AN APPROACH FOR FINDING A GENERAL APPROXIMATION TO THE GROUP SEQUENTIAL BOOTSTRAP TEST

Submitted by
Douglas Ekstedt

A thesis submitted to the Department of Statistics in partial fulfillment of the requirements for a two-year Master of Science degree in Statistics in the Faculty of Social Sciences

Supervisor
Mattias Nordin

Spring, 2022
Randomized experiments are regarded as the gold standard for estimating causal effects. Commonly, a single test is performed using a fixed sample size. However, observations may also be observed sequentially and because of economical and ethical reasons, it may be desirable to terminate the trial early. The group sequential design allows for interim analyses and early stopping of a trial without the need for continuous monitoring of the accumulating data. The implementation of a group sequential procedure requires that the sampling distribution of the test statistic observed at each wave of testing to have a known or asymptotically known sampling distribution. This thesis investigates an approach for finding a general approximation to the group sequential bootstrap test for test statistics with unknown or analytically intractable sampling distributions. There is currently no bootstrap version of the group sequential test. The approach implies approximating the covariance structure of the test statistics over time, but not the marginal sampling distribution, with that of a normal test statistic. The evaluation is performed with a Monte Carlo simulation study where the achieved significance level is compared to the nominal. Evidence from the Monte Carlo simulations suggests that the approach performs well for test statistics with sampling distributions close to a normal distribution.

Keywords: group sequential design, alpha spending, bootstrap, false positive rate
## Contents

1 Introduction  
2 Literature review  
3 Theory  
   3.1 The alpha spending approach  
   3.2 Repeated significance testing with known sampling distribution  
   3.3 Repeated significance testing with unknown sampling distribution  
4 Monte Carlo simulation  
5 Results  
   5.1 Equal increments in statistical information  
   5.2 Unequal increments in statistical information  
6 Conclusion  
References  
Appendices
1 Introduction

Randomized experiments are regarded as the gold standard for estimating causal effects (Rubin, 2007). Commonly, a single test is performed using predetermined sample size. However, observations may also be observed sequentially and because of economical and ethical reasons, it may be desirable to end the trial early. In the sequential design, the sample size is not predetermined and multiple tests are performed. At each stage of testing, a decision is made about whether to collect more data or end the trial early based on a predetermined stopping rule (Bartroff, Lai, & Shih, 2013). The sequential design not only allows companies and organizations to perform efficient analyses of product changes but also to obtain results quickly. Moreover, with these strengths, the sequential design has become popular in A/B-testing, in which fast data-driven decision making is essential for businesses (Liu, Sun, Varshney, & Xu, 2019).

The sequential methods can at large be divided into two different designs. In the strictly sequential procedure, the final sample size is not determined in advance and tests are performed as soon as a new observation is available. On the other hand, in the group sequential test, experimental units are observed in waves, and only a few analyses are performed. The advantages of the group sequential procedure is that it allows for interim analyses and early stopping without the need for continuous monitoring of the accumulating data (Jennison & Turnbull, 1984).

In this thesis, an approach for finding a general approximation to the group sequential test for test statistics with unknown or analytically intractable sampling distributions has been investigated. There is currently no bootstrap version of the group sequential test. The evaluation was conducted with a Monte Carlo simulation study where the achieved significance level was compared to the intended. The results from the simulations suggest that the approach performs well for test statistics with sampling distributions close to a normal distribution.

In the sequential procedure, there is a new risk of conducting a type I error at each time point of testing. The risk arises because as new observations are entering the trial, the sample is overlapping to some extent between interim analyses. In turn, this overlap implies that test statistics observed at different time points are dependent, but not exactly dependent. The variation in the degree of dependence makes it intricate to determine how much the type I error
is inflated. If not handled properly, repeated testing while the data is being collected inflates the false positive risk, the risk of erroneously rejecting the null hypothesis. The critical values in a group sequential test need to be chosen in such a way that the false positive risk for the overall test is the intended (Wassmer, 2000). It is possible to construct conservative tests using the Bonferroni correction in order to maintain the intended type I error (Johnson & Wichern, 2014). However, with several interim analyses, assuming that tests performed at multiple time points are independent, when they in fact are dependent, would result in a decrease in power and an incorrect size for the overall test. For these two reasons, the Bonferroni correction is not considered in this thesis.

In order to obtain the critical values exactly accounted for the correlation between test statistics over time, numerical integration of the distribution functions of the test statistics is required (Armitage, McPherson, & Rowe, 1969). The implementation of the numerical integration procedure requires the sampling distributions of test statistics at each wave of testing to be known or known asymptotically (Kim & Tsiatis, 2020). For test statistics with unknown or analytically intractable sampling distributions, the group sequential test cannot be performed. However, it is often possible to use bootstrap in the non-sequential context to approximate the sampling distribution of test statistics with unknown or analytically intractable sampling distributions (Efron, 1979). In this thesis, an approach for finding a general approximation to the group sequential test that works for arbitrary test statistics is investigated. In the proposed approach, bootstrap is used at each stage of testing to approximate the sampling distribution of the test statistic. The critical values, as derived for a standard normal test statistic, are used to calculate the corresponding alpha value, or significance level, to be used for each test to create bootstrap percentile intervals. If the confidence interval does not cover the value of the null hypothesis, the null is rejected.

The approach implies approximating the covariance structure of the test statistic at each wave of testing, but not the marginal distribution of the test statistic, with that of a standard normal test statistic. However, using the critical values as derived from a standard normal test statistic may provide theoretically incorrect coverage of the confidence interval if the bootstrap distribution is not well approximated by a normal distribution. Even though the approach is not theoretically correct, it may still provide a good approximation. The purpose of this thesis
is to investigate whether the proposed approach provides a good approximation to the group sequential bootstrap test. This thesis aims to answer the following research question:

Can the alpha spent at each interim analysis, as derived for a standard normal test statistic, together with bootstrap provide a good approximation to the group sequential bootstrap test?

The evaluation will be performed with a Monte Carlo simulation study where the achieved false positive rate is compared to the intended.

Section 2 presents an overview of research questions within the field of sequential testing. In Section 3, the implementation of the alpha spending approach for test statistics with known and unknown sampling distributions is introduced. Section 3 also presents the bootstrap percentile interval and how it is used together with the alpha spending approach to perform tests for test statistics with unknown or analytically intractable sampling distributions. The algorithm used for performing the group sequential bootstrap test and the setup for the simulations are presented in Section 4. The results of the Monte Carlo simulations with equal and unequal increments in statistical information are presented in Section 5. Section 6 concludes the thesis.

2 Literature review

Let the random variable $X$ have a probability density function given by $f(x; \theta)$. For testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, the Neyman-Pearson likelihood ratio test statistic is given by:

$$L_N = \prod_{i=1}^{N} \frac{f(x_i; \theta_1)}{f(x_i; \theta_0)}. \quad (1)$$

In the likelihood ratio test, the total sample size is predetermined and the null hypothesis is rejected in favour of the alternative if $L_N > c$. The critical value is denoted by $c$ and determined in such a way that the test has significance level $\alpha$. Of all tests with size less than or equal to $\alpha$, the test is the most powerful test (Kim & Tsiatis, 2020). Based on the Neyman-Pearson lemma, Wald (1947) developed the sequential probability ratio test during the Second World War as a response to an increase in the demand for quality control of anti-aircraft gunnery.

The sequential probability ratio test continuously uses all the available information provided by the observations to determine whether there is sufficient evidence to reject or retain the null.
If there is not sufficient evidence for a decision to be made about rejecting or retaining the null hypothesis, a new observation is gathered and the trial continues (Sébille & Bellissant, 2003). In contrast to the likelihood ratio test, the sequential probability ratio test uses a stopping rule for determining the final sample size (Bartroff et al., 2013). In the sequential probability ratio test, the test statistic for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ is given by:

$$L_n = \prod_{i=1}^{n} \frac{f(x_i; \theta_1)}{f(x_i; \theta_0)}.$$  \hspace{1cm} (2)

For a given number of collected observations, $n$, one of the following decisions is made: continue to gather more observations when $B < L_n < A$ or stop gathering observations. If the accrual of new observations stops, the null is retained if $L_n < B$ or rejected if $L_n > A$. The upper and lower boundaries, $A$ and $B$, are determined by the intended significance level, $\alpha$, and the power of the overall test, $\beta$, and are derived from:

$$A \approx \frac{1 - \beta}{\alpha} \quad B \approx \frac{\beta}{1 - \alpha}. \hspace{1cm} (3)$$

The method developed by Wald (1947) was primarily used in the industrial sector. Armitrage (1954) and Bross (1952) were the first known to have introduced sequential methods for medical trials. Contrary to other advocates of sequential methods, the motivation was ethical concerns rather than reducing the sample size needed for making conclusions about treatment effects in clinical trials. Armitrage (1954) and Bross (1952) emphasized that human suffering could be prevented by early stopping of a trial if there is strong evidence for or against a treatment. If a trial can be terminated early, the time spent with an inferior treatment could potentially be reduced (Mazumdar & Bang, 2007). A disadvantage of the procedure proposed by Wald (1947) is that if no decision can be made at a given stage, new observations can continue to be gathered forever. As a response, Armitrage (1957) imposed a limit to the maximum number of observations entering the trial (Kim & Tsiatis, 2020).

Another disadvantage of the method introduced by Wald (1947) is that with multiple testing on the data as it accumulates, the type I error is inflated. If the critical values from a test with a predetermined sample size are used in a sequential test, the actual significance level of the test will be larger than $\alpha$. This fact is due to that multiple tests are performed with significance level $\alpha$ on the data as it accumulates (Wassmer, 2000). Hence, for each test there is a
new risk of erroneously rejecting the null hypothesis. In order to maintain the overall type I error in repeated testing, Armitage et al. (1969) introduced the repeated significance test as a modification to the sequential probability ratio test and its different variants. The proposed method involves recursive numerical integration of the distribution functions for the test statistics to adjust the critical values for a nominal significance level in such a way that the overall $\alpha$ is maintained. The approach requires the number of interim analyses to be specified before the trial (Bartroff et al., 2013). However, for clinical trials with experimental units assigned to two different treatments, it is not reasonable to continuously perform testing as soon as a new observation has been collected. For this reason, Pocock (1977) proposed a group sequential version of the repeated sequential test.

The general idea of Pocock (1977) was to collect observations in groups rather than single observations. In the procedure by Pocock (1977), a test is performed using all the available observations at the given stage of testing. The approach requires that the maximum number of interim analyses, $K$, to be specified in advance. In addition, the approach requires an equal number of observations in each group, $n$, and that the maximum number of observations collected from each treatment, $N = n \cdot K$, to be determined before the trial. Pocock (1977) showed that constant nominal significance levels as derived for normal responses could be used at each stage of testing to adjust the critical values in such a way that the overall significance level is maintained for a variety of responses. Stricter criteria for early stopping compared to those proposed by Pocock (1977) were later suggested by O’Brien and Fleming (1979), and by Peto et al. (1976).

The approach of O’Brien and Fleming (1979) was stricter than that of Pocock (1977) in the sense that the nominal significance level was increased during the different time points of testing. The given approach implies a greater difficulty in rejecting the null hypothesis in the early stages of testing, but becomes easier as the study prolongs. However, the criteria introduced by Peto et al. (1976) were even stricter for intermediate analyses. In the latter case, the nominal significance level was strict in the early and intermediate stages, such that the nominal significance level at the final test was close to the prespecified significance level.

The methods introduced by Pocock (1977), O’Brien and Fleming (1979), and Peto et al. (1976)
require that the number of observations in each group are equal, such that the accumulated statistical information between two consecutive tests are equal. In his paper, Pocock (1977) suggested that the nominal significance levels for a design with equal group sizes could be used in trials where the number of observations in each group only differs slightly (Jennison & Turnbull, 1984). However, for larger discrepancies in group sizes, the methods proposed by Pocock (1977), O’Brien and Fleming (1979), and Peto et al. (1976) cannot be performed. As a consequence, Slud and Wei (1982) presented a more flexible approach that addressed the problem of unequal group sizes.

In the procedure by Slud and Wei (1982) the overall type I error is arbitrarily divided over each of the tests to maintain the overall type I error, such that $\pi_1 + \ldots + \pi_K = \alpha$. The critical values are derived using the method by Armitage et al. (1969) at each stage of testing in such a way that the probability of erroneously rejecting the null hypothesis at stage $k$ is equal to $\pi_k$. The resulting critical values depend on the observations collected at stage 1 up to stage $k$, but not on the observations collected at stages $k+1$ up until stage $K$. The method by Slud and Wei (1982) enables the time in between tests to differ and contains the general idea of type I error spending. However, the approach has certain limitations.

The disadvantage of the procedure by Slud and Wei (1982) is that it requires the total number of analyses, $K$, and the time points of testing to be specified in advance. If the number of analyses and significance level to be used for each interim analysis are specified in advance, it becomes difficult to adapt the procedure to situations where the accrual of experimental units may be small or very large. For this reason, it may be to preferred that the error is spent in relation to the change in available information to the previous time point of testing. In contrast, for an approach where the spending is based on available information, the type I error can be saved and used for later analyses where more information is available for testing. An alternative to the method by Slud and Wei (1982) for maintaining the type I error in interim analyses with unequal group sizes was suggested by Lan and DeMets (1983).

Lan and DeMets (1983) introduced a spending function for the type I error. The error spending function was formulated as a continuous function of the available information at each stage of testing and depends on current and previous available information. The critical values depend
on the number of analyses conducted and the rate at which the type I error is spent, as indicated by the spending function. The probability for early stopping at calendar time $t$ and the $k$th analysis, for a given spending function, $\alpha^*(\cdot)$, is computed as the amount of the type I error to be spent at time point $t$ minus the type I error already spent in earlier analyses. Specifically, $\alpha^*(t_k) - \alpha^*(t_{k-1}) = \pi_k$ denotes the amount of the type I error that is left to spend at the $k$th analysis. The trial ends with early stopping if the test statistic exceeds the critical values or boundaries corresponding to $\pi_k$. In the approach, only the total or expected total information for the trial needs to be specified. In contrast to the method by Slud and Wei (1982), neither the total number of analyses nor the frequency of the analyses need to be determined in advance.

Since the alpha spending approach was introduced by Lan and DeMets (1983) a variety of alpha spending functions have been formulated in the literature. Notable is the one parameter families of alpha spending functions by Kim and DeMets (1987). The one parameter family of alpha spending functions was later generalised by Hwang, Shih and De Cani (1990) and is presented in Equation 4.

$$\alpha^*_{HSDC}(t_k, \gamma) = \begin{cases} \frac{\alpha(1-e^{-\gamma t_k})}{1-e^{\gamma}}, & \gamma \neq 0 \\ \alpha t_k, & \gamma = 0. \end{cases} \quad (4)$$

For the alpha spending function presented in Equation 4, the parameter $\gamma$ can be set to any value (Jennison & Turnbull, 1999). For $\gamma > 0$, a larger proportion of $\alpha$ is spent in earlier analyses and more importance is given to earlier analyses. On the other hand, for $\gamma < 0$ the type I error is saved for later analyses. The flexibility of the spending function allows for a wide variety of choices for the shape of the boundaries over time.

It is common to use the mean for quantifying the effectiveness of a treatment (Schultzberg & Ankargren, 2022; Liu et al., 2019), mainly because the mean often provides a sufficiently good summary for most metrics. The mean also possesses some theoretical properties that makes it easier to quantify the variance in treatment effects (Schultzberg & Ankargren, 2022; Deng, Longbotham, Walker, & Xu, 2011). The convenient theoretical properties of the mean also imply that it fits well into the two-sample t-test procedure. Even though the mean is a good metric for evaluating treatment effects, other metrics can sometimes be more informative.
In assessing the performance of a treatment, it is often of interest to know who has benefited from the treatment. However, the mean does not provide an informative summary of the tails in the distribution of estimated treatment effects (Schultzberg & Ankargren, 2022; Liu et al., 2019). For example, in evaluating the effects of a new medication for treating high blood pressure, it may be of interest to evaluate the effect of the treatment for those with high blood pressure, not for those defined to have a normal or low blood pressure. Similarly, in online experiments, modifying the user platform to reduce page load time may have the greatest impact on users with the slowest devices or internet connections (Liu et al., 2019). If there is an interest in knowing who or whom has benefited from a treatment, the quantiles may be a more informative metric for comparing treatment effects.

Several methods have been proposed for analysing treatment effects in quantiles using a non-sequential design. Wilcox (1995) proposed an approach for constructing confidence intervals for multiple quantiles. The $(1 - \alpha)100\%$ proposed confidence interval for the difference in quantiles is given by:

$$\hat{y}_p - \hat{x}_p \pm t_{1-\alpha} \sqrt{s_{xp}^2 + s_{yp}^2}.$$  

(5)

In Equation 5, the estimated treatment effect for the experimental method is denoted by $\hat{y}$ and the outcome for the control method by $\hat{x}$. In addition, the respective deciles are denoted by $p$. 

In the procedure of Wilcox (1995), the treatment effects are estimated using the Harrell-Davis (1982) quantile estimator for the given deciles, $p = [0.1, ..., 0.9]$. Moreover, the approach by Wilcox (1995) requires two levels of bootstrap. First, bootstrap is used to estimate the standard error of the treatment effects in each quantile. Second, bootstrap is also used to estimate $t_{1-\alpha}$. The idea of Wilcox was to use the estimated critical value, $\hat{t}_{1-\alpha}$ for normal responses, and also as an approximation for responses with non-normal distributions. The approach by Wilcox (1995) performs well for treatment responses generated from symmetric distributions. However, for treatment responses with heavy tails, the approach was too conservative, resulting in that the achieved type I error being lower than the intended.

The approach considered in this thesis shares some similarities with the paper by Wilcox (1995). The critical values as derived from a normal distribution are also used as an approximation for test statistics with unknown or analytically intractable sampling distributions in this thesis. However, the method by Wilcox (1995) differs from the approach investigated in this
thesis since it applies to the non-sequential design and requires two levels of bootstrap resampling. For large samples, multiple levels of bootstrap can become computationally expensive, which is why the method by Wilcox (1995) is not considered in this thesis. In addition, Wilcox (1995) only focuses on quantiles.

This section has provided a brief introduction to the historical research questions within the field of sequential and group sequential designs. The current theoretical framework of group sequential designs applies for test statistics with known or asymptotically known sampling distributions. In this thesis, a more general approach for performing group sequential tests for test statistics with unknown or analytically intractable sampling distributions is investigated.

3 Theory

In this section, the two theoretical concepts that form the theoretical framework of this thesis are connected. First, the alpha spending approach is presented together with its applications for test statistics with known or asymptotically known sampling distributions. Second, the bootstrap percentile interval is introduced and connected to how it is applied together with the alpha spending approach in this thesis to perform tests for test statistics with unknown or analytically intractable sampling distributions.

3.1 The alpha spending approach

Suppose a group sequential trial is conducted and ended at calendar time $t$, for $t \in [0, T]$, where the scheduled end of the trial is denoted by $T$. At most $K$ analyses are performed. For equal group sizes observed at each interim analysis, the trial includes a maximum of $N = n \cdot k$ observations for each treatment. Furthermore, let the responses from treatment $A$ have mean $\mu_A$ and the responses from treatment $B$ have mean $\mu_B$. Also, let the difference in average treatment effects be denoted by $\theta = \mu_A - \mu_B$. In the trial, a test of $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ is performed at each interim analysis. If the responses from treatments $A$ and $B$ have known common variance $\sigma^2 = 1$, then in many cases, at least approximately for a large sample size by the central limit theorem:

$$Y_k = \frac{\hat{\theta}_k}{\sqrt{2/n}} \sim N \left( \frac{\theta}{\sqrt{2/n}}, 1 \right)$$

(6)
(Kim & Tsiatis, 2020). If the sequence of estimated differences in treatment effects have a multivariate normal distribution, the results of a sequential trial can be expressed in terms of Score or Wald statistics. The Score statistics, $S_k$, and Wald statistics, $Z_k$, are given by:

$$S_k = \sum_{k=1}^{K} Y_k = \sum_{k=1}^{K} \frac{\hat{\theta}_k}{\sqrt{2/n}} \sim N \left( \frac{k\theta}{\sqrt{2/n}}, k \right),$$

$$Z_k = \frac{S_k}{\sqrt{k}} \sim N \left( \frac{\sqrt{k}\theta}{\sqrt{2/n}}, 1 \right).$$

For a group sequential trial, the stopping rules need to satisfy:

$$Pr(\{|S_1| \leq b_1, \ldots, |S_K| \leq b_K\}) = Pr(\{|Z_1| \leq c_1, \ldots, |Z_K| \leq c_K\}) = 1 - \alpha.$$  \hspace{1cm} (9)

In Equation 9, the boundaries are denoted by $b_1, \ldots, b_K$, and the critical values are represented by $c_1, \ldots, c_K$. The boundaries or critical values that defines the stopping rule at each respective interim analysis can be defined by implementing the methods by for example Pocock (1977) or O’Brien and Fleming (1979). However, these methods require that the increments in statistical information, $I(t_1), I(t_2) - I(t_1), \ldots, I(t_K) - I(t_{K-1})$, between each interim analyses are equal. Here, $I(t_k) = n_1 + \ldots + n_k$ represents the total information gathered from the observations collected by the $k$th interim analysis by calendar time $t$. In order to overcome the requirement of equal group sizes, Lan and DeMets (1983) proposed the use of alpha spending functions.

In the alpha spending approach, a spending function is used to describe how the type I error should be spent as a continuous function of information time. The information time, $\tau_k = \frac{I(t_k)}{I(t_K)}$, is defined as the ratio of available information at analysis $k$ by calendar time $t$ divided by the total information for the trial (Bartroff et al., 2013). The rate at which the type I error is spent also defines the stopping criteria for the respective interim analyses. The approach requires the total statistical information or an estimate of such quantity and the alpha spending function to be specified in advance (DeMets & Lan, 1994). The alpha spending functions that approximate those of Pocock (1977) and O’Brien and Fleming (1979) are presented in Equation 10 and 11 (DeMets & Lan, 1994).
\[ \alpha^*_P(\tau) = \begin{cases} \alpha \log(1 + (e - 1)\tau) \\ 0, \tau = 0. \end{cases} \]  
\[ \alpha^*_OBF(\tau) = \begin{cases} 2 - 2\Phi\left(\frac{\tau_{1/2}}{\sqrt{\tau}}\right) \\ 0, \tau = 0. \end{cases} \]

(10) 

In Equation 11 the standard normal distribution function is denoted by \( \Phi \). Lan and DeMets (1983) noted that \( \frac{S_k}{\sqrt{I(t_K)}} \) has the same distribution as a Brownian motion with drift \( \sqrt{I(t_K)}\theta \) observed at information times, \( \tau_k \) for \( k = 1, \ldots, K \), where \( \tau_k = \frac{I(t_k)}{I(t_K)} \) for \( \tau_k \in [0, 1] \) denotes the fraction of the available information at analysis \( k \) by calendar time \( t \) to the total information (Bartroff et al., 2013; Lan & Zucker, 1993). Following the notation by Hamilton (1994), the Brownian motion is defined to be a continuous-time stochastic process which associates each calendar time \( t \in [0, 1] \), such that: \( W(0) = 0 \), for any calendar times \( 0 \leq t_1 < \ldots < t_K \), the differences \( [W(t_2) - W(t_1)], \ldots, [W(u) - W(t)] \sim N(0, u - t) \) and for any given realisation, \( W(t) \) is continuous in \( t \) with probability 1. 

For a group sequential study with unequal increments in statistical information, let \( B(\tau) \) be a Brownian motion. With symmetric boundaries, let \( c_\tau = z_{\alpha/2} \) represent the critical value of interest. Also, let \( s \) denote the first exit time when \( B(\tau) \) exceeds the boundary and the alpha spending function be defined as \( \alpha^*(\tau) = Pr(s \leq \tau) \). The spending function, \( \alpha^*(\tau) \), takes the value 0 if \( \tau = 0 \), and at the end of the trial when \( \tau = 1 \), it takes the value \( \alpha \). In a trial, \( B(\tau) \) will only be observed at discrete information times, \( \tau_k \), for \( k = 1, \ldots, K \). An accumulating probability, \( \alpha^*(\tau_1) \), for \( B(\tau) \) exceeding the boundary at information time \( \tau_1 \) can be derived by finding \( c_1 \) in such a way that:

\[ Pr(B(\tau_1) > c_1) = Pr(s \in [0, \tau_1]) = \alpha^*(\tau_1). \]

(12) 

For the remaining analyses, \( k = 2, \ldots, K \), boundaries \( c_2, \ldots, c_k \) can be found such that:

\[ Pr(B(\tau_2) < c_2, \ldots, B(\tau_{k-1}) < c_{k-1}, B(\tau_k) > c_k) \]

\[ = Pr(s \in [\tau_{k-1}, \tau_k]) = \alpha^*(\tau_k) - \alpha^*(\tau_{k-1}) = \pi_k. \]
In Equation 14, the probability of rejecting the null hypothesis at stage $k$ is represented by $\pi_k$, with $\pi_1 + \ldots + \pi_K = \alpha$, the overall type I error. The first boundary can be computed to be $c_1 = \sqrt{\tau_1} \Phi^{-1}(2\Phi(z_{\alpha/2}/\sqrt{\tau_1}) - 1)$ using the alpha spending function that approximates that of O’Brien and Fleming (1979) (Lan & DeMets, 1983). Finding the remaining boundaries requires the numerical integration procedure by Armitage et al. (1969).

### 3.2 Repeated significance testing with known sampling distribution

For test statistics with known or asymptotically known sampling distributions, the monitoring of a group sequential trial consists of finding boundaries, $b_1, \ldots, b_K$, or critical values, $c_1, \ldots, c_K$. For maintaining the overall intended type I error, the boundaries or critical values need to satisfy:

$$Pr(|S_1| \leq b_1, \ldots, |S_K| \leq b_K) = Pr(|Z_1| \leq c_1, \ldots, |Z_K| \leq c_K) = 1 - \alpha.$$  \hspace{1cm} (15)

Finding boundaries or critical values satisfying Equation 15 equates to solving multivariate integrals on the form:

$$\int_{-b_1}^{b_1} \ldots \int_{-b_K}^{b_K} f(S_1, \ldots, S_K) dS_1 \ldots dS_K = \int_{-c_1}^{c_1} \ldots \int_{-c_K}^{c_K} f(Z_1, \ldots, Z_K) dZ_1 \ldots dZ_K = 1 - \alpha.$$  \hspace{1cm} (16)

However, if the observed sequence of test statistics has independent increments, specifically if:

$$Cov(S_k, S_{k+1}) = Var(S_k) = k \quad \text{or} \quad Cov(S_k, S_{k+1} - S_k) = 0,$$  \hspace{1cm} (17)

the multivariate integrals in Equation 16 become a univariate integral. The sequence of observed test statistics has independent increments if the response variable for the given treatment is only measured once for each observation (Kim & Tsiatis, 2020). With independent increments, the method of Armitage et al. (1969) can be applied to find the boundaries or critical values adjusted for the correlation of test statistics over time in such a way that the overall type I error is the intended (Kim & Tsiatis, 2020).

For a repeated significance test of $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$, with a significance level of 0.05, the trial continues until $|S_k| > b_k$. If the corresponding boundary value $b$ is exceeded at an analysis $k$, $k = 1, \ldots, K$, the null hypothesis is rejected and the gathering of new ob-
servations stops. Furthermore, let $j$ denote the value of $k$ for which the trial ends. In order to compute the probability of exceeding the boundary at a particular analysis $k$, the distribution of $j$ needs to be found. Following the notation by Armitage et al. (1969), let the probability density function of $S_k$ be denoted by $f_k$. The probability density function of $f_k$ can be defined by recursive integration of Equation 18, with $f_1$ being the standard normal density function:

$$f_k(s_k) = \int_{b_{k-1}}^{b_k} f_{k-1}(u) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(s_k-u)^2} du, \text{ for } -b_k \leq s_k \leq b_k. \quad (18)$$

The method proposed by Armitage et al. (1969) in which $f_k$ is evaluated by numerical integration allows the probability of exceeding the bounds for each additional new group of observations to be calculated. The probability of stopping when $k$ interim analyses have been performed or before is given by:

$$Pr_k = Pr(j \leq k) = 1 - Pr(|S_1| \leq b_1, \ldots, |S_k| \leq b_k) = 1 - \int_{-b_k}^{b_k} f_k(u) du. \quad (19)$$

### 3.3 Repeated significance testing with unknown sampling distribution

For the single stage design, suppose a random sample is drawn from an unknown distribution, $F$. Suppose further that the parameter of interest is the mean, denoted by $\theta = s(F)$. The parameter estimate is given by $\hat{\theta} = s(\hat{F})$, with corresponding standard error $\hat{se}$. As the sample size tends to infinity, the distribution of $\hat{\theta}$ becomes closer to that of a normal distribution. For a large sample size, the standardised test statistic $Z$ has an approximately standard normal distribution:

$$Z = \frac{\hat{\theta} - \theta}{\hat{se}} \sim N(0, 1). \quad (20)$$

If the approximation in Equation 20 holds, a $(1 - \alpha)100\%$ confidence interval for $\theta$ is given by:

$$\hat{\theta} \pm z^{1-\alpha/2} \hat{se}. \quad (21)$$

The interval is valid as the sample size tends to infinity, for finite samples it is just an approximation (Tibshirani & Efron, 1993). Moreover, in order to construct the confidence interval, the upper and lower critical values are required.

Even though it may be possible to find the critical values for any test statistic from a sequen-
tial trial with known or asymptotically known sampling distribution, it requires performing a unique numerical integration procedure for every test statistic (Kim & Tsiatis, 2020). Also, some test statistics have analytically intractable sampling distributions. The method by Armitage et al. (1969) requires that the observed test statistic in each interim analysis has a known or asymptotically known sampling distribution. For test statistics with analytically intractable or unknown sampling distributions, the approach by Armitage et al. (1969) cannot be used to find decision regions corrected for the correlation of test statistics over time. However, in the non-sequential setting it is often possible to use bootstrap to approximate the sampling distribution of test statistics with unknown or analytically intractable sampling distributions (Efron, 1979).

In the bootstrap t-approach, \( B \) independent bootstrap samples are obtained from the original data set by random sampling with replacement. Each bootstrap sample consists of \( N \) observations. The parameter estimate of \( \theta \) is computed for each bootstrap sample, \( x^*_{1}, \ldots, x^*_{B} \). Furthermore, let \( \hat{\theta}^*(b) \) represent the estimate of the parameter \( \theta \) obtained from the \( b \)th bootstrap sample \( x^*_b \). It is possible to only use the distribution of \( \hat{\theta}^* \) obtained from the \( B \) bootstrap samples and make inference from that distribution. However, it is common and sometimes recommended to perform a transformation to make the distribution more stable (Tibshirani & Efron, 1993). If a transformation is performed, the standardized test statistic is computed for each bootstrap sample by subtracting the parameter estimate \( \hat{\theta} \) from the estimate of the parameter obtained from the \( b \)th bootstrap sample, \( \hat{\theta}^*(b) \), and divide the difference by the estimated standard error of \( \hat{\theta}^*(b) \), \( \hat{se}^*(b) \).

In order to perform the transformation, the bootstrap standard error of \( \hat{\theta}^* \) is required. For other statistics than the sample mean, the standard error needs to be estimated with bootstrap. Hence, bootstrap resampling is required at two different levels for statistics other than the mean (Tibshirani & Efron, 1993). The bootstrap resampling can become computationally expensive for large sample sizes, which is why in this thesis only the distribution of \( \hat{\theta}^* \) generated from the \( B \) bootstrap samples is considered. Furthermore, an alternative method for constructing confidence intervals that does not require the bootstrap standard error, which will be used in this thesis, is the bootstrap percentile interval.
In the procedure of constructing confidence intervals using the bootstrap percentiles, $B$ bootstrap samples are created by random sampling with replacement from the original sample. The statistic of interest is then calculated for each bootstrap sample $\hat{\theta}^*(b) = s(x^{*b})$, for $b = 1, \ldots, B$. From the ordered list of the $B$ bootstrap replications of $\hat{\theta}^*$, let $\hat{\theta}_B^{(\alpha/2)}$ denote the 100 $\cdot \alpha/2$th percentile of the $\hat{\theta}^*(b)$ values. The corresponding lower and upper bounds for the approximate $1 - \alpha$ percentile interval is then given by:

$$
\hat{\theta}_{\%lower} \approx \hat{\theta}_B^{(\alpha/2)} \quad \text{and} \quad \hat{\theta}_{\%upper} \approx \hat{\theta}_B^{(1-\alpha/2)}.
$$

(22)

The implementation of the alpha spending approach by Lan and DeMets (1983) for performing a group sequential test requires that the sequence of observed test statistics has a known or asymptotically known multivariate normal distribution. Using the critical values from the alpha spending approach to correct for the covariance between test statistics over time assumes that the test statistics have the covariance structure of a Brownian motion. Consequently, the percentiles obtained from the critical values will also be based on the Brownian motion covariance structure. If the percentiles implied by the Brownian motion are used to construct the bootstrap percentile interval and the bootstrap distribution is not well approximated by a normal distribution, the interval will have theoretically incorrect coverage. The theoretically incorrect coverage is caused by the fact that the percentiles are corrected for a normally based covariance structure, but the bootstrap distribution of $\hat{\theta}^*$ for the different time points of testing may not have a multivariate normal distribution.

In this thesis, the covariance structure of test statistics over time is approximated with that of standard normal test statistics. The intervals obtained from the approach will have theoretically incorrect coverage if the bootstrap distribution of the test statistics does not have a normal distribution. However, if the bootstrap distribution is close to that of a normal distribution, the approach may provide a good approximation. This cannot be shown theoretically, which is why the proposed approach in this thesis is evaluated using a Monte Carlo simulation study.

## 4 Monte Carlo simulation

This section presents the algorithm used for implementing the group sequential bootstrap test and the setup for the simulations. The procedure makes use of the alpha spending approach by
Lan and DeMets (1983) and the bootstrap percentile interval by Efron (1979). The algorithm used for the simulations is presented in Algorithm 1. The implementation is made in Julia version 1.7.2 (Bezanson, Edelman, Karpinski, & Shah, 2017). The code for the algorithm can be found here: https://github.com/douglasekstedt/group_sequential_bootstrap_test.

Algorithm 1 The group sequential bootstrap test

**Input:**
- $\tau$: a list of information times at each stage of testing
- $\pi$: a list of critical value inverted alphas
- $dgp$: a data generating process
- $N$: maximum number of observations collected for each treatment in the trial
- $B$: number of bootstrap samples
- $s$: a list of statistics of interest

Generate $Y$ and $X$ under the null hypothesis according to the $dgp$, each of length $N$

for $k$ in 1 : length($\tau$) do
  for $b$ in 1 : $B$ do
    $m \leftarrow$ number of observations from $Y$ observed at analysis $k$, $k \in [1, \text{length}(\tau)]$
    $n \leftarrow$ number of observations from $X$ observed at analysis $k$
    $Y^* \leftarrow$ sample $m$ elements with replacement from $Y$ observed at analysis $k$
    $X^* \leftarrow$ sample $n$ elements with replacement from $X$ observed at analysis $k$
    Compute the vector of statistics of interest, $\hat{\theta}^* = s(Y^*,X^*)$
  end for
  From the bootstrap distribution of the given statistic $\hat{\theta}^*$ obtained from the $B$ bootstrap samples, compute percentile interval using the critical value inverted alpha $[\hat{\theta}^*(\pi[k]/2), \hat{\theta}^*(1−\pi[k]/2)]$
  if $0 \notin [\hat{\theta}^*(\pi[k]/2), \hat{\theta}^*(1−\pi[k]/2)]$ for differences or $1 \notin [\hat{\theta}^*(\pi[k]/2), \hat{\theta}^*(1−\pi[k]/2)]$ for ratios then
    $\text{significant} \leftarrow \text{true}$
    $r + 1$
  else
    $i + 1$
end for

In Algorithm 1, a sample of $N$ observations is generated for each treatment from a data generating process. At each stage of testing, $B$ bootstrap samples of $N \cdot \tau_k$ observations are created for each treatment. Here, $N \cdot \tau_k$ denotes the number of treatment responses observed from each respective treatment at the $k$th interim analysis. The bootstrap distribution is obtained by computing the value of the statistic of interest for each bootstrap sample. The $k$th value in the list of critical value inverted alphas, $\pi_k$, is then used to create the bootstrap percentile interval. If the interval does not cover the value implied by the null, the result of the test is significant. If the result is not statistically significant, $B$ bootstrap samples of size $N \cdot \tau_{k+1}$ are created for each treatment, and the procedure is repeated. The procedure is repeated until a significant result has been obtained or the maximum number of observations collected for the respective treatments has been reached. Before Algorithm 1 is performed, the intended false positive rate $\alpha$, the maximum number of analyses $K$, the maximum number of observations collected for each
treatment $N$, a spending function $\alpha^*$, and a list of information fractions $\tau$ need to be specified.

In the simulations, each trial continues for a maximum of $K = 10$ analyses for an overall intended false positive rate of $\alpha = 0.05$. From the intended false positive rate, the corresponding upper critical values are generated using the alpha spending approach. An implementation of the alpha spending approach by Lan and DeMets (1983) for Julia has been provided by Nordin (2022). From the package, critical values corresponding to the spending functions in the form of uniform, O’Brien and Fleming and Pocock can be obtained. The spending function by Lan and DeMets (1983) that approximates that of O’Brien and Fleming (1979) will be used in the simulations.

The spending function by Lan and DeMets (1983) that approximates that of O’Brien and Fleming (1979) have been chosen since it provides stricter criteria for early stopping of a trial in the early analyses. Furthermore, in Algorithm 1, the upper critical values as implied by the alpha spending function are then used to define the list of corrected alpha values corrected for the covariance structure of test statistics over time, $\pi$. The critical value inverted alphas can be found by solving the integral for each respective analysis, $k \in [1, K]$:

$$\pi_1 = \frac{1}{2\pi} e^{-\frac{1}{2}(z)^2} dz.$$ (23)

In Equation 23 the critical value implied by the Brownian motion covariance structure at the first interim analysis observed at information time $\tau_1$ is denoted by $z(\tau_1)$ and the probability of erroneously rejecting the null hypothesis at the first interim analysis is represented by $\pi_1$. The probability of erroneously rejecting the null hypothesis at an interim analysis $k$ can be computed after an information time for the given analysis has been provided to the alpha spending function.

In the simulations, information times corresponding to equal and unequal increments in statistical information are considered. For equal increments in statistical information the simulations are performed with $\tau_{Equal} = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$ and for unequal using $\tau_{Unequal} = [0.1, 0.15, 0.21, 0.28, 0.3, 0.7, 0.71, 0.87, 0.95, 1.0]$. The alpha spending function by Lan and DeMets (1983) that approximates that of O’Brien and Fleming (1979) is presented in Figure 1 for equal and unequal increments in fractions of statistical information.
Figure 1: The alpha spending function by Lan and DeMets (1983) that approximates that of O’Brien and Fleming (1979) for equal and unequal increments in information times.

The effect of a large increment in the information time on the value of the spending function can be observed in Figure 1. With small increments in statistical information, the test statistics between different interim analyses will almost be perfectly dependent. However, with large increments in statistical information, the test statistics obtained from two consecutive interim analyses will be close to independent. The information times considered in the simulations have been chosen to create variation in the covariance structure. Furthermore, the statistics of interest considered in the simulations are presented together with a definition in Table 1.

Table 1: Statistics considered in the Monte Carlo simulation.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>difference in means</td>
<td>Difference in estimated mean treatment effects, $\bar{Y} - \bar{X}$</td>
</tr>
<tr>
<td>difference in medians</td>
<td>Difference in estimated median treatment effects, $\hat{Y} - \hat{X}$</td>
</tr>
<tr>
<td>difference in 90th percentiles</td>
<td>Difference in estimated 90th percentiles treatment effects, $\hat{p}<em>{Y,90} - \hat{p}</em>{X,90}$</td>
</tr>
<tr>
<td>ratio of means</td>
<td>Ratio of estimated mean treatment effects, $\bar{Y} / \bar{X}$</td>
</tr>
<tr>
<td>difference in variances</td>
<td>Difference in estimated variance of treatment effects, $\hat{s}_Y^2 - \hat{s}_X^2$</td>
</tr>
<tr>
<td>ratio of variances</td>
<td>Ratio of estimated variances of treatment effects, $\hat{s}_Y^2 / \hat{s}_X^2$</td>
</tr>
</tbody>
</table>
The treatment effects are generated under the null hypothesis for the two sample case from the underlying distributions $N(1, 1)$ and $Exp(1)$. For the statistic ratio of means, treatment responses are also generated from the $N(0, 1)$ distribution. In many studies, treatment responses are measured and recorded at multiple time points. In the simulations, treatment responses are only measured and recorded at one time point. The analysis of treatment effects with repeated measurements, as in a longitudinal study, is not considered in this thesis.

Moreover, two different hypotheses are considered for testing the given statistics of interest. For differences, $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ is considered, where $\theta$ is the difference in treatment effects for the statistic of interest. On the other hand, for ratios $H_0 : \theta = 1$ versus $H_1 : \theta \neq 1$ is considered, where $\theta$ is the ratio of treatment effects for the statistic of interest. Furthermore, the maximum number of observations collected for each treatment has been specified to $N = [100; 500; 1,000; 5,000; 10,000]$, and the number of bootstrap samples, $B$, and the number of replications, $R$ is set to 10,000.

In order to quantify the approximation error because of the limited number of replications used, the results for the estimated false positive rate from the simulations are presented together with a normally approximated confidence interval. A two sided 95% normally approximated confidence interval for the Monte Carlo estimated false positive rate is given by:

$$\hat{f}_{pr} \pm 1.96 \cdot \sqrt{\hat{f}_{pr} \cdot (1 - \hat{f}_{pr}) \cdot \frac{1}{R}}.$$  \hspace{1cm} (24)

In Equation 24, the estimated false positive rate, $\hat{f}_{pr}$, denotes the fraction of the total number of significant results to the total number of replications, $R$.

5 Results

This section presents the results from the Monte Carlo simulations with equal and unequal increments in statistical information. The achieved false positive rate is presented for the statistics of interest computed for the different data generating processes and sample sizes considered. The results are evaluated in terms of whether the procedure provides an estimated false positive rate close to the intended false positive rate, $\alpha = 0.05$. 

19
5.1 Equal increments in statistical information

The results from the simulations performed with equal increments in statistical information are presented in Table 2.

Table 2: Estimated false positive rates for equal increments in statistical information.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Statistic</th>
<th>N = 100</th>
<th>N = 500</th>
<th>N = 1,000</th>
<th>N = 5,000</th>
<th>N = 10,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>N(1, 1)</td>
<td>difference in means</td>
<td>0.0649</td>
<td>0.0565</td>
<td>0.0530</td>
<td>0.0588</td>
<td>0.0535</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.060, 0.070)</td>
<td>(0.052, 0.061)</td>
<td>(0.049, 0.057)</td>
<td>(0.054, 0.063)</td>
<td>(0.049, 0.058)</td>
</tr>
<tr>
<td>N(1, 1)</td>
<td>difference in medians</td>
<td>0.0377</td>
<td>0.0432</td>
<td>0.0484</td>
<td>0.0510</td>
<td>0.0488</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.034, 0.041)</td>
<td>(0.039, 0.047)</td>
<td>(0.044, 0.053)</td>
<td>(0.047, 0.055)</td>
<td>(0.045, 0.053)</td>
</tr>
<tr>
<td>N(1, 1)</td>
<td>difference in 90th percentiles</td>
<td>0.0386</td>
<td>0.0409</td>
<td>0.0453</td>
<td>0.0454</td>
<td>0.0493</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.035, 0.042)</td>
<td>(0.037, 0.045)</td>
<td>(0.041, 0.049)</td>
<td>(0.041, 0.049)</td>
<td>(0.045, 0.054)</td>
</tr>
<tr>
<td>N(1, 1)</td>
<td>ratio of means</td>
<td>0.0634</td>
<td>0.0565</td>
<td>0.0531</td>
<td>0.0588</td>
<td>0.0535</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.059, 0.068)</td>
<td>(0.052, 0.061)</td>
<td>(0.049, 0.057)</td>
<td>(0.054, 0.063)</td>
<td>(0.049, 0.058)</td>
</tr>
<tr>
<td>N(0, 1)</td>
<td>ratio of means</td>
<td>0.0181</td>
<td>0.0155</td>
<td>0.0155</td>
<td>0.0164</td>
<td>0.0156</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.015, 0.021)</td>
<td>(0.013, 0.018)</td>
<td>(0.013, 0.018)</td>
<td>(0.014, 0.019)</td>
<td>(0.013, 0.018)</td>
</tr>
<tr>
<td>N(1, 1)</td>
<td>difference in variances</td>
<td>0.0683</td>
<td>0.0546</td>
<td>0.0598</td>
<td>0.0521</td>
<td>0.0511</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.063, 0.073)</td>
<td>(0.050, 0.059)</td>
<td>(0.055, 0.064)</td>
<td>(0.048, 0.057)</td>
<td>(0.047, 0.055)</td>
</tr>
<tr>
<td>N(1, 1)</td>
<td>ratio of variances</td>
<td>0.0683</td>
<td>0.0546</td>
<td>0.0598</td>
<td>0.0521</td>
<td>0.0511</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.063, 0.073)</td>
<td>(0.050, 0.059)</td>
<td>(0.055, 0.064)</td>
<td>(0.048, 0.056)</td>
<td>(0.047, 0.055)</td>
</tr>
<tr>
<td>Exp(1)</td>
<td>difference in means</td>
<td>0.0654</td>
<td>0.0545</td>
<td>0.0535</td>
<td>0.0559</td>
<td>0.0522</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.061, 0.070)</td>
<td>(0.050, 0.059)</td>
<td>(0.049, 0.058)</td>
<td>(0.051, 0.060)</td>
<td>(0.048, 0.057)</td>
</tr>
<tr>
<td>Exp(1)</td>
<td>difference in medians</td>
<td>0.0372</td>
<td>0.0425</td>
<td>0.0457</td>
<td>0.0561</td>
<td>0.0501</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.033, 0.041)</td>
<td>(0.039, 0.046)</td>
<td>(0.042, 0.050)</td>
<td>(0.052, 0.061)</td>
<td>(0.046, 0.054)</td>
</tr>
<tr>
<td>Exp(1)</td>
<td>difference in 90th percentiles</td>
<td>0.0389</td>
<td>0.0423</td>
<td>0.0448</td>
<td>0.0506</td>
<td>0.0503</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.035, 0.043)</td>
<td>(0.038, 0.046)</td>
<td>(0.041, 0.049)</td>
<td>(0.046, 0.055)</td>
<td>(0.046, 0.055)</td>
</tr>
<tr>
<td>Exp(1)</td>
<td>ratio of means</td>
<td>0.0654</td>
<td>0.0545</td>
<td>0.0535</td>
<td>0.0559</td>
<td>0.0522</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.061, 0.070)</td>
<td>(0.050, 0.059)</td>
<td>(0.049, 0.058)</td>
<td>(0.051, 0.060)</td>
<td>(0.048, 0.057)</td>
</tr>
<tr>
<td>Exp(1)</td>
<td>difference in variances</td>
<td>0.0911</td>
<td>0.0689</td>
<td>0.0632</td>
<td>0.0578</td>
<td>0.0543</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.085, 0.097)</td>
<td>(0.064, 0.074)</td>
<td>(0.058, 0.068)</td>
<td>(0.053, 0.062)</td>
<td>(0.050, 0.058)</td>
</tr>
<tr>
<td>Exp(1)</td>
<td>ratio of variances</td>
<td>0.0911</td>
<td>0.0689</td>
<td>0.0632</td>
<td>0.0578</td>
<td>0.0543</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.085, 0.097)</td>
<td>(0.064, 0.074)</td>
<td>(0.058, 0.068)</td>
<td>(0.053, 0.062)</td>
<td>(0.050, 0.058)</td>
</tr>
</tbody>
</table>

Comments: the results are generated using 10,000 replications and 10,000 bootstrap samples. A normally approximated two-sided 95% confidence interval is presented next to each estimated false positive rate in parenthesis. The upper and lower bounds of the confidence intervals are rounded to three decimals.

For the statistics computed from treatment responses generated from the $N(1, 1)$ distribution, it can be observed from Table 2 that the estimated false positive rate is slightly larger than the intended for the statistics difference in means, ratio of means, difference in variances and ratio of variances for $N = 100$. However, as the sample size increases, the estimated false positive rate overall tends to decrease and becomes closer to $\alpha = 0.05$. On the other hand, for the statistics difference in medians and difference in 90th percentiles, the estimated false positive rate is somewhat lower than the intended for $N = 100$, but overall tends towards the intended
false positive rate as the sample size becomes larger. The pattern of the false positive rate becoming closer to the intended for the statistics computed from $N(1, 1)$ distributed responses can also be observed in Figure 2, where the estimated false positive rates is illustrated for the different sample sizes.

![Figure 2: Estimated false positive rates for equal increments in statistical information and normally distributed treatment responses.](image)

The sampling distributions for the statistics computed from treatment responses generated from the $N(1, 1)$ distribution are presented in Appendix A to F. The sampling distributions have been generated using 1,000,000 replications and a sample size of $N = [10; 100; 1,000]$ for each treatment. In addition, a normal probability density function have been fitted to the sampling distributions. From Appendix A to F, it can be observed that the sampling distributions for all statistics computed from treatment responses generated from the $N(1, 1)$ distribution are close to a normal distribution, except for the ratio of means and ratio of variances computed for $N = 10$. However, for larger sample sizes, the sampling distributions for the ratio of means and ratio of variances become close to a normal distribution.

In addition, from Table 2 it can be observed that the estimated false positive rate for the differ-
ence in means computed from treatment responses generated from the $N(1, 1)$ distribution is slightly larger than the intended false positive rate for for $N = 100$. Even tough it theoretically should perform well since the sequence of observed values for the statistic difference in means have a multivariate t-distribution for small sample sizes, but an asymptotically multivariate normal distribution by the central limit theorem. Since the test statistic has an asymptotically normal distribution, the bootstrap distribution should be close to that of a normal distribution. With the bootstrap distribution being close to a normal distribution, the bootstrap percentile should provide approximately correct coverage for the confidence intervals. Consequently, the approach should produce an estimated false positive rate for the difference in means close to the intended. However, it can be observed that estimated false positive rate slightly larger than the intended for $N = 100$. This result indicates that even for the statistic difference in means computed from treatment responses generated from the $N(1, 1)$ distribution, a larger sample size is required for the central limit theorem to apply.

For the statistic ratio of means computed from responses generated from the $N(0, 1)$ distribution, it can from Table 2 be observed that the estimated false positive rate is far below the intended. Moreover, the estimated false positive rate does not become closer to that of the intended as the sample size increases. The result implies that the approach generates too conservative tests for the ratio of means computed from responses generated from the $N(0, 1)$ distribution. In addition, the sampling distribution for the statistic is presented for three different sample sizes in Figure 12 in Appendix G.

From Figure 12 it can be observed that the sampling distribution is centred around zero but also that the distribution has fat tails. The distribution has fat tails since the ratio of two standard normal means has a $Cauchy(0, 1)$ distribution. If the value of the denominator is close to zero, the ratio will become large in absolute value. If the absolute value of the ratio of two means is large, outliers will be created. For a large frequency of outliers, the tails of the bootstrap distribution for the test statistic will become heavy. If the bootstrap distribution has heavy tails, it will not provide a good approximation to a normal distribution. For large discrepancies from a normal distribution, as in this case, the bootstrap percentile interval will provide incorrect coverage. If the bootstrap percentile interval does not provide approximately correct coverage, the approach will not provide an estimated false positive rate close to the intended.
From Table 2 it can be observed that the estimated false positive rate for the statistics computed from treatment responses generated from the $Exp(1)$ distribution is higher than the intended for the statistics difference in means, ratio of means, difference in variances and ratio of variances for $N = 100$. However, as the sample size increases, the estimated false positive rate decreases and becomes closer to the intended. In contrast, for the statistics difference in medians and difference in 90th percentiles, the estimated false positive rate is a bit lower than the intended for $N = 100$, but overall increases towards the intended false positive rate as the sample size becomes larger. The pattern of the estimated false positive rate becoming closer to the intended for the statistics computed from treatment responses generated from the $Exp(1)$ distribution can also be observed graphically in Figure 3.

![Figure 3: Estimated false positive rates for equal increments in statistical information and exponentially distributed treatment responses.](image)

The sampling distributions for the statistics computed from treatment responses with a $Exp(1)$ distribution are presented in Appendix H to M. Similarly to the sampling distributions for the statistics computed from normal responses, the sampling distributions have been generated using 1,000,000 replications and a sample size of $N = [10; 100; 1,000]$ for each treatment. From
Figure 13 to 15, it can be observed that the sampling distributions for the *difference in means*, *difference in medians* and *difference in 90th percentiles* are close to a normal distribution for all sample sizes. Moreover, from Figure 16-18, it can be observed that the sampling distributions for the *ratio of means*, *difference in variances* and *ratio of variances* are not close to a normal distribution for \( N = 100 \). However, for larger sample sizes the sampling distributions for the *ratio of means*, *difference in variances* and *ratio of variances* are close to a normal distribution.

### 5.2 Unequal increments in statistical information

The results from the simulations conducted with unequal increments in statistical information are presented in Table 3.

**Table 3: Estimated false positive rates for unequal increments in statistical information.**

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Statistic</th>
<th>( N = 100 )</th>
<th>( N = 500 )</th>
<th>( N = 1,000 )</th>
<th>( N = 5,000 )</th>
<th>( N = 10,000 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N(1, 1) )</td>
<td><em>difference in means</em></td>
<td>0.0626</td>
<td>0.0517</td>
<td>0.0510</td>
<td>0.0534</td>
<td>0.0536</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.058, 0.067)</td>
<td>(0.047, 0.056)</td>
<td>(0.047, 0.055)</td>
<td>(0.049, 0.058)</td>
<td>(0.049, 0.058)</td>
</tr>
<tr>
<td>( N(1, 1) )</td>
<td><em>difference in medians</em></td>
<td>0.0442</td>
<td>0.0458</td>
<td>0.0469</td>
<td>0.0510</td>
<td>0.0513</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.040, 0.048)</td>
<td>(0.042, 0.050)</td>
<td>(0.043, 0.051)</td>
<td>(0.047, 0.055)</td>
<td>(0.047, 0.056)</td>
</tr>
<tr>
<td>( N(1, 1) )</td>
<td><em>difference in 90th percentiles</em></td>
<td>0.0404</td>
<td>0.0424</td>
<td>0.0430</td>
<td>0.0492</td>
<td>0.0526</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.037, 0.044)</td>
<td>(0.038, 0.046)</td>
<td>(0.039, 0.047)</td>
<td>(0.045, 0.053)</td>
<td>(0.048, 0.057)</td>
</tr>
<tr>
<td>( N(1, 1) )</td>
<td><em>ratio of means</em></td>
<td>0.0601</td>
<td>0.0517</td>
<td>0.0510</td>
<td>0.0534</td>
<td>0.0536</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.055, 0.065)</td>
<td>(0.047, 0.056)</td>
<td>(0.047, 0.055)</td>
<td>(0.049, 0.058)</td>
<td>(0.049, 0.058)</td>
</tr>
<tr>
<td>( N(0, 1) )</td>
<td><em>ratio of means</em></td>
<td>0.0191</td>
<td>0.0158</td>
<td>0.0163</td>
<td>0.0152</td>
<td>0.0135</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.016, 0.022)</td>
<td>(0.013, 0.018)</td>
<td>(0.014, 0.019)</td>
<td>(0.013, 0.018)</td>
<td>(0.011, 0.016)</td>
</tr>
<tr>
<td>( N(1, 1) )</td>
<td><em>difference in variances</em></td>
<td>0.0660</td>
<td>0.0586</td>
<td>0.0579</td>
<td>0.0534</td>
<td>0.0544</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.061, 0.071)</td>
<td>(0.054, 0.063)</td>
<td>(0.053, 0.062)</td>
<td>(0.049, 0.058)</td>
<td>(0.050, 0.059)</td>
</tr>
<tr>
<td>( N(1, 1) )</td>
<td><em>ratio of variances</em></td>
<td>0.0660</td>
<td>0.0586</td>
<td>0.0579</td>
<td>0.0534</td>
<td>0.0544</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.061, 0.071)</td>
<td>(0.054, 0.063)</td>
<td>(0.053, 0.062)</td>
<td>(0.049, 0.058)</td>
<td>(0.050, 0.059)</td>
</tr>
<tr>
<td>( Exp(1) )</td>
<td><em>difference in means</em></td>
<td>0.0666</td>
<td>0.0572</td>
<td>0.0542</td>
<td>0.0539</td>
<td>0.0533</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.062, 0.071)</td>
<td>(0.053, 0.062)</td>
<td>(0.050, 0.059)</td>
<td>(0.049, 0.058)</td>
<td>(0.049, 0.058)</td>
</tr>
<tr>
<td>( Exp(1) )</td>
<td><em>difference in medians</em></td>
<td>0.0385</td>
<td>0.0477</td>
<td>0.0465</td>
<td>0.0513</td>
<td>0.0495</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.035, 0.042)</td>
<td>(0.044, 0.052)</td>
<td>(0.042, 0.051)</td>
<td>(0.047, 0.056)</td>
<td>(0.045, 0.054)</td>
</tr>
<tr>
<td>( Exp(1) )</td>
<td><em>difference in 90th percentiles</em></td>
<td>0.0361</td>
<td>0.0444</td>
<td>0.0411</td>
<td>0.0509</td>
<td>0.0494</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.032, 0.040)</td>
<td>(0.040, 0.048)</td>
<td>(0.037, 0.045)</td>
<td>(0.047, 0.055)</td>
<td>(0.045, 0.054)</td>
</tr>
<tr>
<td>( Exp(1) )</td>
<td><em>ratio of means</em></td>
<td>0.0666</td>
<td>0.0572</td>
<td>0.0542</td>
<td>0.0539</td>
<td>0.0533</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.062, 0.071)</td>
<td>(0.053, 0.062)</td>
<td>(0.050, 0.059)</td>
<td>(0.049, 0.058)</td>
<td>(0.049, 0.058)</td>
</tr>
<tr>
<td>( Exp(1) )</td>
<td><em>difference in variances</em></td>
<td>0.0901</td>
<td>0.0699</td>
<td>0.0621</td>
<td>0.0556</td>
<td>0.0584</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.084, 0.096)</td>
<td>(0.065, 0.075)</td>
<td>(0.057, 0.067)</td>
<td>(0.051, 0.060)</td>
<td>(0.054, 0.063)</td>
</tr>
<tr>
<td>( Exp(1) )</td>
<td><em>ratio of variances</em></td>
<td>0.0901</td>
<td>0.0699</td>
<td>0.0620</td>
<td>0.0556</td>
<td>0.0584</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.084, 0.096)</td>
<td>(0.065, 0.075)</td>
<td>(0.057, 0.067)</td>
<td>(0.051, 0.060)</td>
<td>(0.054, 0.063)</td>
</tr>
</tbody>
</table>

**Comments:** the results are generated using 10,000 replications and 10,000 bootstrap samples. A normally approximated two-sided 95\% confidence interval is presented next to each estimated false positive rate in parenthesis. The upper and lower bounds of the confidence intervals are rounded to three decimals.
For the statistics computed from treatment responses generated from the $N(1, 1)$ distribution in Table 3, the estimated false positive rate overall tends to become closer to the intended rate as the sample size increases. Similar to the results in Table 2 for equal increments in statistical information, the estimated false positive rate is slightly higher for the difference in means, ratio of means, difference in variances and ratio of variances for $N = 100$, but becomes closer to the intended as the sample size increases. For the difference in medians and difference in 90th percentiles on the other hand, the estimated false positive rate is slightly lower than the intended for $N = 100$, but tends to increase as the sample size increases and becomes closer to the intended. Furthermore, the estimated false positive rate does not become close to the intended for the statistic ratio of means computed for treatment responses generated from the standard normal distribution. The pattern of the estimated false positive becoming closer to the intended for the statistics computed from the $N(1, 1)$ can also be observed in Figure 4, where the change of the estimated false positive rate due to different sample sizes is illustrated for treatment responses generated from a normal distribution.

Figure 4: Estimated false positive rates for unequal increments in statistical information and normally distributed treatment responses.
From Table 3, it can also be observed that the estimated false positive rate becomes closer to
the intended for the statistics computed from treatment responses generated from the $Exp(1)$
distribution. The estimated false positive rate is higher than the intended for the statistics difference in means, ratio of means, difference in variances and ratio of variances for $N = 100$. Nevertheless, the estimated false positive rate becomes close to the intended as the sample size increases for the statistics. Moreover, for the statistics difference in medians and difference in 90th percentiles, the estimated false positive rate is a bit lower than the intended for $N = 100$, but overall increases towards the intended false positive rate as the sample size becomes larger.

The overall pattern of the estimated false positive rate becoming closer to the intended as the sample size increases for the statistics computed from treatment responses generated from the $Exp(1)$ distribution can also be observed in Figure 5.

![Figure 5: Estimated false positive rates for unequal increments in statistical information and exponentially distributed treatment responses.](image)

For the simulations performed with unequal increments in statistical information, the estimated false positive rate tends to become closer to the intended for the statistics computed from treatment responses generated from the $N(1, 1)$ and $Exp(1)$ distribution. The results are overall in accordance with the results for the simulations performed with equal increments in statistical
information presented in Table 2. Overall, the resulting estimated false positive rate appears to be unaffected by whether equal or unequal increments in statistical information are considered. In addition, the estimated false positive was a bit further from the intended for the statistics computed from treatment responses generated from the $Exp(1)$ distribution in smaller sample sizes compared to those for the statistics computed from $N(1, 1)$ distributed responses. However, no significant difference can be observed for the estimated false positive rate regarding whether the statistics are computed from treatment responses generated from the $N(1, 1)$ or $Exp(1)$ distribution in larger sample sizes.

6 Conclusion

The benefit of the sequential design is that the duration and sample size required for conducting an experiment may be reduced. Moreover, with these strengths, the sequential design has become popular in online experiments in which fast data-driven decision making is essential for businesses. One of the methods for comparing treatment effects within the sequential design discussed in this thesis is the group sequential test.

The implementation of a group sequential procedure requires the marginal sampling distribution of the test statistic at each interim analysis to be known. For test statistics with unknown or analytically intractable sampling distributions, the group sequential test cannot be performed. In this thesis, an approach for finding a general approximation to the group sequential test that works for arbitrary test statistics has been investigated. The evaluation of the approach was conducted using a Monte Carlo simulation study where the achieved false positive rate was compared to the intended.

The simulations were performed with several different sample sizes and distributions for the treatment responses. Moreover, equal and unequal increments in statistical information were considered. The results suggest that the approach provides an achieved false positive rate close to the intended for test statistics with sampling distributions close to a normal distribution. In addition, no significant difference was found between the simulations performed with equal or unequal increments in statistical information.
The results are representative for the given setup of the simulations concerning the significance level, spending function, total number of analyses, information times, and distributions for the treatment responses. For future studies, the robustness of the approach could be investigated further. It would be of interest to assess how the approach performs for treatment responses generated from more skewed distributions. The treatment responses could for example be generated from a distribution with large extreme values, such as the $\text{Log-normal}(0, 1)$. Another possibility would be to include outliers in the treatment responses and varying the degree of deviation as well as the proportion of the outliers. In addition, it would be of interest to analyze how the approach performs for smaller sample sizes.
Acknowledgements

I would like to thank my supervisor Mattias Nordin and Mårten Schultzberg for their enthusiastic and encouraging support throughout the process of writing this thesis. I am grateful for their insightful comments that helped me to always move forward. To my family and friends, thank you for being there for me.
References


Wassmer, G. (2000). Basic concepts of group sequential and adaptive group sequential test

Appendix A

Figure 6: Sampling distributions for the statistic difference in means computed for treatment responses generated from a Normal(1,1) distribution.
Figure 7: Sampling distributions for the statistic difference in medians computed for treatment responses generated from a Normal(1,1) distribution.
Appendix C

Figure 8: Sampling distributions for the statistic difference in 90th percentiles computed for treatment responses generated from a Normal(1,1) distribution.
Appendix D

Figure 9: Sampling distributions for the statistic ratio of means computed for treatment responses generated from a Normal(1,1) distribution.
Appendix E

Figure 10: Sampling distributions for the statistic difference in variances computed for treatment responses generated from a Normal(1,1) distribution.
Appendix F

Figure 11: Sampling distributions for the statistic ratio of variances computed for treatment responses generated from a Normal(1,1) distribution.
Figure 12: Sampling distributions for the statistic ratio of means computed for treatment responses generated from a Normal(0,1) distribution.
Appendix H

Figure 13: Sampling distributions for the statistic difference in means computed for treatment responses generated from a Exponential(1) distribution.
Appendix I

(a) $\text{Exp}(1)$ for $N = 10$

(b) $\text{Exp}(1)$ for $N = 100$

(c) $\text{Exp}(1)$ for $N = 1,000$

Figure 14: Sampling distributions for the statistic difference in medians computed for treatment responses generated from a Exponential(1) distribution.
Appendix J

Figure 15: Sampling distributions for the statistic difference in 90th percentiles computed for treatment responses generated from a Exponential(1) distribution.
Figure 16: Sampling distributions for the statistic ratio of means computed for treatment responses generated from a Exponential(1) distribution.
Figure 17: Sampling distributions for the statistic difference in variances computed for treatment responses generated from a Exponential(1) distribution.
Appendix M

Figure 18: Sampling distributions for the statistic ratio of variances computed for treatment responses generated from an Exponential(1) distribution.