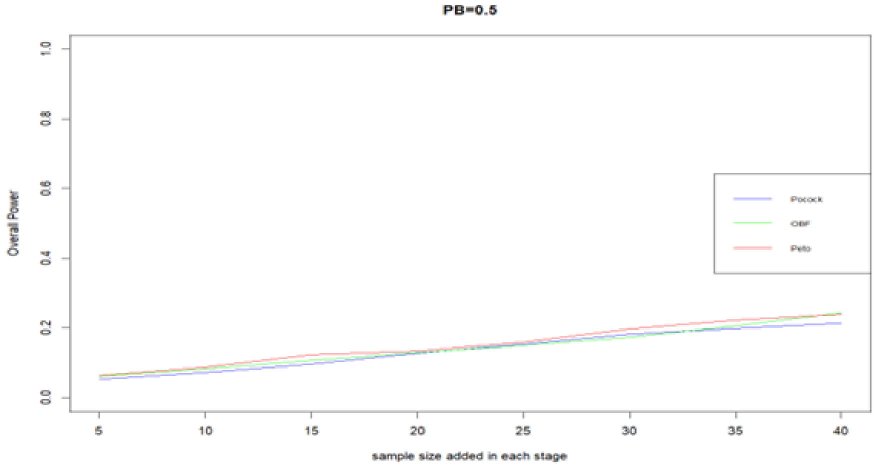

Figure 3.3: Observed Overall Power for various sample sizes when $p_A = 0.4$ and $p_B = 0.5$ (Two stage design based on Pocock, O'Brien & Fleming, Haybittle-Peto methods with confidence level =0.05))

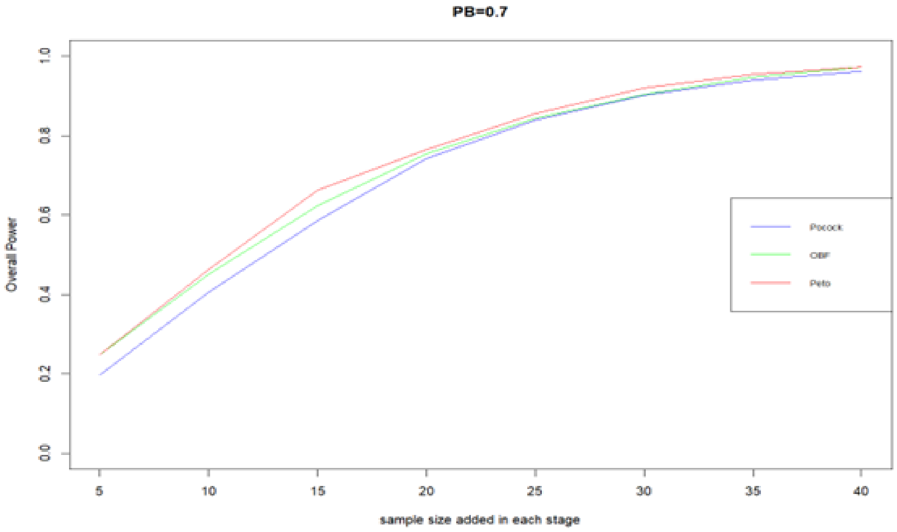

Figure 3.4: Observed Overall Power for various sample sizes when $p_A = 0.4$ and $p_B = 0.7$ (Two stage design based on Pocock, O'Brien & Fleming, Haybittle-Peto methods with confidence level =0.05)

Figures 3.3 and 3.4 illustrate that the overall power of the test increases as the sample size increases for all three approaches. Also, we see that the power of the test is almost equal for all three designs for any given sample size. Most of the time, the Haybittle-Peto design achieves a Power greater than the other two designs. Additionally, by comparing the three lines in Figures 3.3 and 3.4, we can conclude that when the value under a specific alternative hypothesis (p_B) is far away from the value under the null hypothesis (p_0), higher power can be achieved for large sample sizes.

Chapter 4

Summary and future work

4.1 Summary

Randomized clinical trials (RCT) are widely regarded as the most reliable clinical research strategy for comparing treatments. In RCT, the participants are randomly assigned to treatments to compare the effect of a treatment against a control. Randomization is used to ensure that statistical inference at the end of the study is reliable as it assists in reducing systematic bias, ascertainment bias, selection bias and accidental bias. There are several methods for randomly assigning individuals to treatment groups in clinical trials. Simple randomization, block randomization, stratified randomization, and adaptive randomization are the most used randomization methods.

The group sequential designs of RCT enables termination of the study early if preliminary results favour one treatment over the other. The key feature of the group sequential designs of RCT is that the sample size is not specified in advance. The approach performs multiple tests based on cumulative data, and sampling is terminated with a predefined stopping rule as soon as statistically significant findings are revealed. Given the maximum number of inspections, K, a group sequential design will have $(K-1)$ number of 'interim analyses/stages' and a final stage. However, this interim analysis may have hidden consequences. If the analysis does not reveal that the trial can be stopped early, the fact that the interim analysis was carried out could potentially undermine the power of the test. Alpha spending functions approach can be used as a solution to the concerns of accommodating interim analysis. Different functions were proposed to split the type I error rate between interim analyses and the final stage.

The group sequential designs enable early conclusions while maintaining statistical power and controlling the type I error rate of the study. For a group sequential RCT, the total type I error rate is maintained using the critical values that were calculated using the boundaries technique. Pocock (1977, 1982) [32, 33] and O'Brien and Fleming (1979) [29] proposed group sequential test techniques, which are frequently used in clinical research.

The objective of this study is to compute the critical values, type I error rate, and power of group sequential analysis with binary responses. Critical values create the boundaries which separate the acceptance, continuation, and rejection regions. However, in this study, the acceptance region is not considered; therefore, the critical values computed in the study define the boundary between the rejection and the continuation regions. As we consider groups of binary responses, we use the Binomial probability model and log-likelihood ratio as the test statistic in this study. The alpha spending functions used in this study are modified with the idea of the methods proposed by Pocock (1977,1982) [32, 33], O'Brien & Fleming (1979) [29] and Haybittle-Peto (1971,1976) [15, 31] . An iterative Markov chain technique is used to compute critical values that fulfil the alpha spending at each stage of the procedure.

First, the Markov chain technique is briefly explained with the help of a single sample scenario. The equations were developed to determine the test statistic, type I error rate, and power for a three-stage design with binary data. The generalised equations were then proposed, which can be used for any number of stages. A computer programme is created for two-stage and three-stage group sequential designs to calculate the critical values and the associated type I error rates at each stage along with the overall type I error rate. The results were obtained for various sample sizes with the same sample size at each stage. Graphs were created to illustrate the convergence of critical values, type I error rate of each stage, as well as the overall type I error rate as sample size increases. Another R function is written to calculate the power for any given alternative hypothesis and the results were presented for various sample sizes with graphs to demonstrate how the power of the test changes with the sample sizes.

The iterative Markov chain approach was extended to compare two proportions. The technique was well explained for two groups scenario while constructing three different alpha spending functions with the influence of Pocock's design, O'Brien & Fleming's design and Haybittle-Peto's design. Critical values, type I error rate and power for two-stage and three-stage designs have been calculated and results were presented for different sample sizes. The convergence of critical values, type I error rate and Power with increasing sample size was depicted using graphs. Finally, the outcomes of these three designs were analysed in order to compare the effectiveness in sequential designs of clinical trials with binary data.

When all the results are compared, it is clear that the critical values and type I error rate converge as the sample size increases. However, the overall type I error rate obtained by Haybittle-Peto's design converges quickly and with lesser deviations than the values obtained by the other two designs. Additionally, we could observe that when the sample size grows, the power of the test grows as well under different alternative hypotheses. However, the power values obtained under Pocock, O'Brien & Fleming and Haybittle-Peto approaches do not appear to be significantly different from one another for any given alternative hypothesis.

4.2 Future work

While our research is focused on computing critical values, type I error rate and power using an iterative Markov chain for one sample case as well as comparing two treatments, these results can be generalized to more than two treatments. Therefore, in the future, we aim to utilize the Markov chain process, which can be used to compare more than two treatments. Additionally, this study used equal sample sizes at every stage and also for each group when comparing two treatments. However, the proposed method can be generalized for unequal sample sizes at each stage as well as for each group.

Furthermore, this study focused only on three approaches proposed by Pocock (1977,1982) [32, 33], O'Brien Fleming (1979) [29] and Haybittle-Peto (1971,1976) [15, 31]. In Future, new stopping rules can also be proposed based on other popular alpha spending functions.

Bibliography

[1] K. Abt. Poisson sequential sampling modified towards maximal safety in adverse event monitoring. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 40(1):21–41, 1998.

[2] D. G. Altman and J. M. Bland. Treatment allocation in controlled trials: why randomise? *Bmj*, 318(7192):1209–1209, 1999.

[3] D. G. Altman, K. F. Schulz, D. Moher, M. Egger, F. Davidoff, D. Elbourne, P. C. Gøtzsche, T. Lang, and C. Group. The revised consort statement for reporting randomized trials: explanation and elaboration. *Annals of internal medicine*, 134(8):663–694, 2001.

[4] T. W. Anderson. A modification of the sequential probability ratio test to reduce the sample size. *The Annals of Mathematical Statistics*, pages 165–197, 1960.

[5] P. Armitage. Interim analysis in clinical trials. *Statistics in Medicine*, 10(6):925–937, 1991.

[6] P. Armitage, C. McPherson, and B. Rowe. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)*, 132(2):235–244, 1969.

[7] C. Berger and G. Casella. Statistical inference, duxbury press, n. *Scituate, MA.[Google Scholar]*, 2001.

[8] S. Choi. Truncated sequential designs for clinical trials based on markov chains. *Biometrics*, pages 159–168, 1968.

[9] H. F. Dodge and H. G. Romig. A method of sampling inspection. *The Bell System Technical Journal*, 8(4):613–631, 1929.

[10] H. Douke. On sequential design based on markov chains for selecting one of two treatments in clinical trials with delayed observations. *Journal of the Japanese Society of Computational Statistics*, 7(1):89–103, 1994.

[11] J. W. Frane. A method of biased coin randomization, its implementation, and its validation. *Drug information journal: DIJ/Drug Information Association*, 32(2):423–432, 1998.

[12] K. Gordon Lan and D. L. DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663, 1983.

[13] D. C. Hadorn, D. Baker, J. S. Hodges, and N. Hicks. Rating the quality of evidence for clinical practice guidelines. *Journal of clinical epidemiology*, 49(7):749–754, 1996.

[14] D. Harrington. Group sequential methods with applications to clinical trials. christopher jennison and bruce w. turnbull, crc/chapman & hall, uk, 2000. no. of pages: xviii+ 390. price:£ 39.00. isbn 0-849-30316-8, 2001.

[15] J. Haybittle. Repeated assessment of results in clinical trials of cancer treatment. *The British journal of radiology*, 44(526):793–797, 1971.

[16] A. Hayter. Recursive integration methodologies with statistical applications. *Journal of statistical planning and inference*, 136(7):2284–2296, 2006.

[17] D. Hoel, G. Weiss, and R. Simon. Sequential tests for composite hypotheses with two binomial populations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(3):302–308, 1976.

[18] C. Jennison and B. W. Turnbull. Exact calculations for sequential t, x2 and f tests. *Biometrika*, 78(1):133–141, 1991.

[19] C. Jennison and B. W. Turnbull. *Group sequential methods with applications to clinical trials*. CRC Press, 1999.

[20] L. A. Kalish and C. B. Begg. Treatment allocation methods in clinical trials: a review. *Statistics in medicine*, 4(2):129–144, 1985.

[21] J. H. B. Kemperman and G. Weiss. The passage problem for a stationary markov chain. *Physics Today*, 14(9):66, 1961.

[22] K. Kim and D. L. Demets. Design and analysis of group sequential tests based on the type i error spending rate function. *Biometrika*, 74(1):149–154, 1987.

[23] M. Kulldorff, R. L. Davis, M. Kolczak, E. Lewis, T. Lieu, and R. Platt. A maximized sequential probability ratio test for drug and vaccine safety surveillance. *Sequential analysis*, 30(1):58–78, 2011.

[24] A. Kumar and B. S. Chakraborty. Interim analysis: a rational approach of decision making in clinical trial. *Journal of advanced pharmaceutical technology & research*, 7(4):118, 2016.

[25] J. M. Lachin, J. P. Matts, and L. Wei. Randomization in clinical trials: Conclusions and recommendations. *Controlled Clinical Trials*, 9(4):365–374, 1988.

[26] T. Lai. Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: A sequential approach. *Communications in Statistics-Theory and Methods*, 13(19):2355–2368, 1984.

[27] T. L. Lai. Optimal stopping and sequential tests which minimize the maximum expected sample size. *The Annals of Statistics*, pages 659–673, 1973.

[28] C. McPherson and P. Armitage. Repeated significance tests on accumulating data when the null hypothesis is not true. *Journal of the Royal Statistical Society: Series A (General)*, 134(1):15–25, 1971.

[29] P. C. O'Brien and T. R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, pages 549–556, 1979.

[30] E. Paxinou, D. Kalles, C. T. Panagiotakopoulos, and V. S. Verykios. Analyzing sequence data with markov chain models in scientific experiments. *SN Computer Science*, 2(5):1–14, 2021.

[31] R. Peto, M. Pike, P. Armitage, N. Breslow, D. Cox, S. V. Howard, N. Mantel, K. McPherson, J. Peto, and P. Smith. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. i. introduction and design. *British journal of cancer*, 34(6):585–612, 1976.

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

[33] S. J. Pocock. Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics*, pages 153–162, 1982.

[34] S. J. Pocock and R. Simon. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, pages 103–115, 1975.

[35] G. W. Pulford. Markov chain analysis of the sequential probability ratio test for automatic track maintenance. In *Proc. 6th Conf. Info. Fusion, Cairns*, pages 1258–1265, 2003.

[36] K. F. Schulz and D. A. Grimes. Allocation concealment in randomised trials: defending against deciphering. *The Lancet*, 359(9306):614–618, 2002.

[37] W. A. Shewhart. *Economic control of quality of manufactured product.* Macmillan And Co Ltd, London, 1931.

[38] I. R. Silva, M. Kulldorff, and W. Katherine Yih. Optimal alpha spending for sequential analysis with binomial data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1141–1164, 2020.

[39] O. Sverdlov, W. F. Rosenberger, et al. Randomization in clinical trials: can we eliminate bias? *Clinical Investigation*, 3(1):37–47, 2013.

[40] S. Todd. A 25-year review of sequential methodology in clinical studies. *Statistics in medicine*, 26(2):237–252, 2007.

[41] M. E. Valentinuzzi. Friedman lm, furberg cd, demets dl: Fundamentals of clinical trials 3rd edition, 2004.

[42] A. Wald. Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics*, 16(2):117 – 186, 1945.

[43] G. Wassmer and W. Brannath. *Group sequential and confirmatory adaptive designs in clinical trials*, volume 301. Springer, 2016.

[44] K. Weigl and I. Ponocny. Group sequential designs applied in psychological research. *Methodology*, 16(1):75–91, 2020.

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

[46] J. Whitehead and I. Stratton. Group sequential clinical trials with triangular continuation regions. *Biometrics*, pages 227–236, 1983.

[47] J. Whitehead, S. Todd, A. Whitehead, and N. Stallard. Interim analyses in clinical trials. *British Journal of Clinical Pharmacology*, 51(5):393, 2001.

[48] F. Wilcoxon, L. Rhodes, and R. A. Bradley. Two sequential two-sample grouped rank tests with applications to screening experiments. *Biometrics*, pages 58–84, 1963.

[49] W. H. Woodall and M. R. Reynolds Jr. A discrete markov chain representation of the sequential probability ratio test. *Sequential Analysis*, 2(1):27–44, 1983.

[50] Y. Yi. Exact statistical power for response adaptive designs. *Computational Statistics & Data Analysis*, 58:201–209, 2013.

[51] M. Zalene. Randomized consent designs for clinical trails: An upadate. *Stat Med*, 9(6):645–656, 1990.