UNIVERSITY OF CALIFORNIA
RIVERSIDE

# Wilcoxon Rank Sum Tests to Detect One-Sided Mixture Alternatives in Group Sequential Clinical Trials

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

Doctor of Philosophy

in

Applied Statistics

by

Dylan Campbell Friel

June 2022

Dissertation Committee:

Dr. Daniel R. Jeske, Chairperson
Dr. Weixin Yao
Dr. Ramdas Pai

The Dissertation of Dylan Campbell Friel is approved:

---

---

---

Committee Chairperson

University of California, Riverside

## Acknowledgments

I am grateful to my advisor, Prof. Daniel Jeske. Without his help, I would not have been here. I wish to thank Prof. Allan Sampson for opening the doors that set me on this path.

To Sam for all the support.

# ABSTRACT OF THE DISSERTATION

Wilcoxon Rank Sum Tests to Detect One-Sided Mixture Alternatives in Group Sequential Clinical Trials

by

Dylan Campbell Friel

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, June 2022
Dr. Daniel R. Jeske, Chairperson

Group sequential clinical trials offer administrative, economic, and ethical benefits over fixed sample methods when testing a treatment versus control. Traditional methods based on the assumption that the treatment distribution is a pure shift of the control distribution may not always hold. The possibility that an individual from the treatment group may not respond to the treatment motivates the use of a mixture distribution for the treatment group. This work considers two test procedures based on the Wilcoxon Rank Sum statistic for a group sequential design to detect the one-sided mixture alternative. Error spending functions are used for the allocation of error rates at each stage. The two tests are evaluated individually in determination of critical values and arm sizes and asymptotic multivariate normality is shown to hold for both. Upon comparison, the tests are presented to be asymptotically equivalent. Both test statistics maintain the Type I error rate even if $F$ is misspecified in the design alternative. A more general definition of the treatment effect is used with the mixture distribution. Various estimators for the treatment effect are evaluated.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

The setting of this work is that of testing a new treatment versus a control. This may be most simply thought of as developing a new drug to compete with a known drug on the market. Clinical trials are a popular and effective process used to perform these tests while ensuring safety of the participants and general population. Group sequential clinical trials take these ideals a step further by allowing the possibility of terminating a trial early for either efficacy or futility.

Traditional methods of evaluating a treatment versus control assume that the treatment effect should be represented by an additive shift of the control distribution. In this pure shift situation, all treated individuals have responses that come from the shifted distribution. However, this may not always be the case. The proposed methodology seeks to combine these and bring a nonparametric test into the realm of group sequential clinical trials with mixture alternative.

## 1.1 Background

### 1.1.1 Clinical Trials

Clinical trials are a prominent procedure used in experiments to test the efficacy of a new treatment. The steps are deliberate and aim to obtain meaningful results in a prompt yet safe manner, even testing on animals before moving to humans. With the move to testing on humans, the trials enter the first of (at least) three phases. The purpose of Phase I is to establish an appropriate dosage as well as determine the toxicity of a treatment. At this early stage, the small number participants are usually healthy volunteers. Phase II begins the first treatment of diseased patients and analysis of efficacy of the treatment. The number of participants is limited. Phase III is further investigates the efficacy of the treatment and allows many more individuals to participate in the trial. Upon completion of the phases, a final decision may be made about the efficacy of the treatment.

Planning is an essential component in the implementation of clinical trials. This has not only administrative but also statistical benefits. Statistically, it needs to be determined the treatment effect size the experimenters are trying to detect in order to formulate the appropriate hypotheses. Then a suitable test statistic can be chosen from which we would calculate the sample size necessary to detect the alternative hypothesis with a specified amount of power. These sample size calculations are an important part of planning an experiment to ensure the test has sufficient power to detect a difference in the control and treatment groups. This paper will focus on scenarios where the size of the groups are equal.

A common model used for representing a difference between control and treatment is the location-shift model. In this setting, the control group has some density $f(x)$ with

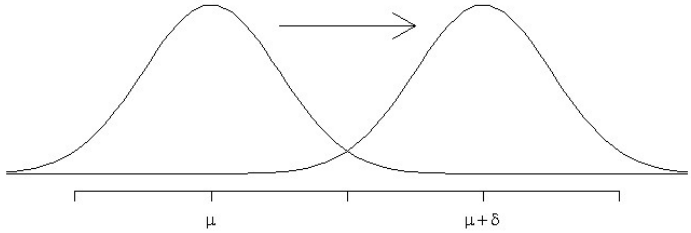


Figure 1.1: A normal density centered at $\mu$ then shifted by parameter $\delta$.

center $\mu$. Then the treatment group has the density $g(x) = f(x-\delta)$. The parameter $\delta$ shifts the density from its center at $\mu$ to the new center $\mu+\delta$. The shape and spread of the density remain unchanged. Figure 1.1 shows a simple example using the normal distribution where the original density is $N(\mu, \sigma^2)$ is shifted by $\delta$ to become $N(\mu + \delta, \sigma^2)$.

This paper will focus on scenarios where a treatment provides improvement over a control. Without loss of generality, the hypotheses will be designed as upper one-sided tests to correspond to larger values exhibiting an effective improvement. In our location shift setting, this will be represented by $\delta > 0$. Two-sided tests would look to show that a treatment does better or worse than the control. A lower one-sided test may be used to determine that a treatment performs at least as well as the control.

Here, we provide a simple example of a test procedure to detect a location-shift. In the fixed sample setting under normal theory, we have a control group $X_1, \ldots, X_m \overset{iid}{\sim} N(\mu_X, \sigma^2)$ and a treatment group $Y_1, \ldots, Y_m \overset{iid}{\sim} N(\mu_Y, \sigma^2)$ where $\sigma^2$ is known. In order to test whether the treatment is an improvement over the control, we can use a simple

difference of two means test using the following hypotheses:

$$H_0 : \mu_Y - \mu_X = 0$$
$$H_A : \mu_Y - \mu_X > 0 \tag{1.1}$$

If we believe the treatment group is a location-shift of the control group then we can rewrite the setup as $X_1, \ldots, X_m \overset{iid}{\sim} N(\mu, \sigma^2)$ and $Y_1, \ldots, Y_m \overset{iid}{\sim} N(\mu + \delta, \sigma^2)$ with the hypotheses that more clearly represent what is being tested:

$$H_0 : \delta = 0$$
$$H_A : \delta > 0 \tag{1.2}$$

When $\sigma^2$ is known, the difference of two means test statistic is

$$Z = \frac{1}{\sqrt{2m\sigma^2}} \left( \sum_{i=1}^{m} Y_i - \sum_{i=1}^{m} X_i \right) = \sqrt{\frac{m}{2\sigma^2}} (\overline{Y} - \overline{X}) \tag{1.3}$$

where $Z \sim N(\delta \sqrt{\frac{m}{2\sigma^2}}, 1)$. We will reject the null hypothesis when $Z > z_\alpha$ for a specified Type I error $\alpha$.

In order to prepare to detect the desired difference for the above test, we can determine the necessary sample size with the specified $\delta$, $\alpha$, and power $(1 - \beta)$ using

$$m = \frac{2\sigma^2 (\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta))^2}{\delta^2} \tag{1.4}$$

where $\Phi(u)$ is the cumulative distribution function (CDF) of the standard normal distribution. Note that we can define $\delta = K\sigma$, then $K$ is the size of the shift in terms of $\sigma$. When using this formulation, Equation (1.4) does not depend on $\sigma$.

### 1.1.2 Wilcoxon Rank Sum

The difference of two means test works well but requires the experimenter to have (or assume they have) normal data. When the distribution of the data is unknown and the sample size is small, it would not be appropriate to use the above test.

If the distribution of the data is known to be non-normal or the distribution of the data is unknown then it would be recommended to use a nonparametric test. Nonparametric test statistics offer an alternative option to the same or similar hypotheses as parametric test statistics while forgoing specific distributional assumptions. A trade-off from requiring less assumptions is the test may have slightly less power than its parametric equivalent [4].

For the location-shift setting, the Wilcoxon Rank Sum test is the recommended nonparametric test [4, 24]. We keep the independent and identically distributed assumptions for the observations within groups and the independence assumption between groups. The only assumption we make for the properties of the distributions are that they are continuous. The samples are $X_1, \ldots, X_m \overset{iid}{\sim} F$ and $Y_1, \ldots, Y_m \overset{iid}{\sim} G$ and the null hypothesis is $H_0 : F = G$ or $H_0 : F(u) = G(u)$ for all $u$.

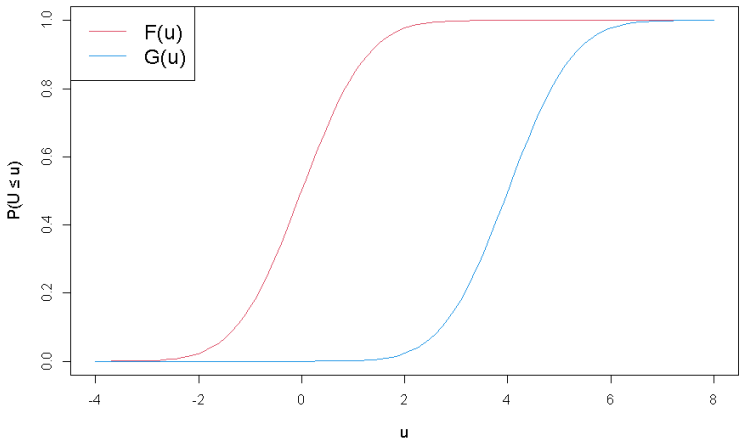


Figure 1.2: An example where $G(u) \leq F(u)$ for the CDFs $F$ and $G$. If $X \sim F$ and $Y \sim G$, then we may say that $Y$ is stochastically larger than $X$.

The Wilcoxon Rank Sum test is designed for the general alternative hypothesis. In the one-sided case, this is represented by either of the following:

$$H_A : F(u) \leq G(u) \text{ for all } u \qquad \text{or} \qquad H_A : F(u) \geq G(u) \text{ for all } u$$
$$F(u) < G(u) \text{ for some } u \qquad\qquad F(u) > G(u) \text{ for some } u$$

The case of $G(u) \leq F(u)$ for all $u$ and $G(u) < F(u)$ for some $u$, shown in Figure 1.2, may also be described by saying that $Y$ is stochastically larger than $X$. Simply, it is the setting where values of $Y$ tend to be larger than values of $X$.

Thus, the location-shift model where $Y = X + \delta$ for some $\delta > 0$ is a subset of this scenario. We can write our hypotheses to show this in our test with

$$\begin{aligned} H_0 &: G(u) = F(u) \qquad \text{for all } u \\ H_A &: G(u) = F(u - \delta) \quad \text{for all } u \text{ and some } \delta > 0 \end{aligned} \tag{1.5}$$

or as in Equation (1.2) above, clearly showing this test can be used for the same hypotheses as used for the difference of two means test.

In order to implement the Wilcoxon Rank Sum test, let $R_i$ be the rank of $Y_i$ in the combined group of the $X$'s and $Y$'s. Then the test statistic is

$$W = \sum_{i=1}^{m} R_i \tag{1.6}$$

and we will reject when $W > w_\alpha$. The exact distribution of $W$ can be determined relatively easily for small sample sizes, but can require much computation for larger sample sizes. There are $\binom{2m}{m}$ possible arrangements of the $X$'s and $Y$'s.

The distribution of $W$ is known to be asymptotically normal for large $m$ [4]. Using the mean, $\frac{m(2m+1)}{2}$, and variance, $\frac{m^2(2m+1)}{12}$, of $W$ for equal sample sizes under the null hypothesis, then

$$\frac{W - m(2m+1)/2}{m^2(2m+1)/12} \quad \xrightarrow{d} \quad N(0,1) \tag{1.7}$$

With the normal approximation, we are able to calculate a sample size $N = m+m$ for the above test. Let $p = P(X < Y)$ and $m = cN$, where in the equal sample size case

$c = 1/2$. Then for a specified $\alpha$ and Type II error $\beta$ we find the upper $z_\alpha$ and $z_\beta$ quantiles of the standard normal distribution. Finally, the sample size for detecting the alternative corresponding to $p$ can be calculated as

$$N = \frac{(z_\alpha + z_\beta)^2}{12c(1-c)(p-\frac{1}{2})^2} \tag{1.8}$$

When using the normal approximation, the experimenter may apply a continuity correction to their test statistic. This may be desired since the distribution of $W$ is discrete, while the normal distribution is continuous.

The asymptotic relative efficiency (ARE) is a measure of comparison between two statistics. An ARE of one signifies the tests perform equivalently, a value less than one indicates worse performance, and greater than one indicates better performance. In the case of testing a location-shift, a possible parametric test would be the two-sample $t$ test. It is known that the ARE when comparing the Wilcoxon Rank Sum test to the two-sample $t$ test is never lower than 0.864. The interpretation of this value is that the Wilcoxon Rank Sum test has at worst 86.4% of the performance of the two-sample $t$ test. The ARE when comparing these two tests is even closer, 0.955, when the distribution of the data is normal. There are even situations when the Wilcoxon Rank Sum test outperforms the $t$ test. Two examples of this occur when the distribution has heavier tails. If the distribution is logistic, then the ARE is 1.09, and if it is Laplace, then the ARE is 1.50 [4].

## 1.2 Mixture Models and Nonresponders

We know that a treatment will not have the same effect on all individuals, and this is represented by the variance of a distribution. However, the possibility that an individual is unaffected by the treatment is not truly present in the location-shift model. Good [5] and Boos and Brownie [1] both show instances of "nonresponders" to a given treatment. Here, nonresponders are individuals that have been given the treatment and were unaffected by it. If an individual is not responding to the treatment, they are considered as those who did not receive the treatment. Therefore, they have responses that come from the same distribution as the control group responses. Good [5] introduces the idea of the treatment distribution being a combination of responders and nonresponders. The proportion of nonresponders is considered unknown, and with their existence, the treatment distribution will be represented by a mixture of the control distribution and a location shift applied to the control distribution.

A mixture distribution can be thought of a as population that is made of subpopulations [11]. When observing the data, we wish to record both the variable of interest, $X$, and its subpopulation, $V$, such that we have the pair $(X_i, V_i)$ for the $i$th individual from the sample, $i = 1, \ldots, m$. Let $\mu_v$ be a parameter with value specific to the $v$th distribution. If the subpopulation from which an observation comes is known then

$$P(X = x|V = v) = f(x; \mu_v) \tag{1.9}$$

However, in practice the subpopulation may be unknown. Let the probability that an observation comes from the $v$th population be $P(V = v) = \theta_v$ such that $\theta_v \geq 0$ and $\sum_v \theta_v = 1$. Then the joint distribution of $(X, V)$ is

$$P(X = x, V = v) = P(X = x|V = v)P(V = v) = f(x; \mu_v)\theta_v \tag{1.10}$$

Finally we consider all the possible subpopulations, and the mixture distribution comes together as

$$h(x; \boldsymbol{\theta}, \boldsymbol{\mu}) = \sum P(X = x|V = v)P(V = v) = \sum \theta_v f(x; \mu_v) \tag{1.11}$$

For the simple two-component mixture model, let $f$ and $g$ be the densities of the two subpopulations with probabilities $(1 - \theta)$ and $\theta$, respectively, $\theta \in (0, 1)$. If $X$ is a random variable from this mixture distribution then we may write

$$X \sim (1 - \theta)f + \theta g \tag{1.12}$$

as a simpler alternative to Equation (1.11). A possible instance of this two-component scenario is pictured in Figure 1.3.

The two-component mixture model is introduced by Good [5] as the distribution for the treatment population. Using Figure 1.3, we may have $f$ as the density of the control group and $g$ as the shift of $f$. Then $\theta$ represents the proportion of responders who come from

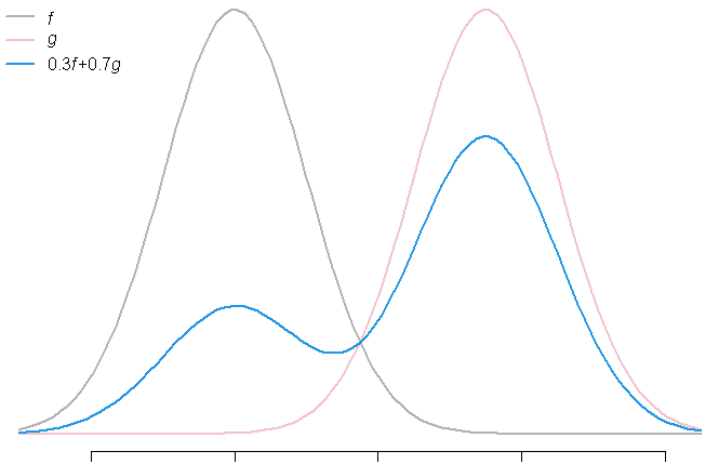


Figure 1.3: A mixture density where the observations could come from subdensity $f$ or subdensity $g$ with $\theta = 0.7$.

the shifted component while the nonresponders come from subdensity $f$ with proportion $(1 - \theta)$.

When under the false assumption that all individuals in the treatment group will respond to the treatment, the test will be underpowered. The sample size required to detect an alternative that has nonresponders is larger than the sample size needed for the pure shift alternative. As seen in Figure 1.4, the difference between the required sample sizes grows quickly as the proportion of nonresponders increases.

### 1.2.1 Parametric Test

Under normal theory, we sample $X_1, \ldots, X_m$ from the control group distribution, $F \equiv N(\mu, \sigma^2)$ and $Y_1, \ldots, Y_m$ from the treatment population whose distribution is $G \equiv$

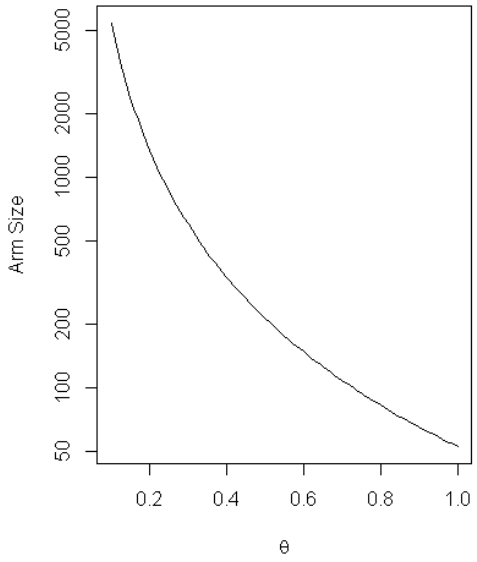


Figure 1.4: Arm size needed to detect a shift as $\theta$ increases, all else held equal.

$(1-\theta)N(\mu,\sigma^2)+\theta N(\mu+\delta,\sigma^2)$. The assumptions in place are that all observations are independent, $\sigma^2$ is known, $\mu \in (-\infty,\infty)$, $\delta > 0$, and $\theta \in (0,1)$. We now write the hypotheses as

$$\begin{aligned} &H_0 : F = G \\ &H_A : F \geq G(u) \text{ for all u, } F > G(u) \text{ for some u} \end{aligned} \tag{1.13}$$

where $G(u) = (1-\theta)F(u) + \theta F(u-\delta)$. For this test, one can use the same test statistic as above for the pure location-shift alternative under normal theory seen in Equation (1.3).

Under the null hypothesis, the test statistic will have the standard normal distribution. Therefore, we will reject the null hypothesis if the test statistic is larger than the upper $\alpha$ quantile of the standard normal distribution. Under the alternative hypothesis for

large $m$, the test statistic is asymptotically distributed as

$$N\left(\theta\delta\sqrt{\frac{m}{2\sigma^2}}, 1+\frac{\theta(1-\theta)\delta^2}{2\sigma^2}\right) \tag{1.14}$$

We can then find the necessary arm size required to detect the alternative for specified $\theta$, $\delta$, Type I error $\alpha$, and Type II error $\beta$. Let $z_\alpha$ and $z_\beta$ be the upper $\alpha$ and upper $\beta$ quantiles of the standard normal distribution, respectively. Then the arm size is

$$m = \frac{(\sqrt{2}z_\alpha + \sqrt{2+\theta(1-\theta)(\delta/\sigma)^2}z_\beta)^2}{\theta^2(\delta/\sigma)^2} \tag{1.15}$$

### 1.2.2 Nonparametric Test

Jeske and Yao [9] investigated the use of the Wilcoxon Rank Sum test statistic with a mixture alternative. Moving out of normal theory and into nonparametrics means we will sample from some continuous density so $X_1, \ldots, X_m \sim F$ and $Y_1, \ldots, Y_m \sim G$. The hypotheses will be the same as above, i.e. we are testing if $G(u) = (1-\theta)F(u) + \theta F(u-\delta)$. They use the standardized test statistic

$$Z = \frac{W - m(2m+1)/2}{\sqrt{m^2(2m+1)/12}} \tag{1.16}$$

where $W$ is the Wilcoxon Rank Sum test statistic. Since the null hypothesis and test statistic are the same as the standardized Wilcoxon Rank Sum test statistic for the pure location-shift alternative, this test statistic has an asymptotic standard normal distribution for large $m$ under $H_0$.

Jeske and Yao state that under the mixture alternative, the test statistic has the following limiting normal distribution

$$\frac{\sqrt{2m}}{m^2}\Big(W - (m(m\gamma(F,G) + (m+1)/2))\Big) \quad \xrightarrow{d} \quad N\left(0, \frac{\xi_1(F,G)}{\lambda} + \frac{\xi_2(F,G)}{1-\lambda}\right) \tag{1.17}$$

where

$$\gamma(F,G) = P(X_1 < Y_1) \qquad \xi_1(F,G) = P(X_1 < Y_1, X_1 < Y_2) - \gamma^2(F,G)$$

$$\lambda = \lim_{(2m)\to\infty} m/(2m) < 1 \qquad \xi_2(F,G) = P(X_1 < Y_1, X_2 < Y_1) - \gamma^2(F,G)$$

With the mixture alternative $G(u) = (1-\theta)F(u) + \theta F(u-\delta)$, they rewrite $\gamma(F,G) \equiv \gamma(\theta,\delta,F)$, $\xi_1(F,G) \equiv \xi_1(\theta,\delta,F)$, and $\xi_2(F,G) \equiv \xi_2(\theta,\delta,F)$. Using the asymptotic distribution, Jeske and Yao [9] show that the formula for determining the sample size necessary to detect the mixture alternative is

$$m = \left(\frac{z_\alpha\sqrt{(\rho+1)/12\rho} + z_\beta\sqrt{\xi_1(\theta,\delta,F) + \xi_2(\theta,\delta,F)/\rho}}{\gamma(\theta,\delta,F) - 1/2}\right)^2 \tag{1.18}$$

where $\rho$ is the ratio of the arm sizes. In the scenario where the arm sizes are equal, the above equation simplifies to

$$m = \left(\frac{z_\alpha/\sqrt{6} + z_\beta\sqrt{\xi_1(\theta,\delta,F) + \xi_2(\theta,\delta,F)}}{\gamma(\theta,\delta,F) - 1/2}\right)^2 \tag{1.19}$$

| | $K$ | | | |
|---|---|---|---|---|
| $\theta$ | 0.25 | 0.5 | 0.75 | 1.0 |
| 0.5 | 730 | 189 | 89 | 54 |
| 0.6 | 507 | 131 | 62 | 37 |
| 0.7 | 372 | 96 | 45 | 27 |
| 0.8 | 285 | 73 | 34 | 20 |
| 0.9 | 225 | 58 | 27 | 16 |
| 1.0 | 182 | 46 | 21 | 13 |

Table 1.1: Standardized Logistic fixed sample sizes for $\alpha = 0.05$ and 80% power; $\delta = K\sigma$.

In Table 1.1, we see arm sizes needed for various combinations of $\theta$ and shift size using the Wilcoxon Rank Sum test statistic with a mixture alternative as determined by Equation (1.19). The change in the required sample size depending on $\theta$ is easily seen, highlighting the importance of using the mixture alternative when appropriate.

Jeske and Yao [9] focus on location-scale families where the CDF $\Psi(u)$ has zero mean and unit variance and the corresponding density function is $\psi(u)$. Then the $X$ observations have the distribution $F(u) = \Psi((u-\mu_F)/\sigma_F)$ and could be equivalently defined as $X = \mu_F + \sigma_F Z$, where $Z \sim \Psi$. In this way, $\mu_F$ and $\sigma_F$ can be interpreted as the mean and standard deviation for any choice of $F$. The treatment group may then be written as $Y = (1-\theta)[\mu_F + \sigma_F Z] + \theta[\mu_F + \delta + \sigma_F Z]$.

## 1.3 Group Sequential Clinical Trials

Group sequential clinical trials further break down a phase from a clinical trial into several more stages. Instead of doing a single fixed sample experiment, data is analyzed at a preset number of times during the study. It is common for the study to have 2-5 stages. New individuals are recruited for each stage. For the first stage, a test statistic is calculated

and used to determine whether the null hypothesis should be rejected, accepted, or if there is not enough evidence to reject/accept and the experiment should continue to the next stage. At each stage after the first, a test statistic is calculated based on the observations gathered at the current stage as well as all previous stages. Then the decision of whether to reject, accept, or continue the experiment is made. At the final stage, there is only the decision to reject or accept the null hypothesis. It is common in this setting to use the "accept" terminology instead of "fail to reject" in regard to the null hypothesis in this setting.

Implementing group sequential methods provides several benefits over fixed sample experiments. If the treatment turns out to be unsafe for patients, the experiment may be ended at an early stage with fewer participants exposed. Even if the treatment is not unsafe but there is early evidence that it is not effective, the trial may end earlier than fixed sample methods. This would allow research to move in a new direction to discover a treatment that is effective. On the other hand, if there is evidence that the treatment is particularly effective, it may get released to the public sooner. These ethical benefits of early stopping are accompanied by economical benefits. Ending at an early stage would mean less costs are wasted if the treatment is ineffective and money could be earned sooner if it is effective. Finally, evaluation at several points throughout the study means it may be closely monitored to ensure everything is operating as planned.

The setup for the group sequential hypothesis tests will be the same as the fixed sample setting except for there will be $2mS$ observations in total where $m$ is the arm size for each group per stage and $S$ is the total number of stages. The potential control group

sample is $X_1, \ldots, X_{mS}$ and the potential treatment group sample is $Y_1, \ldots, Y_{mS}$. Because the number of stages is random, it is possible that not all potential observations will be realized.

To illustrate the group sequential trial testing procedure, we may consider a case where the treatment group observations come from a pure shift of the control group distribution. An example test statistic could be the following

$$Z_s = \frac{1}{\sqrt{2ms\sigma^2}} \left( \sum_{i=1}^{ms} Y_i - \sum_{i=1}^{ms} X_i \right) \tag{1.20}$$

where $S$ is the total number of stages and $s = 1, \ldots, S$.

In this setting, we may reject when the test statistic is larger than an upper critical value or fail to reject if the test statistic is smaller than a lower critical value. This could happen at any stage and would end the experiment early if it happens at a stage prior to stage $S$. Seen in Figure 1.5, the critical values $r_1, \ldots, r_{S-1}, a_1, \ldots, a_{S-1}$, and $u$ need to be determined. Here, the $r_i$'s, $i = 1, \ldots, S-1$, represent the upper critical values at which we would reject the null hypothesis, the $a_i$'s, $i = 1, \ldots, S-1$, represent the lower critical values at which we would accept the null hypothesis, and $u$ is the critical value for the final stage. After stage $s = 1, \ldots, S-1$, we make the following comparisons and corresponding

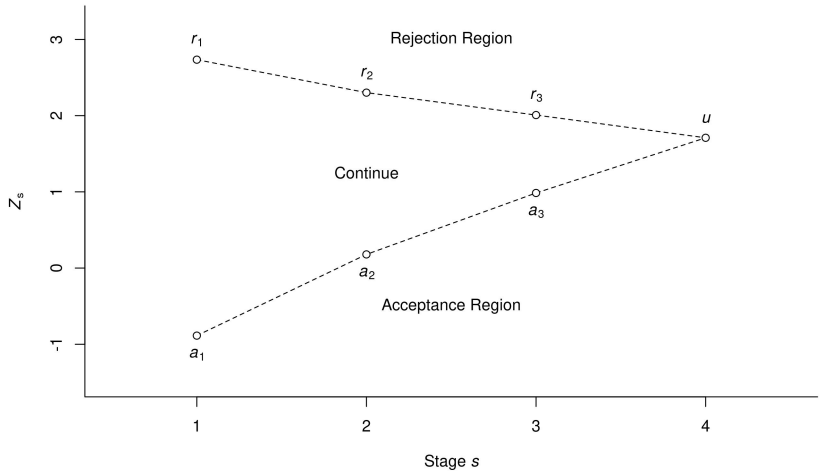


Figure 1.5: An example of the critical values for the test statistic of a 4-stage group sequential trial.

decisions

$$\text{if } Z_s \geq r_s \qquad \text{reject, end experiment} \tag{1.21}$$

$$\text{if } Z_s \leq a_s \qquad \text{accept, end experiment} \tag{1.22}$$

$$\text{otherwise} \qquad \text{continue to next stage} \tag{1.23}$$

After stage $S$

$$\text{if } Z_S \geq u \qquad \text{reject, end experiment} \tag{1.24}$$

$$\text{if } Z_S < u \qquad \text{accept, end experiment} \tag{1.25}$$

The critical values are found using both the Type I error ($\alpha$) and Type II error ($\beta$) set by the experimenter at the beginning of the study and the joint distribution of $(Z_1, \ldots, Z_S)$ under both the null and alternative hypotheses. The Type I and Type II error should be decided for the overall experiment as well as per stage. There are various ways to split the per stage Type I and Type II error. We will use error spending functions established by Jennison and Turnbull [8].

We define functions $f(t)$ and $g(t)$ to set the Type I and Type II errors at each stage, respectively. These nondecreasing functions are chosen such that they are equal to zero for $t = 0$ and $f(t) = \alpha$ and $g(t) = \beta$ for $t \geq 1$. Following Jennison and Turnbull, we have

$$f(t) = \min(\alpha, \alpha t^{\rho}) \qquad\qquad g(t) = \min(\beta, \beta t^{\rho}) \tag{1.26}$$

for $t \in (0, 1)$ and $\rho > 0$. With equal increases in the arm sizes, $t$ is defined at stage $s$ as $t_s = s/S$. The experimenter may choose $\rho = 1$ for equal spending at each stage and larger $\rho$ to have smaller Type I and Type II error for early stages. Values such as $\rho = 2$ or $\rho = 3$ may be a desired strategy in order to make it more likely to avoid an early false positive or false negative.

The resulting Type I and Type II errors for each stage are calculated as follows

$$\alpha_1 = f(t_1) \qquad\qquad \beta_1 = g(t_1) \tag{1.27}$$

$$\alpha_s = f(t_s) - f(t_{s-1}) \qquad\qquad \beta_s = g(t_s) - g(t_{s-1}) \tag{1.28}$$

where $s = 2, \ldots, S$. Then the overall Type I and Type II errors for the entire study are $\alpha = \sum_{s=1}^{S} \alpha_s$ and $\beta = \sum_{s=1}^{S} \beta_s$

Once the desired Type I and Type II errors are set, the critical values are found using the joint distribution of $(Z_1, \ldots, Z_S)$ under both the null and alternative hypotheses. They are chosen to satisfy the following set of equations

$$P_{H_0}(Z_1 \geq r_1) = \alpha_1 \tag{1.29}$$

$$P_{H_A}(Z_1 \leq a_1) = \beta_1 \tag{1.30}$$

For $s = 2, \ldots, S-1$,

$$P_{H_0}(a_1 < Z_1 < r_1, \ldots,\ a_{s-1} < Z_{s-1} < r_{s-1},\ Z_s \geq r_s) = \alpha_s \tag{1.31}$$

$$P_{H_A}(a_1 < Z_1 < r_1, \ldots,\ a_{s-1} < Z_{s-1} < r_{s-1},\ Z_s \leq a_s) = \beta_s \tag{1.32}$$

Finally, at stage $S$,

$$P_{H_0}(a_1 < Z_1 < r_1, \ldots,\ a_{S-1} < Z_{S-1} < r_{S-1},\ Z_S \geq u) = \alpha_S \tag{1.33}$$

$$P_{H_A}(a_1 < Z_1 < r_1, \ldots,\ a_{S-1} < Z_{S-1} < r_{S-1},\ Z_S < u) = \beta_S \tag{1.34}$$

The critical values are best found in pairs beginning with $r_1$ and $a_1$ then moving through subsequent stages until the final stage is reached.

### 1.3.1 Parametric Test

Peng et al. [16] developed a group sequential test for detecting a one-sided mixture alternative with a location-shift component under normal theory. The procedure assumes that the control consists of $N(\mu, \sigma^2)$ random variables and the treatment come from the mixture distribution $(1-\theta)N(\mu, \sigma^2) + \theta N(\mu+\delta, \sigma^2)$ where $\theta \in (0,1)$, $\delta > 0$, and $\sigma^2$ is known. The test statistic is the same as the one used in Equation (1.20). For large $m$, the joint distribution is multivariate normal

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_S \end{pmatrix} \xrightarrow{d} N_S \left( \begin{pmatrix} \theta\delta\sqrt{m/(2\sigma^2)} \\ \theta\delta\sqrt{2m/(2\sigma^2)} \\ \vdots \\ \theta\delta\sqrt{Sm/(2\sigma^2)} \end{pmatrix}, \left(1 + \frac{\theta(1-\theta)\delta^2}{2\sigma^2}\right) \begin{pmatrix} 1 & \sqrt{1/2} & \dots & \sqrt{1/S} \\ \sqrt{1/2} & 1 & \dots & \sqrt{2/S} \\ & & \ddots & \\ \sqrt{1/S} & \sqrt{2/S} & \dots & 1 \end{pmatrix} \right) \tag{1.35}$$

R code is available for the user to input the number of stages, overall Type I error, overall Type II error, $\rho$ for the error spending function, the proportion of the population that will respond to treatment, and the desired shift in terms of standard deviations from the mean. The function in R uses the asymptotic multivariate distribution and numerical integration to determine and output the critical values and arm size to be used per group at each stage. Table 1.2 displays arm sizes for various scenarios calculated using this R code.

| | | $K$ | | | |
|---|---|---|---|---|---|
| $\theta$ | Stages | 0.25 | 0.5 | 0.75 | 1.0 |
| 0.5 | 1 | 794 | 200 | 90 | 52 |
| | 2 | 414 | 105 | 47 | 27 |
| | 3 | 284 | 72 | 33 | 19 |
| | 4 | 216 | 55 | 25 | 14 |
| 0.6 | 1 | 551 | 139 | 63 | 36 |
| | 2 | 288 | 73 | 33 | 19 |
| | 3 | 197 | 50 | 23 | 13 |
| | 4 | 150 | 38 | 17 | 10 |
| 0.7 | 1 | 405 | 102 | 46 | 27 |
| | 2 | 212 | 54 | 24 | 14 |
| | 3 | 145 | 37 | 17 | 10 |
| | 4 | 110 | 28 | 13 | 8 |
| 0.8 | 1 | 310 | 78 | 35 | 20 |
| | 2 | 162 | 41 | 19 | 11 |
| | 3 | 111 | 28 | 13 | 8 |
| | 4 | 85 | 22 | 10 | 6 |
| 0.9 | 1 | 245 | 62 | 28 | 16 |
| | 2 | 128 | 32 | 15 | 9 |
| | 3 | 88 | 22 | 10 | 6 |
| | 4 | 67 | 17 | 8 | 5 |
| 1 | 1 | 198 | 50 | 22 | 13 |
| | 2 | 104 | 26 | 12 | 7 |
| | 3 | 71 | 18 | 8 | 5 |
| | 4 | 54 | 14 | 6 | 4 |

Table 1.2: Arm sizes calculated using `GSDMix()` for $\alpha = 0.05$ and 80% power with $\rho = 2$ and $\delta = K\sigma$.

### 1.3.2 Nonparametric Tests

There is not currently a nonparametric test designed for testing a mixture alternative in group sequential clinical trials. However, there are various existing nonparametric tests designed for other scenarios in group sequential clinical trials. Wilcoxon, Rhodes, and Bradley [25] developed two methods for testing the Lehmann alternative. Spurrier and Hewett [21] use the Wilcoxon Rank Sum statistic in group sequential methods to detect the general alternative $F < G$ in situations with only two stages. Madsen and Hewett [14] create an alternative rank-based statistic to test pure shift alternatives in multiple stages. Lee and DeMets [10] create a nonparametric test for group sequential designs with repeated measurements on individuals. Su and Lachin [22] present work with multivariate observations. Using the test statistic from Spurrier and Hewett [21] in an arbitrary number of stages, Shuster, Chang, and Tian [19] focus on ordinal categorical data. Yuan, Zheng, Huang and Tan [26] allow for the consideration of covariate information in their test statistic. Huang and Tan [7] develop methods for the experimenter interested in multiple primary endpoints. In this work, we look to develop an $S$-stage group sequential clinical trial test to detect the mixture alternative with some continuous $F$.

The rest of the dissertation is organized as follows. In Chapter 2, we introduce our proposed test procedure including discussion of the test statistic, its distribution, the normal approximation of its distribution, and computation of arm size. The power and robustness properties are also examined. Chapter 3 investigates a comparison of the test statistic used by Shuster et al. [19] and the test statistic we introduce in Section 2. Chapter 4 introduces a multiplicative treatment effect in the mixture setting as an alternative to the

location-shift model. In Chapter 5, we explore estimation of the treatment effect in both the location-shift and multiplicative treatment effect mixture settings. We conclude with a brief summary and considerations for future work in Chapter 6.

# Chapter 2

# Sequential Average Rank

The purpose of this work is to develop nonparametric ranking techniques to be used in group sequential clinical trials where the distribution of the treatment group is a mixture distribution. As we look to build upon the work of Jeske and Yao [9], an issue to be determined is how to combine information across the stages in the statistic. Here we examine the proposed group sequential method, the Sequential Average Rank procedure.

## 2.1 Test Statistic

The Sequential Average Rank (SAR) test statistic is calculated by finding the Wilcoxon Rank Sum statistic for each stage using only the observations within that stage and averaging these statistics across stages. We will focus on the scenario where the number of observations, $m$, are equal between groups and stages. Let $X_{11}, \ldots, X_{1m}, \ldots, X_{S1}, \ldots, X_{Sm} \overset{iid}{\sim} F$ represent potential observations from the control group and $Y_{11}, \ldots, Y_{1m}, \ldots, Y_{S1}, \ldots, Y_{Sm} \overset{iid}{\sim} G$ represent potential observations from the treatment group. Because

the number of stages is random, it is possible that not all potential observations will be realized. We are interested in testing the null hypothesis $H_0 : F = G$ against the one-sided alternative, $H_A : G$ is the mixture distribution. We take $G(u) = (1-\theta)F(u) + \theta F(u-\delta)$ where $\theta \in (0,1)$ is the proportion of responders and $\delta = K\sigma_F$ is the shift size for $K > 0$ and $\sigma_F > 0$. With $\sigma_F$ defined as the standard deviation of $F$, $K$ represents the size of the shift in the terms of number of standard deviations. For identifiability, define the null case as the point $(\theta, \delta) = (0, 0)$. For $R_{s1}, \ldots, R_{sm}$, the ranks of the $Y$ observations at stage $s$, we get the Wilcoxon Rank Sum statistic at stage $s$, $W_s = \sum_{j=1}^{m} R_{sj}$.

With equal arm sizes for each group and at each stage, the mean and variance under the null hypothesis of each $W_s$ is

$$\mu = \frac{m(2m+1)}{2} \qquad\qquad \sigma^2 = \frac{m^2(2m+1)}{12} \tag{2.1}$$

Using these moments, we can easily calculate the standardized test statistic we desire. The SAR test statistic at stage $s = 1, \ldots, S$ is

$$Z_s = \frac{\frac{1}{s}\sum_{i=1}^{s} W_i - \mu}{\sigma/\sqrt{s}} \tag{2.2}$$

The marginal mean and variance at each stage are zero and one, respectively, under the null hypothesis. Since the $W_i$'s are independent, the covariance of the SAR under the null

hypothesis between stages $s$ and $s'$ for $s \leq s'$, can be found in the following way

$$\begin{aligned}\mathrm{Cov}(Z_s, Z_{s'}) &= \mathrm{Cov}\left(\frac{\frac{1}{s}\sum_{i=1}^{s} W_i - \mu}{\sigma/\sqrt{s}},\ \frac{\frac{1}{s'}\sum_{j=1}^{s'} W_j - \mu}{\sigma/\sqrt{s'}}\right) = \frac{\sqrt{ss'}}{ss'\sigma^2}\sum_{i=1}^{s}\mathrm{Var}(W_i) \\ &= \frac{s\sigma^2\sqrt{ss'}}{ss'\sigma^2} = \sqrt{\frac{s}{s'}} \end{aligned} \tag{2.3}$$

In order to determine the mean and variance of the SAR under the alternative hypothesis, we define

$$\begin{aligned} p &= P(X < Y) \\ p_1 &= P(X_1 < Y_1,\ X_1 < Y_2) \\ p_2 &= P(X_1 < Y_1,\ X_2 < Y_1) \end{aligned} \tag{2.4}$$

When testing the mixture alternative $G(u) = (1-\theta)F(u) + \theta F(u-\delta)$, these probabilities can be written as the following one-dimensional integrals:

$$\begin{aligned} p &= \int_{-\infty}^{\infty}\int_{-\infty}^{y} f(x)g(y)\ dx\ dy = \int_{-\infty}^{\infty} F(u)\ dG(u) \\ &= \int_{-\infty}^{\infty} F(u)[(1-\theta)f(u) + \theta f(u-\delta)]\ du \end{aligned} \tag{2.5}$$

$$p_1 = \int_{-\infty}^{\infty}[1-G(u)]^2\ dF(u) = \int_{-\infty}^{\infty}[1-(1-\theta)F(u) - \theta F(u-\delta)]^2 f(u)\ du \tag{2.6}$$

$$p_2 = \int_{-\infty}^{\infty}[F(u)]^2\ dG(u) = \int_{-\infty}^{\infty} F^2(u)[(1-\theta)f(u) + \theta f(u-\delta)]\ du \tag{2.7}$$

where $f(u)$ is the probability density function of $F$.

The mean and variance for the Wilcoxon Rank Sum statistic with equal arm sizes under the alternative hypothesis are

$$\mu_A = m\left(mp + \frac{(m+1)}{2}\right) \tag{2.8}$$

$$\sigma_A^2 = m^2\big(p(1-p) + (m-1)(p_1 - p^2) + (m-1)(p_2 - p^2)\big) \tag{2.9}$$

Note that under the null hypothesis, $p = 1/2$ and $p_1 = p_2 = 1/3$, and using these probabilities result in the previous mean and variance [4].

These lead us to the mean, variance, and covariance of the SAR under the alternative hypothesis.

$$\mathrm{E}_A[Z_s] = \mathrm{E}_A\left[\frac{\frac{1}{s}\sum_{i=1}^{s} W_i - \mu}{\sigma/\sqrt{s}}\right] = \frac{\mu_A - \mu}{\sigma/\sqrt{s}} \tag{2.10}$$

$$\mathrm{Var}_A(Z_s) = \mathrm{Var}_A\left(\frac{\frac{1}{s}\sum_{i=1}^{s} W_i - \mu}{\sigma/\sqrt{s}}\right) = \frac{\sigma_A^2}{\sigma^2} \tag{2.11}$$

$$\mathrm{Cov}_A(Z_s, Z_{s'}) = \mathrm{Cov}_A\left(\frac{\frac{1}{s}\sum_{i=1}^{s} W_i - \mu}{\sigma/\sqrt{s}},\ \frac{\frac{1}{s'}\sum_{j=1}^{s'} W_j - \mu}{\sigma/\sqrt{s'}}\right) = \frac{\sigma_A^2}{\sigma^2}\sqrt{\frac{s}{s'}} \tag{2.12}$$

## 2.2 Monte Carlo Simulation of the Exact Distribution

The exact joint distribution of $(Z_1, \ldots, Z_S)$ under the null hypothesis does not depend on the distribution of the data since it inherits the distribution-free properties of the Wilcoxon Rank Sum statistic. Therefore, the SAR test statistic is nonparametric in this regard. The joint distribution is needed in order to determine the critical values and arm size required to detect the alternative for a given level of power. Under alternative hypothesis, the joint distribution depends on the choice of $F$. Furthermore, the joint distribution and, therefore, the critical values depend on $m$ under both the null and alternative hypotheses. The proper solutions to Equations (1.29)-(1.34) cannot be achieved until $m$ is large enough.

Determining the full joint distribution can be computationally intensive, although we can use simulation to approximate the distribution and obtain the appropriate critical values and arm size. The algorithm used to determine the critical values and arm size when simulating the joint distribution is visualized in the flowchart in Figure 2.1. It follows the general procedure established in Section 1.3.

The algorithm starts at an arm size $m = m_0$. This can easily be chosen to be $m = 0$, and the algorithm will eventually reach the necessary arm size to detect the given design alternative. However, the arm sizes for small shifts can be large, resulting in many iterations of the algorithm when increasing $m$ by one. Through empirical evidence, the arm sizes for two stages are found to be roughly half of the fixed sample arm size, the arm sizes for three stages are roughly one third of the fixed sample arm size, for four stages are roughly one fourth, etc. This phenomenon can be seen in Table 1.2 and persists for our method. Using this observation along with the fixed sample arm size formula from Jeske

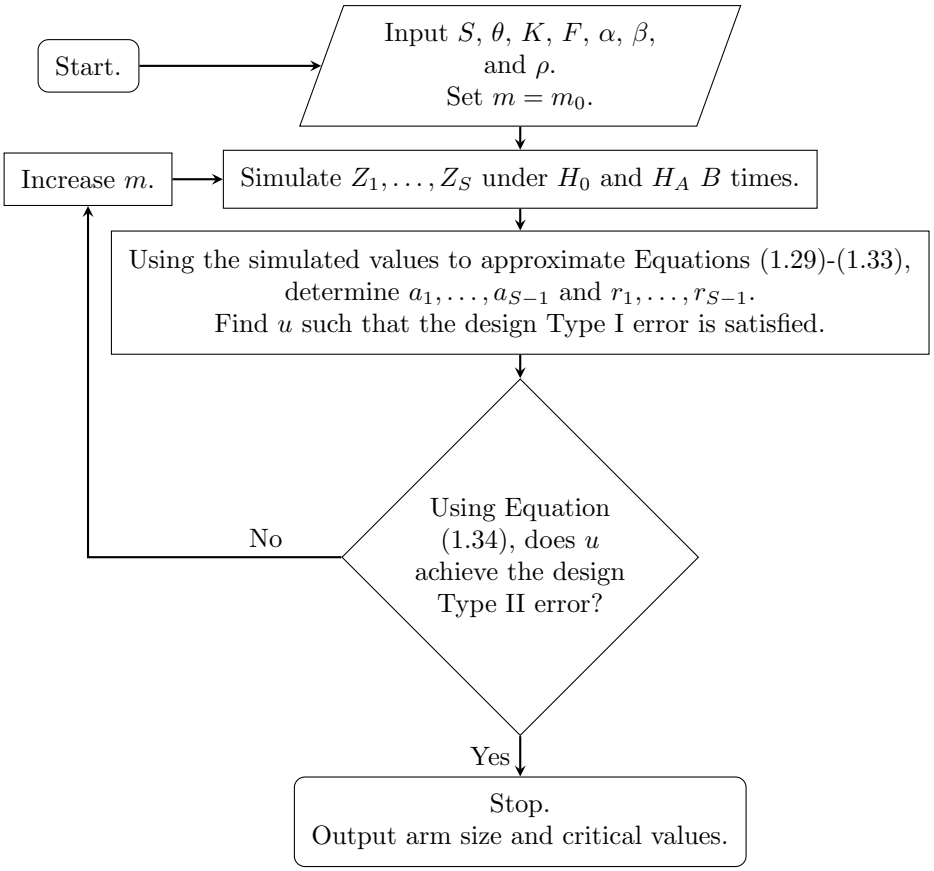


Figure 2.1: Flow chart of the algorithm used to find critical values when simulating the exact joint distribution of the test statistic.

and Yao [9], one can quickly find an improved starting value. This approach is implemented in the algorithm to start $m = m_0$ with this approximate value instead of zero to reduce the number of iterations to achieve the final result.

The exact joint and marginal distributions of the SAR are discrete with the number of possible values increasing as the arm size increases. In situations with small arm sizes and especially at early stages, it may not be possible to achieve the design Type I error or Type II error. In these cases, the code running the algorithm is designed to assign an upper critical value $+\infty$ or lower critical value $-\infty$ rather than have an inflated Type I or Type II error. Consider a two-stage scenario where $r_1 = +\infty$, and we have some attainable values for $a_1$ and $u$. Thus the resulting test would be impossible to reject at stage 1, while still leaving the possibility of accepting the null hypothesis at the first stage. These situations are uncommon. They are more likely to arise with the Type I error, since it is often much smaller than the Type II error, and depend on the error spending function, proportion of responders, and shift size. When a critical value is assigned $\pm\infty$, the Type I or Type II error is neither spent nor pushed to a future stage. Therefore, the overall error rates will be strictly less than the values use for the design alternative.

The discrete nature of the distribution of the SAR test statistic makes achieving the exact Type I and Type II errors impossible in most situations. Because of this, the algorithm will result in critical values that correspond to Type I and Type II errors that are less than or equal to the design values.

## 2.3 Multivariate Normal Approximation

The Wilcoxon Rank Sum statistic at stage $s$, $W_s$, used for the SAR is calculated using only the observations at stage $s$. Therefore $W_s$ and the Wilcoxon Rank Sum statistic calculated at stage $s'$ are independent. Furthermore, since the arm sizes are the same at stage $s$ and $s'$, $W_s$ and $W_{s'}$ are identically distributed. Under the null hypothesis, the distribution of $W_s$ can be approximated by a normal distribution for large $m$ [4]. Working under this assumption, we are interested in the limiting distribution of the SAR at stage $s$, $Z_s$. It can be shown that the sum of independent normal random variables has a normal distribution [17]. First, let $Y_i = \frac{W_i - \mu}{\sigma}$, then $Y_i \sim N(0,1)$. Now $Z_s = \frac{1}{\sqrt{s}} \sum_{i=1}^{s} Y_i$. Let $M_{Z_s}$ be the moment generating function of $Z_s$ and $M_{Y_i}$ be the moment generating function of the $Y_i$'s. Using the moment generating functions we get

$$M_{Z_s}(t) = M_{\sum Y_i/\sqrt{s}}(t) = \prod_{i=1}^{s} M_{\frac{1}{\sqrt{s}}Y_i}(t) = \prod_{i=1}^{s} M_{Y_i}(\frac{t}{\sqrt{s}}) \tag{2.13}$$

$$= \left[ M_{Y_1}(\frac{t}{\sqrt{s}}) \right]^s \tag{2.14}$$

$$= \left[ e^{t^2/2s} \right]^s \tag{2.15}$$

$$= e^{t^2/2} \tag{2.16}$$

which is the moment generating function of a $N(0,1)$ random variable, $s = 1, \ldots, S$.

Now we wish to consider the joint distribution of the $Z_s$'s under the null hypothesis. It is possible to show that a vector of random variables has a multivariate normal distribution by showing that linear functions of the random variables are univariate normal [15]. Let $\vec{\mathbf{Y}} = (Y_1, \ldots, Y_S)$, where $Y_s$ is defined as above for $s = 1, \ldots, s$, then $\vec{\mathbf{Y}}$ has the following

$S$-variate normal distribution

$$\vec{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_S \end{bmatrix} \xrightarrow{d} N_S(\vec{\mathbf{0}}, \mathbf{I}_S) \tag{2.17}$$

since $\vec{\mathbf{c}}\vec{\mathbf{Y}}$ is univariate normal for all fixed $S$-vectors $\vec{\mathbf{c}}$, where $\boldsymbol{I}_S$ is the $S \times S$ identity matrix.

The SAR can be written in vector form as

$$\vec{\mathbf{Z}} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 & \dots & 0 \\ & & & \vdots & \\ 1/\sqrt{S} & 1/\sqrt{S} & 1/\sqrt{S} & \dots & 1/\sqrt{S} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_S \end{bmatrix} = \mathbf{M}\vec{\mathbf{Y}} \tag{2.18}$$

Then

$$\vec{\mathbf{Z}} \xrightarrow{d} N_S(\mathbf{M}\vec{\mathbf{0}}, \mathbf{M}\mathbf{I}_S\mathbf{M'}) \equiv N_s(\vec{\mathbf{0}}, \mathbf{M}\mathbf{M}^\top) \tag{2.19}$$

Thus under the null hypothesis for large $m$, the SAR has a multivariate normal distribution. This holds for any conformable matrix $\mathbf{M}$. Note this is regarding constants, such as the case that $\mathbf{M}$ is a zero matrix, as degenerate forms of the normal distribution.

The Wilcoxon Rank Sum statistic is known to follow a normal distribution asymptotically under the general alternative [4]. Following the same procedure as above under the alternative hypothesis, the distribution of the SAR statistic has a limiting multivariate nor-

mal distribution. Now, each $Y_s$ will have mean $(\mu_A - \mu)/(\sigma/\sqrt{s})$ and variance $\sigma_A^2/\sigma^2$. Using the same $\mathbf{M}$ as above, the distribution of $\vec{\mathbf{Z}}$ under the alternative hypothesis can be approximated by a normal distribution with mean vector $\mathbf{M}((\mu_A - \mu)/\sigma)\vec{\mathbf{1}}$ and variance-covariance matrix $\sigma_A^2/\sigma^2\mathbf{M}\mathbf{M}^\top$.

Using the normal approximation to obtain the critical values and arm size for a given alternative will follow a similar algorithm as simulating the exact joint distribution, seen in Figure 2.2. This algorithm will utilize the same approach to determine a starting value $m_0$ as the algorithm for the simulation of the exact distribution. Instead of approximating Equations (1.29)-(1.34) empirically using the simulation, we now may use the multivariate normal densities.

In R, we can use the `mvtnorm` [3] package for the distribution function of a multivariate normal random variable. Its `pmvnorm()` function allows us to find the cumulative probability for a candidate critical value at the current stage. Therefore, we use this in combination with a root finding function to find the critical value that satisfies the design Type I or Type II error.

## 2.4 Comparison of Exact Distribution and Normal Approximation

When simulating the exact joint distribution of the text statistic, the goal is to achieve the true critical values. However, the joint distribution can be extensive. With such a vast distribution that grows with the arm size, it can take much time and resources to

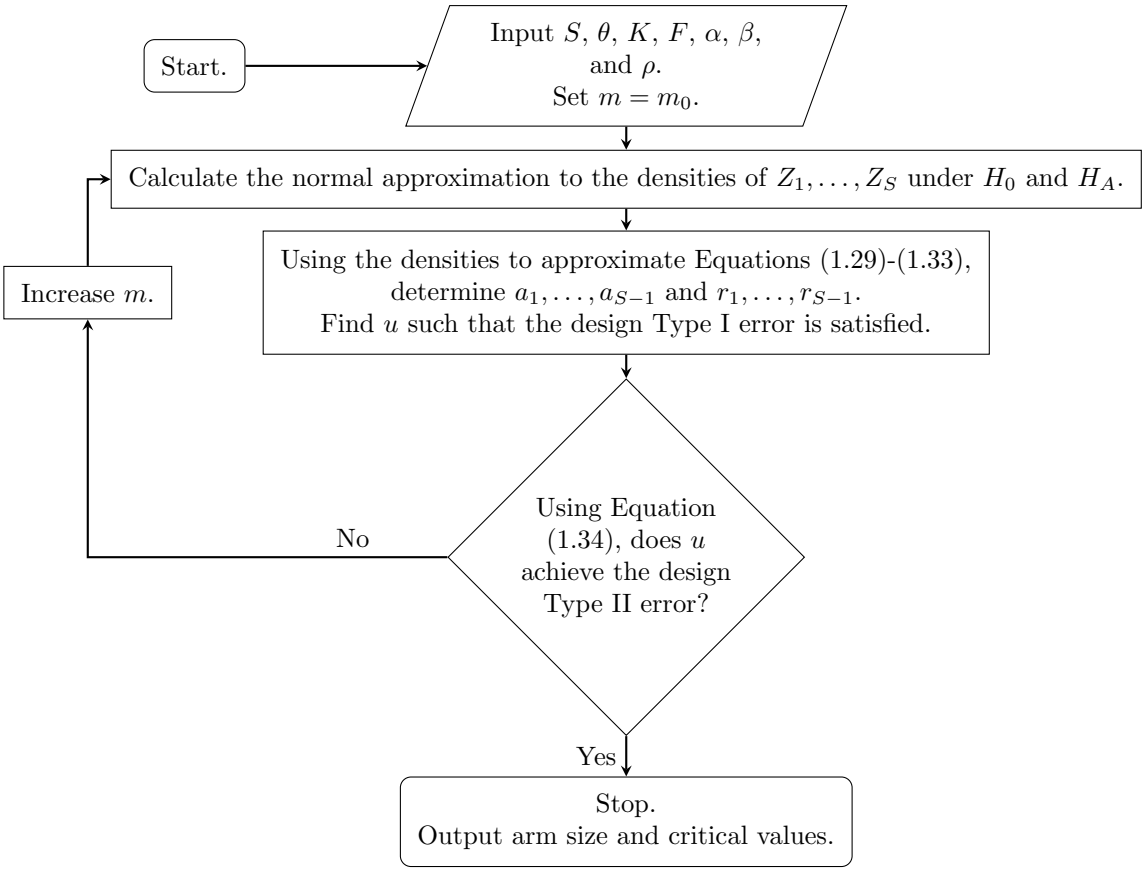


Figure 2.2: Flow chart of the algorithm used to find critical values and arm size when using the multivariate normal approximation of the joint distribution of the SAR.

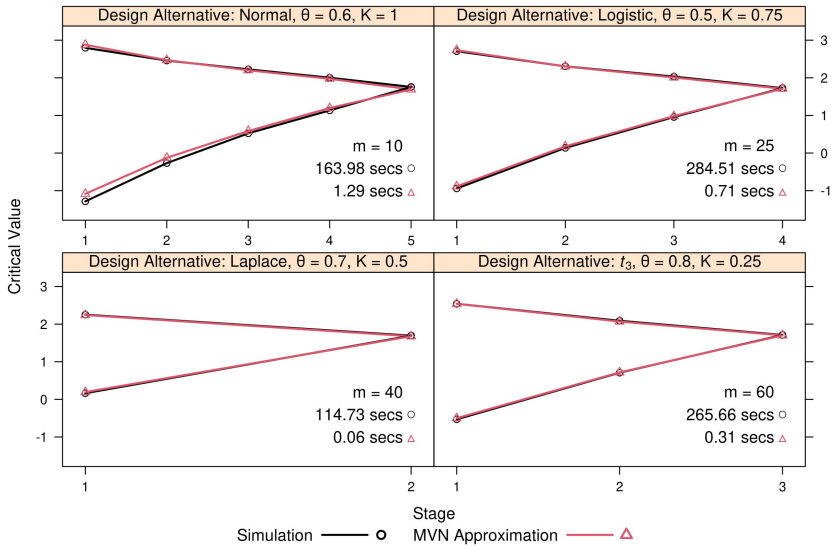


Figure 2.3: Comparison of algorithms for the SAR test statistic for various scenarios using 100,000 simulations for the exact simulation with $\alpha = 0.05$, $\beta = 0.2$, and $\rho = 2$.

simulate. It quickly becomes possible that the user is not simulating the entire exact distribution depending on the arm size and number of simulations. Simulating an appropriate number of times to get a good representation of the distribution is a concern. Without a large enough number, simulation of the exact distribution is itself an approximation.

The multivariate approximation offers a great increase in computation speed. Figure 2.3 highlights the computation time difference. Similarities in the solutions between the two algorithms are also apparent in Figure 2.3. They recommend the same arm size in each case and the critical values become virtually indistinguishable for large $m$.

It would be recommended in most cases to use the normal approximation. This algorithm gives a great increase in speed while still offering the same arm sizes and often virtually identical critical values. It can be done without the uncertainty of running sufficient simulations to represent the distribution.

## 2.5 Properties

We wish to show that there are some properties of the SAR that make it even more useful and competitive with existing methods, namely the test statistic from Peng et al. [16].

### 2.5.1 Arm Sizes

The arm sizes needed to detect the mixture alternative using both algorithms of the SAR for various combinations of number of stages, $\theta$, $\delta = K\sigma_F$, and $F$ are shown in Tables 2.1-2.4. The fixed sample arm sizes using the normal approximation in these tables were determined by the algorithm in Figure 2.2 to be consistent with the rest of the formulation of the tables, rather than using Equation (1.19).

Arm sizes from the simulation of the exact distribution were obtained by running the simulation three times and taking the largest arm size of the three runs. The variation in the arm size from this method is perhaps due to the number of simulations. This may be an explanation for some differences between the algorithms, especially in large arm size situations.

The focus here is on standardized location-scale family distributions following the definitions from Jeske and Yao mentioned in Section 1.2.2. For the normal distribution, $\Psi$ is the CDF of the standard normal distribution. The logistic distribution will have $\Psi(u) = (1 + \exp(-cu))^{-1}$ with $c = \pi/3$. The Laplace distribution will use $\Psi(u) = (e^{cu}/2)I(u < 0) + (1 - e^{-cu}/2)I(u \geq 0)$ with $c = \sqrt{2}$. Finally, the $t$ distribution with 3 degrees of freedom

| | | $K$ | | | |
|---|---|---|---|---|---|
| $\theta$ | Stages | 0.25 | 0.5 | 0.75 | 1.0 |
| 0.5 | 1 | 837 (837) | 215 (215) | 101 (100) | 60 (60) |
| | 2 | 437 (437) | 113 (113) | 53 (53) | 32 (32) |
| | 3 | 299 (299) | 78 (77) | 37 (36) | 22 (22) |
| | 4 | 228 (228) | 59 (59) | 28 (28) | 17 (17) |
| | 5 | 185 (184) | 48 (48) | 23 (23) | 14 (14) |
| 0.6 | 1 | 581 (580) | 149 (149) | 69 (69) | 42 (41) |
| | 2 | 304 (303) | 78 (78) | 37 (37) | 22 (22) |
| | 3 | 208 (208) | 54 (54) | 26 (25) | 16 (15) |
| | 4 | 159 (158) | 41 (41) | 20 (20) | 12 (12) |
| | 5 | 128 (128) | 34 (34) | 16 (16) | 10 (10) |
| 0.7 | 1 | 426 (426) | 110 (109) | 51 (50) | 30 (30) |
| | 2 | 223 (223) | 58 (58) | 27 (27) | 16 (16) |
| | 3 | 153 (153) | 40 (40) | 19 (19) | 12 (11) |
| | 4 | 117 (117) | 31 (30) | 15 (15) | 9 (9) |
| | 5 | 95 (94) | 25 (25) | 12 (12) | 8 (7) |
| 0.8 | 1 | 327 (326) | 84 (83) | 39 (38) | 23 (22) |
| | 2 | 171 (171) | 44 (44) | 21 (21) | 13 (12) |
| | 3 | 117 (117) | 31 (30) | 15 (14) | 9 (9) |
| | 4 | 90 (89) | 24 (23) | 12 (11) | 7 (7) |
| | 5 | 73 (72) | 19 (19) | 10 (9) | 6 (6) |
| 0.9 | 1 | 258 (257) | 66 (65) | 30 (30) | 18 (17) |
| | 2 | 135 (135) | 35 (16) | 17 (16) | 10 (10) |
| | 3 | 93 (93) | 24 (24) | 12 (11) | 7 (7) |
| | 4 | 71 (71) | 19 (19) | 9 (9) | 6 (6) |
| | 5 | 58 (57) | 16 (15) | 8 (7) | 5 (5) |
| 1 | 1 | 210 (208) | 53 (53) | 25 (24) | 15 (14) |
| | 2 | 110 (109) | 28 (28) | 14 (13) | 8 (8) |
| | 3 | 75 (75) | 20 (20) | 10 (9) | 6 (6) |
| | 4 | 58 (57) | 15 (15) | 8 (7) | 5 (5) |
| | 5 | 47 (47) | 13 (12) | 7 (6) | 4 (4) |

Table 2.1: Arm sizes needed to detect the mixture alternative using the SAR where $F$ is the standard normal distribution for $\alpha = 0.05$ and 80% power with $\rho = 2$ determined with 1,000,000 simulations to represent the exact joint distribution. Arm sizes determined by the normal approximation are in parentheses.

| | | $K$ | | | |
|---|---|---|---|---|---|
| $\theta$ | Stages | 0.25 | 0.5 | 0.75 | 1.0 |
| 0.5 | 1 | 732 (731) | 190 (189) | 90 (90) | 55 (54) |
| | 2 | 382 (382) | 99 (99) | 47 (47) | 29 (29) |
| | 3 | 262 (261) | 69 (68) | 33 (33) | 21 (20) |
| | 4 | 199 (199) | 52 (52) | 25 (25) | 16 (16) |
| | 5 | 162 (161) | 43 (42) | 21 (20) | 13 (13) |
| 0.6 | 1 | 508 (507) | 132 (132) | 62 (62) | 38 (38) |
| | 2 | 265 (265) | 69 (69) | 33 (33) | 20 (20) |
| | 3 | 182 (182) | 48 (48) | 23 (23) | 14 (14) |
| | 4 | 139 (139) | 37 (36) | 18 (18) | 11 (11) |
| | 5 | 112 (112) | 30 (30) | 15 (14) | 9 (9) |
| 0.7 | 1 | 374 (373) | 97 (96) | 45 (45) | 28 (28) |
| | 2 | 195 (195) | 51 (51) | 24 (24) | 15 (15) |
| | 3 | 134 (134) | 35 (35) | 17 (17) | 11 (10) |
| | 4 | 102 (102) | 27 (27) | 14 (13) | 8 (8) |
| | 5 | 83 (83) | 22 (22) | 11 (11) | 7 (7) |
| 0.8 | 1 | 286 (285) | 74 (74) | 35 (34) | 21 (21) |
| | 2 | 149 (149) | 39 (39) | 19 (19) | 11 (11) |
| | 3 | 102 (102) | 27 (27) | 13 (13) | 8 (8) |
| | 4 | 78 (78) | 21 (21) | 10 (10) | 7 (6) |
| | 5 | 64 (63) | 17 (17) | 9 (8) | 6 (5) |
| 0.9 | 1 | 225 (225) | 58 (58) | 27 (26) | 17 (16) |
| | 2 | 118 (118) | 31 (31) | 15 (15) | 9 (9) |
| | 3 | 81 (81) | 22 (21) | 11 (10) | 7 (6) |
| | 4 | 62 (62) | 17 (17) | 8 (8) | 6 (5) |
| | 5 | 51 (50) | 14 (14) | 7 (7) | 5 (4) |
| 1 | 1 | 182 (182) | 47 (47) | 22 (21) | 14 (13) |
| | 2 | 96 (96) | 25 (25) | 12 (12) | 8 (7) |
| | 3 | 66 (66) | 18 (17) | 9 (8) | 6 (5) |
| | 4 | 51 (50) | 14 (13) | 7 (7) | 5 (4) |
| | 5 | 41 (41) | 11 (11) | 6 (6) | 4 (4) |

Table 2.2: Arm sizes needed to detect the mixture alternative using the SAR where $F$ is the logistic distribution for $\alpha = 0.05$ and 80% power with $\rho = 2$ determined with 1,000,000 simulations to represent the exact joint distribution. Arm sizes determined by the normal approximation are in parentheses.

| | | $K$ | | | |
|---|---|---|---|---|---|
| $\theta$ | Stages | 0.25 | 0.5 | 0.75 | 1.0 |
| 0.5 | 1 | 546 (546) | 149 (149) | 74 (74) | 48 (48) |
| | 2 | 286 (285) | 78 (78) | 40 (39) | 26 (25) |
| | 3 | 196 (195) | 54 (54) | 27 (27) | 18 (18) |
| | 4 | 149 (149) | 41 (41) | 21 (21) | 14 (14) |
| | 5 | 121 (121) | 34 (33) | 17 (17) | 12 (11) |
| 0.6 | 1 | 379 (379) | 103 (103) | 51 (51) | 33 (33) |
| | 2 | 199 (198) | 54 (54) | 28 (27) | 18 (18) |
| | 3 | 136 (136) | 38 (37) | 19 (19) | 13 (12) |
| | 4 | 104 (104) | 29 (29) | 15 (15) | 10 (10) |
| | 5 | 84 (84) | 24 (23) | 12 (12) | 8 (8) |
| 0.7 | 1 | 279 (278) | 76 (76) | 38 (38) | 24 (24) |
| | 2 | 146 (146) | 40 (40) | 20 (20) | 13 (13) |
| | 3 | 100 (100) | 28 (28) | 14 (14) | 9 (9) |
| | 4 | 77 (76) | 22 (21) | 11 (11) | 8 (7) |
| | 5 | 62 (62) | 18 (17) | 9 (9) | 7 (6) |
| 0.8 | 1 | 213 (213) | 58 (58) | 29 (29) | 18 (18) |
| | 2 | 112 (112) | 31 (31) | 16 (15) | 10 (10) |
| | 3 | 77 (77) | 21 (21) | 11 (11) | 8 (7) |
| | 4 | 59 (59) | 17 (16) | 9 (8) | 6 (6) |
| | 5 | 48 (48) | 14 (13) | 7 (7) | 5 (5) |
| 0.9 | 1 | 169 (168) | 46 (45) | 23 (22) | 15 (14) |
| | 2 | 89 (88) | 25 (24) | 12 (12) | 8 (8) |
| | 3 | 61 (61) | 17 (17) | 9 (9) | 6 (6) |
| | 4 | 47 (46) | 13 (13) | 7 (7) | 5 (5) |
| | 5 | 38 (38) | 11 (11) | 6 (6) | 4 (4) |
| 1 | 1 | 136 (136) | 37 (37) | 18 (18) | 12 (11) |
| | 2 | 72 (72) | 20 (20) | 10 (10) | 7 (6) |
| | 3 | 49 (49) | 14 (14) | 7 (7) | 5 (5) |
| | 4 | 38 (38) | 11 (11) | 6 (6) | 4 (4) |
| | 5 | 31 (31) | 9 (9) | 5 (5) | 4 (3) |

Table 2.3: Arm sizes needed to detect the mixture alternative using the SAR where $F$ is the Laplace distribution for $\alpha = 0.05$ and 80% power with $\rho = 2$ determined with 1,000,000 simulations to represent the exact joint distribution. Arm sizes determined by the normal approximation are in parentheses.

| | | $K$ | | | |
|---|---|---|---|---|---|
| $\theta$ | Stages | 0.25 | 0.5 | 0.75 | 1.0 |
| 0.5 | 1 | 430 (428) | 117 (116) | 59 (58) | 39 (38) |
| | 2 | 224 (224) | 61 (61) | 31 (31) | 21 (20) |
| | 3 | 154 (153) | 42 (42) | 22 (21) | 15 (14) |
| | 4 | 117 (117) | 33 (32) | 17 (17) | 11 (11) |
| | 5 | 95 (95) | 26 (26) | 14 (14) | 10 (9) |
| 0.6 | 1 | 298 (297) | 81 (80) | 40 (40) | 27 (26) |
| | 2 | 156 (156) | 43 (42) | 22 (21) | 15 (14) |
| | 3 | 107 (107) | 30 (29) | 15 (15) | 10 (10) |
| | 4 | 82 (81) | 23 (23) | 12 (12) | 8 (8) |
| | 5 | 66 (66) | 19 (18) | 10 (10) | 7 (7) |
| 0.7 | 1 | 219 (218) | 59 (58) | 30 (29) | 20 (19) |
| | 2 | 115 (114) | 31 (31) | 16 (16) | 11 (10) |
| | 3 | 79 (78) | 22 (22) | 12 (11) | 8 (7) |
| | 4 | 60 (60) | 17 (17) | 9 (9) | 6 (6) |
| | 5 | 49 (49) | 14 (14) | 8 (7) | 5 (5) |
| 0.8 | 1 | 167 (166) | 45 (45) | 23 (22) | 15 (14) |
| | 2 | 88 (88) | 24 (24) | 12 (12) | 9 (8) |
| | 3 | 61 (60) | 17 (17) | 9 (9) | 6 (6) |
| | 4 | 46 (46) | 13 (13) | 7 (7) | 5 (5) |
| | 5 | 38 (37) | 11 (11) | 6 (6) | 4 (4) |
| 0.9 | 1 | 132 (131) | 36 (35) | 18 (17) | 12 (11) |
| | 2 | 70 (69) | 19 (19) | 10 (10) | 7 (6) |
| | 3 | 48 (48) | 14 (13) | 7 (7) | 5 (5) |
| | 4 | 37 (37) | 11 (10) | 6 (5) | 4 (4) |
| | 5 | 30 (30) | 9 (9) | 5 (5) | 4 (3) |
| 1 | 1 | 107 (106) | 29 (28) | 14 (14) | 9 (9) |
| | 2 | 56 (56) | 16 (15) | 8 (8) | 6 (5) |
| | 3 | 39 (39) | 11 (11) | 6 (6) | 4 (4) |
| | 4 | 30 (30) | 9 (8) | 5 (4) | 4 (3) |
| | 5 | 24 (24) | 7 (7) | 4 (4) | 3 (3) |

Table 2.4: Arm sizes needed to detect the mixture alternative using the SAR where $F$ is the $t_3$ distribution for $\alpha = 0.05$ and 80% power with $\rho = 2$ determined with 1,000,000 simulations to represent the exact joint distribution. Arm sizes determined by the normal approximation are in parentheses.

has $\psi(u) = 2/\pi(1+u^2)^2$. With these locations-scale formulations, each $\Psi$ has zero mean and unit variance. Then $F(u) = \Psi((u-\mu_F)/\sigma_F)$ with mean $\mu_F$ and standard deviation $\sigma_F$.

We note the effect of $\theta$. As $\theta$ increases with all else equal, the arm size decreases. If there are nonresponders that are unaccounted for in the treatment group, the experimenter would not have the necessary arm size if they worked under the pure shift assumption ($\theta = 1$). We also see that if $K$ increases, the required arm size decreases. This expected as the control group distribution and shifted distribution are growing farther apart, making it easier to detect the shift.

Only Table 2.1, where $F$ is normal, contains arm sizes larger than those in Table 1.2 for the same design alternative. It is unsurprising that a method designed under normal theory would require less than a method designed for any distribution when $F$ is normal. However, if we know that $F$ is logistic, Laplace, or $t_3$, then the experiment can be performed with a much smaller arm size than previous methods.

In fact, the arm size is decreasing as $F$ changes from normal to logistic to Laplace to $t_3$ with all else equal. One might imagine the normal distribution would offer the smallest arm sizes since it has lighter tails than the other distributions. However, this ordering of the distributions is explained by Jeske and Yao by showing the Kullback-Leibler (K-L) distances between the null and mixture distributions from smallest to largest are the normal, logistic, Laplace, and $t_3$. A larger K-L distance corresponds to distributions that less similar.

Here, we offer an illustration to explain the ordering of the K-L distances. By starting with each distribution having zero mean and unit variance we might imagine "squeezing"

these distributions with heavier tails. This results in the normal distribution having the heaviest tails. Consider a density centered at zero and a pure shift of the density to have mean two. Figure 2.4 displays this setting for three different distributions as well as their K-L distances and area of overlap. The default $t_3$ distribution has the heaviest tails and most overlap between the mean zero and shifted densities. When we use the standardized location-scale form of the $t_3$ to reduce the base variance to one, there is much less overlap between the density with mean zero and the density with mean two as pictured in the bottom of Figure 2.4. This occurs because it is a scale change of the default $t_3$ distribution. Therefore, a probability, such as being with one standard deviation of the mean, is unchanged. The density becomes more concentrated around its center because the standard deviation being equal to one is smaller than the default standard deviation. Furthermore, the shift for the default $t_3$ in Figure 2.4 is presented as an absolute shift of two whereas it is a shift of two standard deviations for both the normal and location-scale $t_3$. If the shift of the default $t_3$ were presented in terms of standard deviations (i.e. the shifted distribution having mean $2\sqrt{3}$), the K-L distance and area of overlap would be the same as the location-scale $t_3$ values. Like the default $t_3$, the standardized location-scale $t_3$ has heavier tails than the normal distribution. The use of "$t_3$" anywhere else in this paper refers to the standardized location-scale form of the $t_3$ distribution.

The average sample number shows what the total sample size will be on average for a given design. It accounts for the random number of stages associated with the trial. Average sample numbers for a variety of designs are displayed in Table 2.5. The values are calculated by simulating each setting until termination, finding the average stage at which

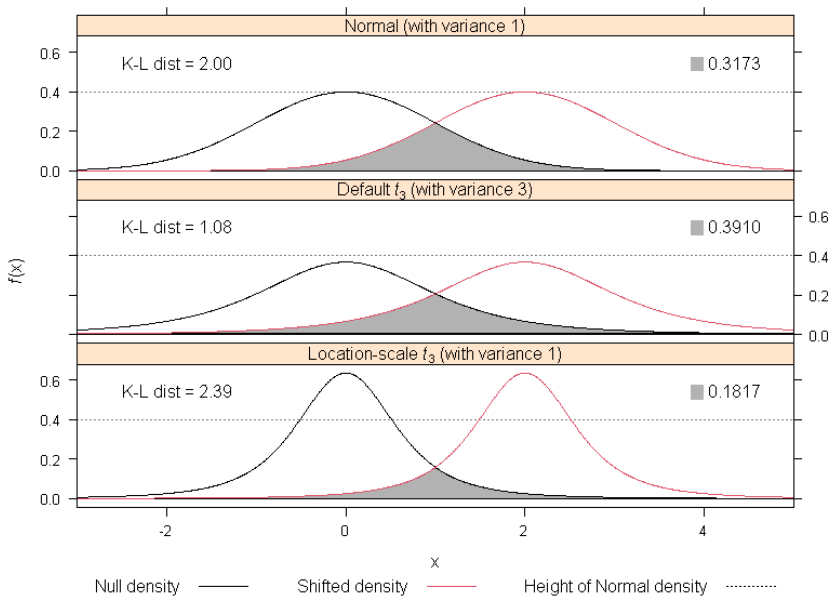


Figure 2.4: Example of overlap between location-shifted densities.

it stopped, then multiplying the average by the arm size per stage, $2m$. From the table, it is apparent that group sequential methods can offer smaller average sample sizes over fixed sample methods.

### 2.5.2 Power Simulation

In this section, we evaluate the power of the SAR test statistic. We set the design alternative with a particular $S$, $\theta$, $K$, $\alpha$, $\beta$, and $\rho$. While the SAR test statistic is nonparametric in the sense that its distribution is the same under the null hypothesis for any choice of $F$, an $F$ must be considered when calculating the power, arm size, and critical values. We do this for the situation where the $F$ is normal, logistic, Laplace, or $t_3$. Each distribution is scaled to have zero mean and unit variance. For each distribution, we get the

| | | $K = 0.5$ | | | | $K = 1$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | Stages | Normal | Logistic | Laplace | $t_3$ | Normal | Logistic | Laplace | $t_3$ |
| 0.6 | 1 | 294.0 | 260.0 | 204.0 | 158.0 | 80.0 | 72.0 | 64.0 | 50.0 |
| | 2 | 253.4 | 223.9 | 175.5 | 136.6 | 71.7 | 65.0 | 58.0 | 45.6 |
| | 3 | 239.5 | 213.0 | 164.4 | 129.0 | 67.1 | 62.3 | 53.4 | 44.2 |
| | 4 | 231.0 | 202.8 | 162.6 | 128.2 | 67.6 | 61.8 | 55.5 | 44.7 |
| 0.8 | 1 | 164.0 | 144.0 | 112.0 | 86.0 | 44.0 | 40.0 | 34.0 | 26.0 |
| | 2 | 143.2 | 126.4 | 100.3 | 77.9 | 39.3 | 36.1 | 32.4 | 26.3 |
| | 3 | 133.7 | 120.0 | 93.6 | 75.0 | 40.1 | 35.7 | 31.1 | 26.3 |
| | 4 | 129.9 | 117.5 | 90.5 | 72.7 | 39.0 | 34.4 | 32.9 | 27.7 |
| 1 | 1 | 104.0 | 90.0 | 70.0 | 54.0 | 26.0 | 24.0 | 20.0 | 16.0 |
| | 2 | 91.0 | 81.3 | 64.5 | 48.9 | 26.3 | 23.5 | 19.5 | 16.8 |
| | 3 | 88.5 | 76.0 | 61.9 | 48.7 | 26.6 | 22.3 | 20.9 | 18.3 |
| | 4 | 84.5 | 73.7 | 61.2 | 45.6 | 27.5 | 23.4 | 22.0 | 17.0 |

Table 2.5: Average sample numbers to detect the mixture alternative with SAR for $\alpha = 0.05$ and 80% power with $\rho = 2$ determined with 100,000 simulations. Critical values and arm sizes were determined by normal approximation.

arm size and critical values via simulation for the chosen design alternative. Then we use the same distribution to simulate scenarios where $K$ is different than the design alternative. That is, we are calculating the power for different $K$.

The design alternative uses $S = 2$, $\theta = 0.8$, $K = 0.5$, $\alpha = 0.05$, $\beta = 0.2$, and $\rho = 2$ for each $F$. The power curves are shown in Figure 2.5. We calculated the power curves for the SAR test statistic as well as the test statistic from Peng et al. [16]. Since the procedure from Peng et al. [16] does not use $F$ when determining the critical values and arm size, they are the same in all four panels. Clearly, both test statistics provide basically equal power curves. The large arm size used for Peng et al. [16] invokes the Central Limit Theorem, and in doing so, the test can handle data that is not normal.

The noticeable difference between the two methods is the arm size. The method from Peng et al. [16] will recommend the same arm size for different $F$, all else equal. The

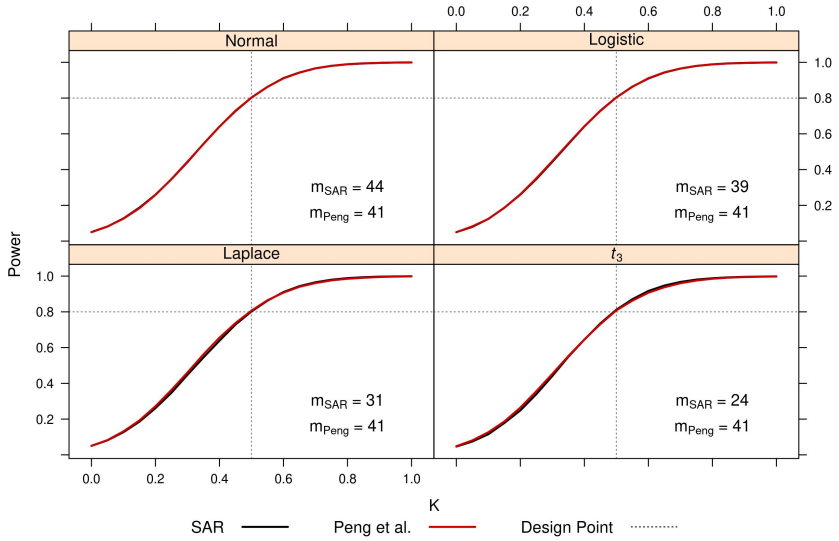


Figure 2.5: Power curves for both SAR and the test statistic from Peng et al. from a design alternative with $S = 2$, $\theta = 0.8$, $K = 0.5$, $\alpha = 0.05$, $\beta = 0.2$, $\rho = 2$, and $F$ is Normal, Logistic, Laplace, or $t_3$.

SAR is able to take a hypothesized distribution of the data into account, thus adapting its design to the data. This results in a different arm size needed to detect the alternative as well as different critical values (not shown).

It can be seen that the order of distributions from greatest to least arm size is normal, logistic, Laplace, $t_3$. The only $F$ for which the SAR has a larger arm size than the competition is the normal distribution. Peng et al. [16] developed their test statistic under normal theory, thus it is more efficient than the SAR when $F$ is normal. If we believe the distribution of the data is logistic, Laplace, or $t_3$ then using the SAR test statistic allows us to design the test appropriately and use a reduced arm size.

Under the alternative hypothesis, the joint distribution of the SAR test statistic depends on the choice of $F$. This is due to the fact that the integrals in Equation (2.6) vary with $F$. Thus, an $F$ must be chosen when designing the test in order to determine

the critical values and arm size. However, under the null hypothesis, the joint distribution of the SAR test statistic inherits the distribution-free property of the Wilcoxon Rank Sum statistic. This is rooted in the probabilities from Equation (2.4) having the values 1/2, 1/3, and 1/3, respectively, under the null hypothesis for any $F$. Consequently, the SAR procedure guarantees a test with size $\alpha$, even if the hypothesized $F$ for the alternative is misspecified.

### 2.5.3 Robustness

Next, we explore the robustness of the SAR test procedure. This will involve two parts: evaluating how the SAR responds if a hypothesized value of the design alternative, namely $\theta$ or $F$, differs from the true value.

Like the power study, we set $S = 2$, $\theta = 0.8$, $K = 0.5$, $\alpha = 0.05$, $\beta = 0.2$, and $\rho = 2$. We do this for each of the four previous location-scale distributions. Then we evaluate the power for a range of $K$ as well as a different $\theta$ than the design. For the design with $\theta = 0.8$, we look at a situation where the true proportion of responders is lower, $\theta = 0.5$, and a situation where it is higher, $\theta = 1$.

The power curves for $\theta$ less than the design alternative can be seen in Figure 2.6, and power curves for $\theta$ greater than the design alternative are in Figure 2.7. Again we compare SAR with the test statistic from Peng et al. [16] and see that performance is roughly equal between the two methods for either situation. The power curves change as we would expect with a different $\theta$ than the design: if there are more nonresponders (lower $\theta$) than the design, the power of the test will be lower, and power will be higher if there are more responders (higher $\theta$) than the design. With more responders there will be more

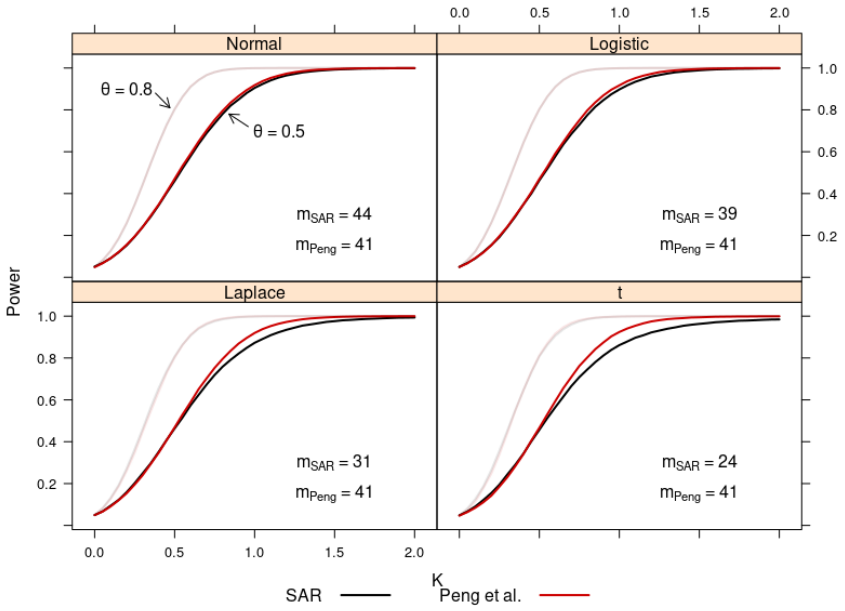


Figure 2.6: Power curves where the true $\theta$ value is less than the design alternative, $\theta = 0.8$.

evidence of the shifted component of the treatment distribution which should make it easier to detect the effect.

Further evidence to support these results lie in the examination of Tables 2.1-2.4. The arm sizes are decreasing for increasing $\theta$. Therefore, the test will not have the necessary arm size to detect the shift if $\theta$ is less than the design value, resulting in lower power.

The second component of exploring the robustness of the SAR is to see how the power curves react when the true $F$ is different than the design alternative. We use the same design alternative setup as above and create plots where each $F$ is used for the design alternative, seen in Figure 2.8. Then we simulate to evaluate the power for different $K$ and a different $F$ than the design alternative.

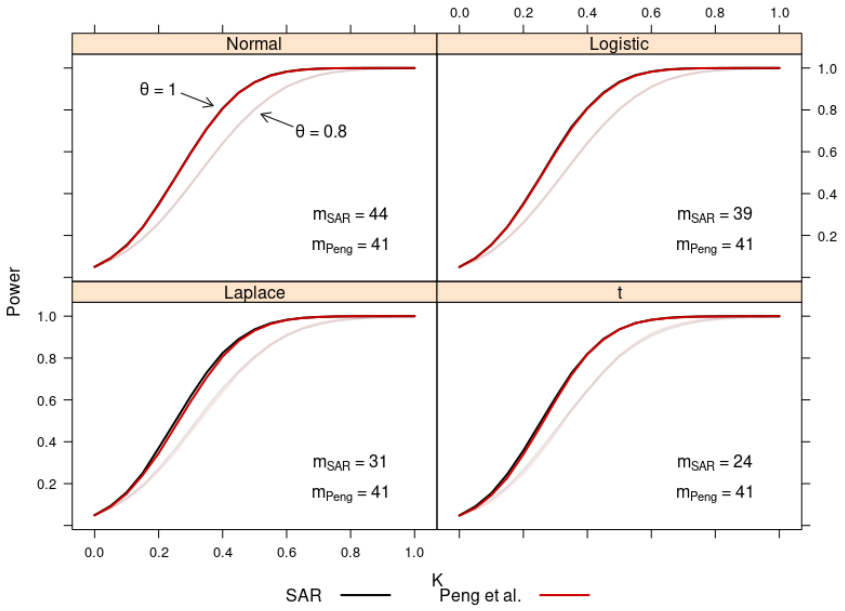


Figure 2.7: Power curves where the true $\theta$ value is greater than the design alternative, $\theta = 0.8$.

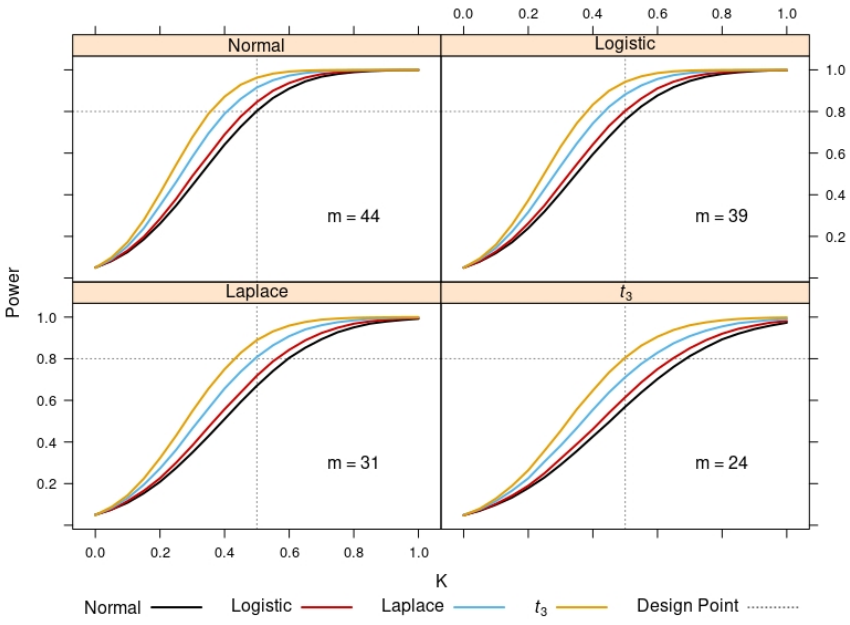


Figure 2.8: Power curves where the distribution is different than the design.

Throughout the plots, there are distinctly separated power curves for each distribution. We see that if the data actually comes from the normal distribution it has the lowest power curve, irrespective of the design alternative $F$. After normal, comes logistic, Laplace, and finally $t_3$ with the highest power curves. This is the same type of ordering as the arm sizes from the power study and corresponds to a smaller arm size being capable of detecting the same size shift if the distribution was $t_3$ compared to normal. Thus, it would be possible to have more or less power than desired depending on the hypothesized $F$ versus nature.

# Chapter 3

# Sequential Rerank

The second method we wish to consider to combine the observations at each stage we will deem the Group Sequential Rerank procedure. This method is explored by Spurrier and Hewett [21] for two stages and general alternative. It is used by Shuster et al. [19] for ordinal categorical data with an arbitrary number of stages and general alternative. We investigate the test statistic in an arbitrary number of stages with continuous data under a mixture alternative.

## 3.1 Test Statistic

For this method, we will again be finding the Wilcoxon Rank Sum statistic at each stage. However, the Sequential Rerank (SR) statistic at stage $s$ will be using the Wilcoxon Rank Sum statistic calculated from all observations up to and including stage $s$ in contrast to the SAR using only the observations at stage $s$. The SR will take into account the information from each stage by directly both using within-stage and between-stage ranks.

Let $X_{11}, \ldots, X_{1m}, \ldots, X_{S1}, \ldots, X_{Sm} \sim F$ be the potential observations from the control group and $Y_{11}, \ldots, Y_{1m}, \ldots, Y_{S1}, \ldots, Y_{Sm} \sim G$ be the potential observations from the treatment group for a group sequential clinical trial with $S$ stages. We wish to detect the alternative that $G(u) = (1-\theta)F(u) + \theta F(u-\delta)$ for $\theta \in (0, 1]$ and $\delta > 0$. For identifiability, define the null case as the point $(\theta, \delta) = (0, 0)$. Let $R_{i1}, \ldots, R_{im}$ be the ranks of $Y_{i1}, \ldots, Y_{im}$ among all $X$ and $Y$ observations up to and including stage $i$, $i = 1, \ldots, s$. That is, we find the ranks of $Y_{i1}, \ldots, Y_{im}$ in the combined sample of $2sm$ observations. At each stage, we calculate the Wilcoxon Rank Sum statistic using all observations from previous stages as well as the current stage. At stage $s$, the mean and variance are found by extending the mean and variance of the fixed sample Wilcoxon Rank Sum statistic. We simply replace $m$ with $sm$ in the fixed sample formulas to find the mean and variance of the of the Wilcoxon Rank Sum statistic at stage $s$ under the null hypothesis are

$$\mu_s = \frac{sm(2sm+1)}{2} \qquad \sigma_s^2 = \frac{(sm)^2(2sm+1)}{12} \tag{3.1}$$

Using these, we define the standardized SR test statistic at stage $s$, $s = 1, \ldots, S$

$$\tilde{Z}_s = \frac{\sum_{i=1}^{s}\sum_{j=1}^{m} R_{ij} - \mu_s}{\sigma_s} \tag{3.2}$$

This statistic has zero mean and unit variance under the null hypothesis. Hewett and Spurrier [6] provide the covariance for $\tilde{Z}_1$ and $\tilde{Z}_2$ under the null hypothesis and their methods can be use to establish the covariance between any two stages $s$ and $s'$. Let $s \leq s'$.

Then the covariance between $\tilde{Z}_s$ and $\tilde{Z}_{s'}$ is

$$\mathrm{Cov}(\tilde{Z}_s, \tilde{Z}_{s'}) = \frac{(sm)^2(sm + (s'-s)m + sm + (s'-s)m + 1)}{12\sigma_s\sigma_{s'}} = \frac{(sm)^2(2s'm+1)}{12\sigma_s\sigma_{s'}} \tag{3.3}$$

Under the alternative hypothesis, we use the probabilities defined in Equation (2.4). Then we can extend the fixed sample mean and variance of the Wilcoxon Rank Sum statistic from Equations (2.8) and (2.9) to have $sm$ observations.

$$\mathrm{E}_A[\tilde{Z}_s] = \frac{sm\left(smp + \frac{sm+1}{2}\right) - \mu_s}{\sigma_s} \tag{3.4}$$

$$\begin{aligned}\mathrm{Var}_A(\tilde{Z}_s) &= \frac{(sm)^2}{\sigma_s^2}\big(p(1-p) + (sm-1)(p_1 - p^2) + (sm-1)(p_2 - p^2)\big) \\ &= \frac{(sm)^2}{\sigma_s^2}\big(p(1-p) + (sm-1)(p_1 + p_2 - 2p^2)\big)\end{aligned} \tag{3.5}$$

The covariance of the SR statistic between stages $s$ and $s'$, $s \leq s'$, can be found using the Mann-Whitney formulation of the Wilcoxon Rank Sum statistic. The result is the following:

$$\mathrm{Cov}_A(\tilde{Z}_s,\ \tilde{Z}_{s'}) = \frac{(sm)^2}{\sigma_s\sigma_{s'}}\big(p(1-p) + (s'm-1)(p_1 + p_2 - 2p^2)\big) \tag{3.6}$$

The same arm size algorithm from Figure 2.1 can be used to determine the arm size and critical values where we replace the SAR statistic with the SR statistic.

## 3.2 Discussion of the Joint Distribution

The number of possible values in the exact distribution grows quickly. For a two-stage test, "there are $(4m)!/(m!m!m!m!)$ possible rankings of the first and second samples of the $X$'s and $Y$'s" [6]. As for SAR, we consider a multivariate normal approximation to circumvent the need for running the large number of simulations that could be necessary to enumerate the exact distribution.

It is clear that the marginal distribution at each stage is asymptotically normal both under the null and alternative hypothesis, given that each is a standardized Wilcoxon Rank Sum statistic. Spurrier and Hewett [21] show that the SR test statistic for the first and second stages has a limiting bivariate normal distribution under the null and alternative hypotheses. Their methods can be applied to any pairwise combination of stages. This results in asymptotic bivariate normality for any two $(\tilde{Z}_s, \tilde{Z}_{s'})$. Shuster et al. [19] show that the SR statistic will have an asymptotic multivariate normal distribution for three or more stages under both hypotheses.

We present an empirical illustration of the limiting multivariate normal of the SR test statistic. First, we generate the joint distribution of the SR under both the null and alternative hypothesis for a four-stage design alternative where the data comes from a Laplace distribution with $\theta = 0.8$, $K = 0.25$. We use an arm size of 58, the necessary arm size based on our algorithm to detect alternative with $\alpha = 0.05$, $\beta = 0.2$, and $\rho = 2$. This was done with 100,000 simulations where a single simulation is calculation of the test statistic through all four stages. Figure 3.1 shows scatterplots of the values of the SR test statistic for each pairwise combination of the stages. Each plot shows the points forming

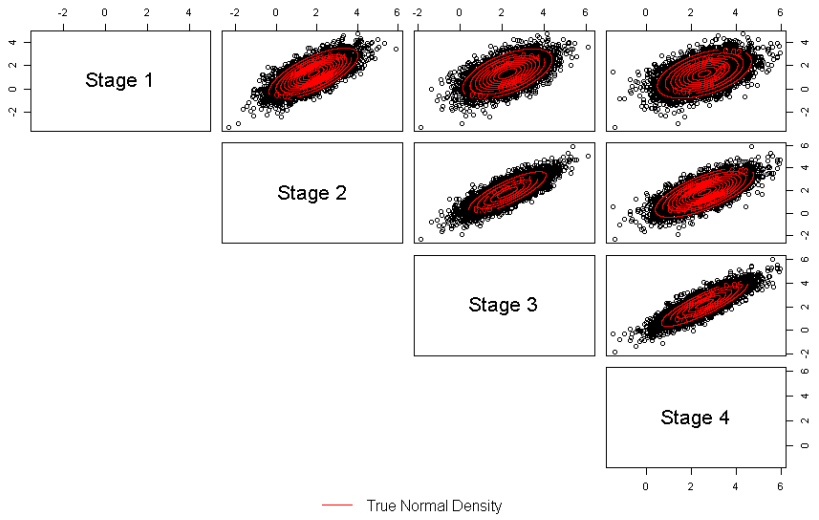


Figure 3.1: Pairwise scatterplots of the values of the SR at each stage with contours of the bivariate normal distribution overlaid.

a clear ellipse, suggesting bivariate normality. The bivariate normal density contours are overlaid using the mean vector and variance-covariance matrices determined by Equations (3.4) and (3.6), respectively. The similarity of the points and the curves suggest bivariate normality holds for each pairwise combination of the stages.

In order to provide evidence of multivariate normality, we consider an inverse Cholesky transformation. If we have $\mathbf{X} = (X_1, \ldots, X_S) \sim N_S(\boldsymbol{\mu}, \Sigma)$, then the Cholesky decomposition of the covariance matrix is $\Sigma = \Gamma'\Gamma$. Let $\mathbf{Z} = (\Gamma')^{-1}(\mathbf{X} - \boldsymbol{\mu})$. Then $\mathbf{Z}$ will be $N_S(\mathbf{0}, \mathbf{I}_S)$, where $\mathbf{I}_S$ is the $S \times S$ identity matrix. Thus, if we have multivariate normal data to begin with, the inverse Cholesky transformation will result in independent standard normal random variables. We aim to show that since performing the Cholesky transformation on the joint distribution of the SR test statistic results in independent standard normal random variables, the joint distribution must have been multivariate normal from the start.

We will consider this transformation under both the null and alternative hypotheses. The joint distribution of the SR statistic is simulated as above and we perform the transformation on the variance-covariance matrices calculated by Equations (3.3) and (3.6). Figures 3.2 and 3.3 show the results of the transformation under the null and alternative hypotheses, respectively. In the plots on the diagonals in these figures, we see the univariate histograms at each stage with the $N(0,1)$ density overlaid. In the upper triangle of the grids, we present scatterplots of each pairwise combination of the stages. These have $N_2(\mathbf{0}, \mathbf{I}_2)$ density curves overlaid. The similarity of the transformed data to the density curves in both the univariate and bivariate plots suggests that the transformation did indeed result in independent standard normal random variables. In the lower triangle of the figures are the results of a nonparametric test for independence. The Spearman test for independence tests the null hypothesis of independence versus the alternative that there is some relation between the variables. We perform a Bonferroni correction to adjust for multiple testing. The corrected $p$-values are all extremely large. Therefore, it suggests that the joint distribution of the SR is multivariate normal under both the null and alternative hypotheses.

We note that Figures 3.1-3.3 used a random subset of 5,000 simulations from the original 100,000 due to computation constraints.

More empirical support that the distribution of the SR under both the null and alternative hypotheses is multivariate normal can be seen in Figure 3.4. Critical values and arm sizes were determined by using the multivariate normal approximation of the joint distribution and compared to the values obtained by the simulation of the joint distribution.

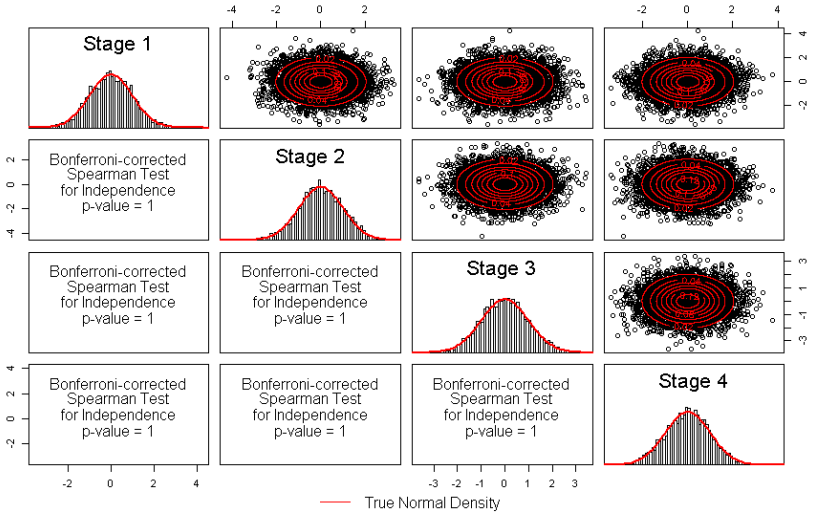


Figure 3.2: Results of the Cholesky transformation on the SR test statistic under the null hypothesis.

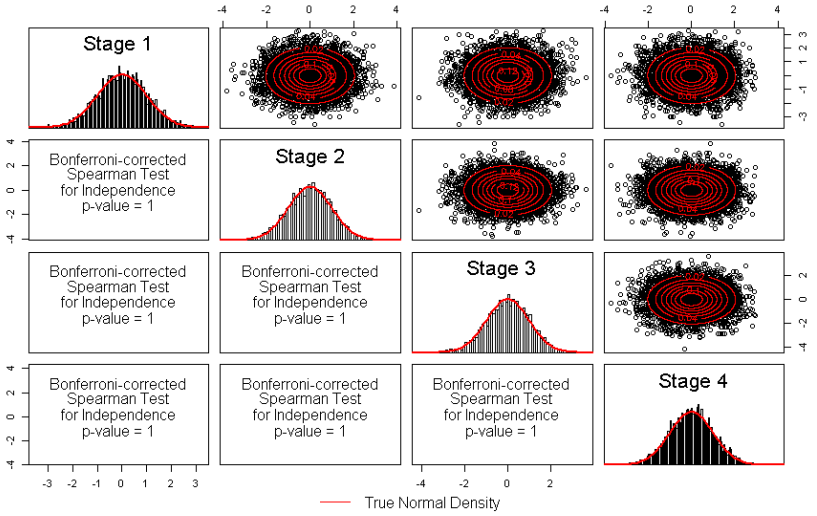


Figure 3.3: Results of the Cholesky transformation on the SR test statistic under the alternative hypothesis.

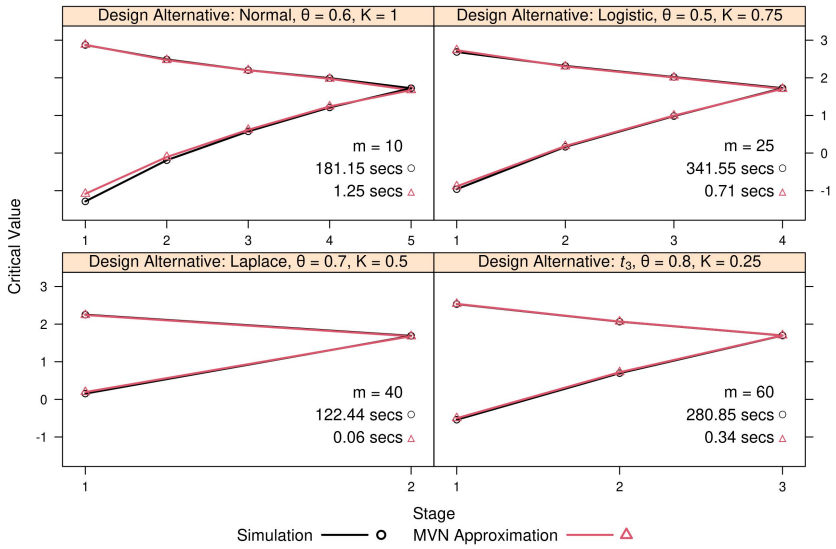


Figure 3.4: Comparison of the multivariate normal approximation and simulation of the exact distribution for the SR using 100,000 simulations with $\alpha = 0.05$, $\beta = 0.2$, and $\rho = 2$.

The means and variance-covariance matrices from Section 3.1 were used for the multivariate normal distribution. Replacing the SAR statistic with the SR statistic, the same steps as the algorithm in Figure 2.2 were used to determine the arm size and critical values.

In the various scenarios, both algorithms give the exact same arm size. There are some small differences in the critical values in the smaller arm size situations. This is perhaps attributed to the discreteness of the test statistic when there are small arm sizes. The critical values become indistinguishable between the two algorithms as the arm size increases. As with the SAR, we see a great improvement in computation time by using the normal approximation.

Tables 3.1-3.4 display arm sizes needed to detect a number of design alternatives while using the SR statistic. The fixed sample arm sizes using the normal approximation in these tables were determined by the algorithm in Figure 2.2 to be consistent with the rest of

the formulation of the tables, rather than using Equation (1.19). The arm sizes determined by simulation in these tables were found by running the simulation three times and taking the largest arm size. Some differences between the arm sizes from the simulation and normal approximation may be attributed to the randomness of three simulations. The normal approximation was used under both the null and alternative hypotheses. Equal arm sizes between the simulation algorithm and the normal approximation algorithm provide further evidence that the distribution of the SR statistic can be approximated by the multivariate normal distribution.

## 3.3 Comparison of SAR and SR

With both the SAR and SR being based on the Wilcoxon Rank Sum statistic, it is of interest to see how they compare and whether one may be universally preferred over the other. Note that the SR statistic and SAR statistic are the same for the first stage of a group sequential clinical trial, as they are both essentially the fixed sample Wilcoxon Rank Sum test at that time. We also keep in mind that the comparisons of the observations at each stage to determine the ranks for the SAR statistic are a subset of the comparisons done for the SR statistic since both use the within-stage ranks.

We begin by looking at the limits of the mean, variance, and covariance of each statistic. At each stage, $s$ and $p$ are fixed, since $s$ is the stage number and $p = P(X < Y)$ is determined by $\theta$, $\delta$, and $F$. First, we compare the means by looking at the limit of their ratio for increasing $m$. Under the null hypothesis, both test statistics have zero mean.

| | | $K$ | | | |
|---|---|---|---|---|---|
| $\theta$ | Stages | 0.25 | 0.5 | 0.75 | 1.0 |
| 0.5 | 1 | 837 (837) | 215 (215) | 101 (100) | 60 (60) |
| | 2 | 437 (437) | 113 (113) | 53 (53) | 32 (32) |
| | 3 | 300 (299) | 77 (77) | 36 (36) | 22 (22) |
| | 4 | 228 (228) | 59 (59) | 28 (28) | 17 (17) |
| | 5 | 184 (184) | 48 (48) | 23 (22) | 14 (14) |
| 0.6 | 1 | 581 (580) | 149 (149) | 69 (69) | 42 (42) |
| | 2 | 304 (303) | 78 (78) | 37 (36) | 22 (22) |
| | 3 | 208 (207) | 54 (54) | 25 (25) | 15 (15) |
| | 4 | 159 (158) | 41 (41) | 19 (19) | 12 (12) |
| | 5 | 128 (128) | 33 (33) | 16 (16) | 10 (10) |
| 0.7 | 1 | 426 (426) | 110 (109) | 51 (51) | 30 (30) |
| | 2 | 223 (223) | 57 (57) | 27 (27) | 16 (16) |
| | 3 | 153 (152) | 40 (39) | 19 (18) | 11 (11) |
| | 4 | 116 (116) | 30 (30) | 14 (14) | 9 (9) |
| | 5 | 94 (94) | 25 (24) | 12 (12) | 7 (7) |
| 0.8 | 1 | 327 (326) | 84 (84) | 39 (39) | 23 (23) |
| | 2 | 171 (171) | 44 (44) | 21 (20) | 12 (12) |
| | 3 | 117 (117) | 30 (30) | 14 (14) | 9 (9) |
| | 4 | 89 (89) | 23 (23) | 11 (11) | 7 (7) |
| | 5 | 72 (72) | 19 (19) | 9 (9) | 6 (5) |
| 0.9 | 1 | 258 (257) | 66 (66) | 30 (30) | 18 (18) |
| | 2 | 135 (135) | 35 (35) | 16 (16) | 10 (10) |
| | 3 | 92 (92) | 24 (24) | 11 (11) | 7 (7) |
| | 4 | 70 (70) | 18 (18) | 9 (9) | 6 (5) |
| | 5 | 57 (57) | 15 (15) | 7 (7) | 5 (4) |
| 1 | 1 | 210 (209) | 53 (53) | 25 (24) | 15 (14) |
| | 2 | 109 (109) | 28 (28) | 13 (13) | 8 (8) |
| | 3 | 75 (75) | 20 (19) | 9 (9) | 6 (5) |
| | 4 | 57 (57) | 15 (15) | 7 (7) | 5 (4) |
| | 5 | 46 (46) | 12 (12) | 6 (6) | 4 (4) |

Table 3.1: Arm sizes needed to detect the mixture alternative using the SR where $F$ is the standard normal distribution for $\alpha = 0.05$ and 80% power with $\rho = 2$ determined with 1,000,000 simulations to represent the exact joint distribution. Arm sizes determined by the normal approximation are in parentheses.

| | | $K$ | | | |
|---|---|---|---|---|---|
| $\theta$ | Stages | 0.25 | 0.5 | 0.75 | 1.0 |
| 0.5 | 1 | 732 (730) | 190 (190) | 90 (90) | 55 (55) |
| | 2 | 383 (382) | 100 (99) | 47 (47) | 29 (29) |
| | 3 | 262 (261) | 68 (68) | 33 (32) | 20 (20) |
| | 4 | 199 (199) | 52 (52) | 25 (25) | 15 (15) |
| | 5 | 162 (161) | 42 (42) | 20 (20) | 13 (12) |
| 0.6 | 1 | 508 (507) | 132 (133) | 62 (62) | 38 (37) |
| | 2 | 265 (265) | 69 (69) | 33 (33) | 20 (20) |
| | 3 | 182 (181) | 47 (47) | 23 (22) | 14 (14) |
| | 4 | 138 (138) | 36 (36) | 17 (17) | 11 (11) |
| | 5 | 112 (112) | 29 (29) | 14 (14) | 9 (9) |
| 0.7 | 1 | 374 (373) | 97 (96) | 45 (45) | 28 (28) |
| | 2 | 195 (195) | 51 (51) | 24 (24) | 15 (15) |
| | 3 | 134 (133) | 35 (35) | 17 (17) | 10 (10) |
| | 4 | 102 (102) | 27 (27) | 13 (13) | 8 (8) |
| | 5 | 82 (82) | 22 (22) | 10 (10) | 7 (6) |
| 0.8 | 1 | 286 (285) | 74 (74) | 35 (35) | 21 (21) |
| | 2 | 149 (149) | 39 (39) | 19 (18) | 11 (11) |
| | 3 | 102 (102) | 27 (27) | 13 (13) | 8 (8) |
| | 4 | 78 (78) | 20 (20) | 10 (10) | 6 (6) |
| | 5 | 63 (63) | 17 (17) | 8 (8) | 5 (5) |
| 0.9 | 1 | 225 (225) | 58 (58) | 27 (27) | 17 (16) |
| | 2 | 118 (118) | 31 (31) | 15 (14) | 9 (9) |
| | 3 | 81 (81) | 21 (21) | 10 (10) | 6 (6) |
| | 4 | 62 (62) | 16 (16) | 8 (8) | 5 (5) |
| | 5 | 50 (50) | 13 (13) | 7 (6) | 4 (4) |
| 1 | 1 | 182 (182) | 47 (47) | 22 (22) | 14 (13) |
| | 2 | 96 (95) | 25 (25) | 12 (12) | 7 (7) |
| | 3 | 66 (65) | 17 (17) | 8 (8) | 5 (5) |
| | 4 | 50 (50) | 13 (13) | 7 (6) | 4 (4) |
| | 5 | 41 (40) | 11 (11) | 5 (5) | 4 (3) |

Table 3.2: Arm sizes needed to detect the mixture alternative using the SR where $F$ logistic distribution for $\alpha = 0.05$ and 80% power with $\rho = 2$ determined with 1,000,000 simulations to represent the exact joint distribution. Arm sizes determined by the normal approximation are in parentheses.

| | | $K$ | | | |
|---|---|---|---|---|---|
| $\theta$ | Stages | 0.25 | 0.5 | 0.75 | 1.0 |
| 0.5 | 1 | 546 (546) | 149 (149) | 74 (74) | 48 (48) |
| | 2 | 286 (285) | 78 (78) | 39 (39) | 25 (25) |
| | 3 | 195 (195) | 54 (53) | 27 (27) | 18 (17) |
| | 4 | 149 (149) | 41 (41) | 21 (21) | 14 (13) |
| | 5 | 121 (120) | 33 (33) | 17 (17) | 11 (11) |
| 0.6 | 1 | 379 (379) | 103 (103) | 51 (51) | 33 (33) |
| | 2 | 198 (198) | 54 (54) | 27 (27) | 18 (17) |
| | 3 | 136 (136) | 37 (37) | 19 (19) | 12 (12) |
| | 4 | 103 (103) | 29 (28) | 14 (14) | 9 (9) |
| | 5 | 84 (84) | 23 (23) | 12 (12) | 8 (8) |
| 0.7 | 1 | 279 (278) | 76 (76) | 38 (38) | 24 (24) |
| | 2 | 146 (146) | 40 (40) | 20 (20) | 13 (13) |
| | 3 | 100 (100) | 27 (27) | 14 (14) | 9 (9) |
| | 4 | 76 (76) | 21 (21) | 11 (11) | 7 (7) |
| | 5 | 62 (62) | 17 (17) | 9 (9) | 6 (6) |
| 0.8 | 1 | 213 (213) | 58 (58) | 29 (29) | 18 (18) |
| | 2 | 111 (111) | 31 (30) | 15 (15) | 10 (10) |
| | 3 | 77 (76) | 21 (21) | 11 (11) | 7 (7) |
| | 4 | 58 (58) | 16 (16) | 8 (8) | 6 (5) |
| | 5 | 47 (47) | 13 (13) | 7 (7) | 5 (4) |
| 0.9 | 1 | 169 (168) | 46 (46) | 23 (23) | 15 (14) |
| | 2 | 88 (88) | 24 (24) | 12 (12) | 8 (8) |
| | 3 | 61 (60) | 17 (17) | 9 (8) | 6 (5) |
| | 4 | 46 (46) | 13 (13) | 7 (6) | 4 (4) |
| | 5 | 38 (37) | 11 (10) | 6 (5) | 4 (4) |
| 1 | 1 | 136 (136) | 37 (37) | 18 (18) | 12 (11) |
| | 2 | 72 (71) | 20 (19) | 10 (10) | 7 (6) |
| | 3 | 49 (49) | 14 (13) | 7 (7) | 5 (4) |
| | 4 | 37 (37) | 11 (10) | 6 (5) | 4 (3) |
| | 5 | 30 (30) | 9 (8) | 5 (4) | 3 (3) |

Table 3.3: Arm sizes needed to detect the mixture alternative using the SR where $F$ is the Laplace distribution for $\alpha = 0.05$ and 80% power with $\rho = 2$ determined with 1,000,000 simulations to represent the exact joint distribution. Arm sizes determined by the normal approximation are in parentheses.

| | | $K$ | | | |
|---|---|---|---|---|---|
| $\theta$ | Stages | 0.25 | 0.5 | 0.75 | 1.0 |
| 0.5 | 1 | 430 (428) | 117 (116) | 59 (58) | 39 (38) |
| | 2 | 224 (224) | 61 (61) | 31 (31) | 20 (20) |
| | 3 | 154 (153) | 42 (42) | 21 (21) | 14 (14) |
| | 4 | 117 (117) | 32 (32) | 16 (16) | 11 (11) |
| | 5 | 95 (94) | 26 (26) | 13 (13) | 9 (9) |
| 0.6 | 1 | 298 (297) | 81 (80) | 40 (40) | 27 (26) |
| | 2 | 156 (155) | 42 (42) | 22 (21) | 14 (14) |
| | 3 | 107 (106) | 29 (29) | 15 (15) | 10 (10) |
| | 4 | 82 (81) | 22 (22) | 12 (11) | 8 (8) |
| | 5 | 66 (66) | 18 (18) | 9 (9) | 6 (6) |
| 0.7 | 1 | 219 (218) | 59 (59) | 30 (29) | 20 (19) |
| | 2 | 114 (114) | 31 (31) | 16 (16) | 11 (10) |
| | 3 | 78 (78) | 21 (21) | 11 (11) | 7 (7) |
| | 4 | 60 (60) | 16 (16) | 9 (8) | 6 (6) |
| | 5 | 48 (48) | 13 (13) | 7 (7) | 5 (5) |
| 0.8 | 1 | 167 (167) | 45 (45) | 23 (22) | 15 (15) |
| | 2 | 87 (87) | 24 (24) | 12 (12) | 8 (8) |
| | 3 | 60 (60) | 17 (16) | 9 (8) | 6 (6) |
| | 4 | 46 (46) | 13 (13) | 7 (6) | 5 (4) |
| | 5 | 37 (37) | 10 (10) | 6 (5) | 4 (4) |
| 0.9 | 1 | 132 (132) | 36 (35) | 18 (18) | 12 (11) |
| | 2 | 69 (69) | 19 (19) | 9 (9) | 7 (6) |
| | 3 | 47 (47) | 13 (13) | 7 (7) | 5 (4) |
| | 4 | 36 (36) | 10 (10) | 5 (5) | 4 (3) |
| | 5 | 29 (29) | 8 (8) | 4 (4) | 3 (3) |
| 1 | 1 | 107 (107) | 29 (29) | 14 (14) | 9 (9) |
| | 2 | 56 (56) | 15 (15) | 8 (8) | 5 (5) |
| | 3 | 38 (38) | 11 (11) | 6 (5) | 4 (4) |
| | 4 | 30 (29) | 8 (8) | 4 (4) | 3 (3) |
| | 5 | 24 (24) | 7 (7) | 4 (3) | 3 (2) |

Table 3.4: Arm sizes needed to detect the mixture alternative using the SR where $F$ is the $t_3$ distribution for $\alpha = 0.05$ and 80% power with $\rho = 2$ determined with 1,000,000 simulations to represent the exact joint distribution. Arm sizes determined by the normal approximation are in parentheses.

Under the alternative hypothesis, the mean of the SAR statistic is

$$\begin{aligned} \mathrm{E}_A[Z_s] &= \frac{\mu_A - \mu}{\sigma/\sqrt{s}} = \frac{\sqrt{s}\left(m\left(mp + \frac{m+1}{2}\right) - \frac{m(2m+1)}{2}\right)}{\sqrt{m^2(2m+1)/12}} \\ &= \frac{\sqrt{12s}\left(p - \frac{1}{2}\right)m^2}{\sqrt{2m^3 + m^2}} \end{aligned} \tag{3.7}$$

and the mean of the SR statistic is

$$\begin{aligned} \mathrm{E}_A[\tilde{Z}_s] &= \frac{sm\left(smp + \frac{sm+1}{2}\right) - \mu_s}{\sigma_s} = \frac{sm\left(smp + \frac{sm+1}{2}\right) - \frac{sm(2sm+1)}{2}}{\sqrt{(sm)^2(2sm+1)/12}} \\ &= \frac{\sqrt{12}\left(p - \frac{1}{2}\right)(sm)^2}{\sqrt{2(sm)^3 + (sm)^2}} \end{aligned} \tag{3.8}$$

Using these means, we examine their ratio as $m \to \infty$

$$\lim_{m\to\infty} \frac{\mathrm{E}_A[Z_s]}{\mathrm{E}_A[\tilde{Z}_s]} = \lim_{m\to\infty} \left(\frac{\sqrt{12s}\left(p - \frac{1}{2}\right)m^2}{\sqrt{2m^3 + m^2}}\right)\left(\frac{\sqrt{2(sm)^3 + (sm)^2}}{\sqrt{12}\left(p - \frac{1}{2}\right)(sm)^2}\right) \tag{3.9}$$

$$= \lim_{m\to\infty} \frac{\sqrt{s}\left(\sqrt{2(sm)^3 + (sm)^2}\right)}{s^2\sqrt{2m^3 + m^2}} \tag{3.10}$$

$$= \frac{\sqrt{s}\left(\sqrt{s^3}\right)}{s^2} = 1 \tag{3.11}$$

When $m \to \infty$, it can be seen that the ratio of the means goes to one. Therefore the means under both the null and alternative hypotheses are asymptotically equivalent. Figure 3.5 provides a visual of the ratio of the means under a particular design alternative.

Next, we examine the covariance of the test statistics. We will consider stages $s$ and $s'$, for $s \leq s'$. Equality allows us to consider the variance. Recall that both test statistics have unit variance under the null hypothesis. Under the null hypothesis, the covariance of

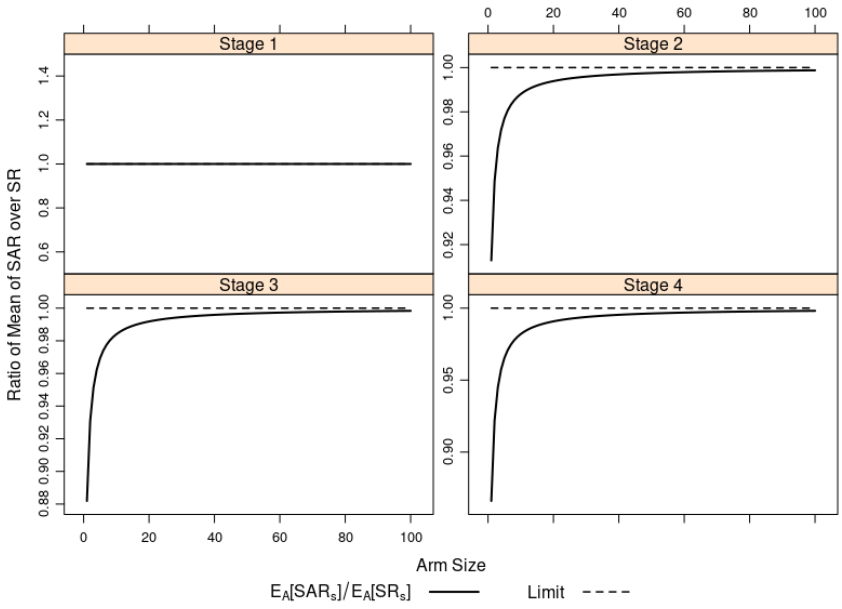


Figure 3.5: Ratio of the means of the SAR and SR under the alternative hypothesis as $m \to \infty$ for a design with $S = 4$, $\theta = 0.8$, $K = 0.5$, and $F$ is the normal distribution.

the SAR between stages $s$ and $s'$ is a fixed number, $\sqrt{s/s'}$, only depending on the number of the stages being used. The covariance of the SR depends on the stages as well as $m$. With some simplification of this covariance under the null hypothesis we get

$$\begin{aligned}\mathrm{Cov}(\tilde{Z}_s, \tilde{Z}_{s'}) &= \frac{(sm)^2(2s'm+1)}{12\sigma_s\sigma_{s'}} = \frac{(sm)^2(2s'm+1)}{12\sqrt{((sm)^2(2sm+1)/12)((s'm)^2(2s'm+1)/12)}} \\ &= \frac{s(2s'm+1)}{s'\sqrt{4ss'm^2+2sm+2s'm+1}} \end{aligned} \tag{3.12}$$

This has the limit

$$\lim_{m\to\infty} \mathrm{Cov}(\tilde{Z}_s, \tilde{Z}_{s'}) = \sqrt{\frac{s}{s'}} \tag{3.13}$$

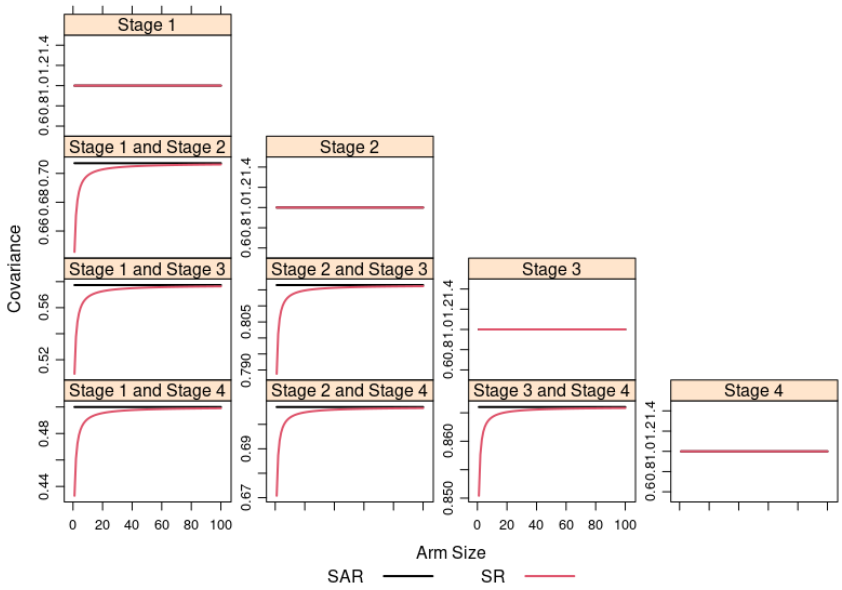


Figure 3.6: Covariance of the SAR and SR under the null hypothesis as $m \to \infty$ for a four-stage design.

Thus, we see that the asymptotic covariance of the SR is exactly the covariance of the SAR under the null hypothesis. The behavior of this covariance as $m \to \infty$ can be seen in Figure 3.6.

The general covariance of the SAR statistic and the SR statistic are those used under the alternative hypothesis. Both depend on the number of stages, the arm size, and other aspects of the design alternative: $\theta$, $K$, and $F$. Consider the following calculations of the limit of the ratio of covariances under the alternative hypothesis as $m \to \infty$. First, the

covariance for the SAR statistic:

$$\begin{aligned}
\mathrm{Cov}_A(Z_s, Z_{s'}) &= \Big(\frac{s}{s'}\Big)^{\frac{1}{2}} \frac{\sigma_A^2}{\sigma^2} \\
&= \Big(\frac{s}{s'}\Big)^{\frac{1}{2}} \frac{m^2\big(p(1-p) + (m-1)(p_1 - p^2) + (m-1)(p_2 - p^2)\big)}{m^2(2m+1)/12} \\
&= \Big(\frac{s}{s'}\Big)^{\frac{1}{2}} \frac{12\big(p(1-p) + (m-1)(p_1 - p^2 + p_2 - p^2)\big)}{2m+1} \qquad (3.14)
\end{aligned}$$

For the covariance of the SR statistic, we have

$$\begin{aligned}
\mathrm{Cov}_A(\tilde{Z}_s, \tilde{Z}_{s'}) &= \frac{(sm)^2}{\sigma_s \sigma_{s'}} \big(p(1-p) + (s'm-1)(p_1 - p^2) + (s'm-1)(p_2 - p^2)\big) \\
&= \frac{(sm)^2\big(p(1-p) + (s'm-1)(p_1 - p^2) + (s'm-1)(p_2 - p^2)\big)}{\sqrt{\big((sm)^2(2sm+1)/12\big)\big((s'm)^2(2s'm+1)/12\big)}} \\
&= \frac{12s\big(p(1-p) + (s'm-1)(p_1 - p^2 + p_2 - p^2)\big)}{s'\sqrt{4ss'm^2 + 2sm + 2s'm + 1}} \qquad (3.15)
\end{aligned}$$

The limit of their ratio is

$$\begin{aligned}
\lim_{m\to\infty} \frac{\mathrm{Cov}_A(Z_s, Z_{s'})}{\mathrm{Cov}_A(\tilde{Z}_s, \tilde{Z}_{s'})} &= \lim_{m\to\infty} \frac{\big(\frac{s}{s'}\big)^{\frac{1}{2}} \frac{12\left(p(1-p)+(m-1)(p_1-p^2+p_2-p^2)\right)}{2m+1}}{\frac{12s\left(p(1-p)+(s'm-1)(p_1-p^2+p_2-p^2)\right)}{s'\sqrt{4ss'm^2+2sm+2s'm+1}}} \qquad (3.16) \\
&= \lim_{m\to\infty} \left[\sqrt{\frac{s'}{s}} \frac{p(1-p) + (m-1)(p_1 + p_2 - 2p^2)}{p(1-p) + (s'm-1)(p_1 + p_2 - 2p^2)}\right. \qquad (3.17) \\
&\qquad \left. \cdot \frac{\sqrt{4ss'm + 2sm + 2s'm + 1}}{2m+1}\right] \qquad (3.18) \\
&= \sqrt{\frac{s'}{s}} \left(\frac{1}{s'}\right) \sqrt{ss'} = 1 \qquad (3.19)
\end{aligned}$$

Thus, the two standardized test statistics both have zero mean, unit variance, and the same asymptotic covariance under the null hypothesis and have the same asymptotic

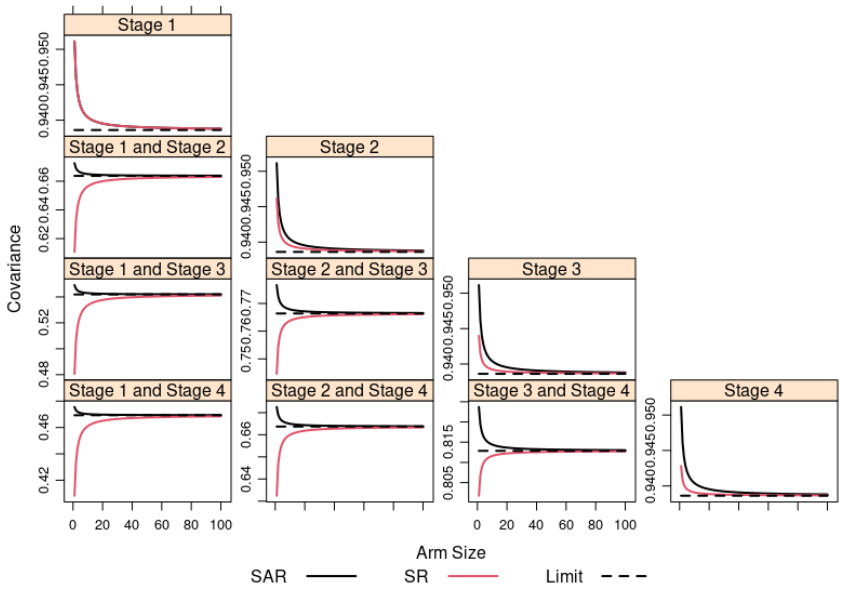


Figure 3.7: Covariance of the SAR and SR under the alternative hypothesis as $m \to \infty$ for a design with $S = 4$, $\theta = 0.8$, $K = 0.5$, and $F$ is the normal distribution.

mean and variance-covariance matrix under the alternative hypothesis. Figure 3.7 shows the behavior of the variance and covariance for a particular scenario as $m$ gets large.

With the asymptotic equality of means, variances, and covariances shown above and the asymptotic behavior of the bivariate distribution, it can be seen that the SAR and Rerank methods have the same bivariate limiting distributions under both the null and alternative hypothesis. This is most notable for two-stage designs, however it holds for pairwise groupings of any two stages. From the discussion of the full limiting joint distribution, we know the SAR and SR test statistics have the same asymptotic multivariate normal distribution.

As an illustration of the similarity between the SAR and SR procedures, consider the arm sizes needed to detect a given design alternative for each method in Table 3.5.

The arm sizes were determined by the normal approximation of the joint distributions for both the SAR statistic and SR statistic. Many of the arm sizes presented are equal. Any inequality between the arm sizes is only a difference of one in favor of the SR statistic.

| | | $K = 0.5$ | | | | $K = 1$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | Stages | Normal | Logistic | Laplace | $t_3$ | Normal | Logistic | Laplace | $t_3$ |
| 0.6 | 1 | 147 (147) | 130 (130) | 102 (102) | 79 (79) | 40 (40) | 36 (36) | 32 (32) | 25 (25) |
| | 2 | 78 (78) | 69 (69) | 54 (54) | 42 (42) | 22 (22) | 20 (20) | 17 (18) | 14 (14) |
| | 3 | 54 (54) | 47 (48) | 37 (37) | 29 (29) | 15 (15) | 14 (14) | 12 (12) | 10 (10) |
| | 4 | 41 (41) | 36 (36) | 28 (29) | 22 (23) | 12 (12) | 11 (11) | 9 (10) | 8 (8) |
| 0.8 | 1 | 82 (82) | 72 (72) | 56 (56) | 43 (43) | 22 (22) | 20 (20) | 17 (17) | 13 (13) |
| | 2 | 44 (44) | 39 (39) | 30 (31) | 24 (24) | 12 (12) | 11 (11) | 10 (10) | 8 (8) |
| | 3 | 30 (30) | 27 (27) | 21 (21) | 16 (17) | 9 (9) | 8 (8) | 7 (7) | 6 (6) |
| | 4 | 23 (23) | 20 (21) | 16 (16) | 13 (13) | 7 (7) | 6 (6) | 5 (6) | 4 (5) |
| 1 | 1 | 52 (52) | 45 (45) | 35 (35) | 27 (27) | 13 (13) | 12 (12) | 10 (10) | 8 (8) |
| | 2 | 28 (28) | 25 (25) | 19 (20) | 15 (15) | 8 (8) | 7 (7) | 6 (6) | 5 (5) |
| | 3 | 19 (20) | 17 (17) | 13 (14) | 11 (11) | 5 (6) | 5 (5) | 4 (5) | 4 (4) |
| | 4 | 15 (15) | 13 (13) | 10 (11) | 8 (8) | 4 (5) | 4 (4) | 3 (4) | 3 (3) |

Table 3.5: Arm sizes necessary to detect the mixture alternative with SR statistic for $\alpha = 0.05$ and 80% power with $\rho = 2$ determined by the normal approximation. Arm sizes for the SAR statistic determined by the normal approximation are in parentheses.

Although we are not using the between-stage ranks with the SAR, these comparisons do not seem to provide much more information than the within-stage ranks used in both the SAR and SR. To shed some light on this, we can examine the Mann-Whitney formulation of the Wilcoxon Rank Sum statistics used in the SAR and the SR. We consider a two-stage group sequential design, but the ideas hold for more than two stages. For the SAR at stage 2, the Mann-Whitney $U$ is

$$U_2 = \sum_{i=1}^{m}\sum_{j=1}^{m} I(X_{1i} < Y_{1j}) + \sum_{i=1}^{m}\sum_{j=1}^{m} I(X_{2i} < Y_{2j}) \tag{3.20}$$

where $I(\cdot)$ is one if the argument is true and zero otherwise. The SR statistic at stage $s$ uses Mann-Whitney $\tilde{U}$

$$\tilde{U}_2 = \sum_{i=1}^{m}\sum_{j=1}^{m} I(X_{1i} < Y_{1j}) + \sum_{i=1}^{m}\sum_{j=1}^{m} I(X_{2i} < Y_{2j}) + \tag{3.21}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} I(X_{1i} < Y_{2j}) + \sum_{i=1}^{m}\sum_{j=1}^{m} I(X_{2i} < Y_{1j}) \tag{3.22}$$

Obviously, the quantities that constitute the SAR are a subset of those used for the SR. More importantly, the additional quantities, the between-stage comparisons, used for the SR are not independent of the within-stage comparisons. This dependence is likely the reason that there is little information gained with their inclusion.

An added benefit of the SAR, is the ability to only record the Wilcoxon Rank Sum statistic at each stage whereas all observations need to be used throughout the entirety of the experiment with the SR. With the evidence above of asymptotic equivalence, we will focus solely on the SAR in the remainder of this work.

# Chapter 4

# Multiplicative Treatment Effect

While the previous chapters focused on distributions that are continuous and symmetric with support on the real line, this may not be the data type for all experimenters. In this chapter, we wish to explore the efficacy of the SAR statistic when the data comes from a continuous, non-negative distribution.

## 4.1 Scale Family and Multiplicative Treatment Effect

To define a scale family, let $Z$ have CDF $\Psi$ with scale parameter equal to one without loss of generality. If $X = \sigma_F Z$ has CDF $F(x) = \Psi(x/\sigma_F)$ for $\sigma_F > 0$, then $X$ has scale parameter $\sigma_F$. Without loss of generality, we can assume $\sigma_F = 1$. It is common for a distribution in the scale family to also be parameterized by a second value, a shape parameter. Applying the SAR in this setting is appropriate if the distribution is in a scale family whose range is $[0, \infty)$ and whose shape parameter remains fixed. As in the location-shift treatment effect setting, a larger observation will correspond to an improvement. With

| Distribution | Gamma | Weibull | Folded normal | Log logistic |
|---|---|---|---|---|
| Density function | $\frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1}\exp(-x/\beta)$ | $\frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}\exp\left(-(x/\beta)^{\alpha}\right)$ | $\sqrt{\frac{2}{\pi}}\exp\left(-\frac{x^2}{2\sigma^2}\right)$ | $\frac{\alpha x^{\alpha-1}\lambda}{[1+\lambda x^{\alpha}]^2}$ |

Table 4.1: Probability density functions of various scale family distributions. All parameters $> 0$.

these attributes, the distribution will be stochastically increasing with an increasing scale parameter, thus fulfilling the requirements to use the Wilcoxon Rank Sum test [2]. Since

Note that rather than an additive shift as in the previous location-shift setting, we are now working with a multiplicative treatment effect. Both the mean and variance of the responders will be greater than the mean and variance of the control group. However, the shape of the distribution is preserved when a random variable is multiplied by a constant. Simpson et al. [20] provide motivation for a multiplicative treatment effect with their investigation of data regarding rain clouds.

Some possible choices of $F$ are the gamma distribution with a given shape parameter, the Weibull distribution with a given shape parameter, the folded normal distribution, and the log logistic distribution for a given shape parameter. Their densities are presented in Table 4.1.

## 4.2 Test Statistic and Distribution

The test statistic used in this setting will be the SAR test statistic from Equation (2.2). With any choice of applicable distribution, the test statistic remains distribution-free under the null hypothesis and the Type I error is preserved, even if $F$ is misspecified. The selection of a shape parameter mirrors the assumed equal variance in the location shift setting. We are interested in testing the null hypothesis $H_0 : F = G$ against the one-sided

alternative, $H_A : G$ is the mixture distribution. We take $G(u) = (1-\theta)F(u)+\theta F(u/\delta)$ where $\theta \in (0, 1]$ is the proportion of responders and $\delta > 1$. The control group will be represented by $X$ as defined in Section 4.1. Then an observation from the treatment group will be $Y = (1-\theta)X + \theta\delta X$. For identifiability, define the null case as the point $(\theta, \delta) = (0, 1)$.

Under the null hypothesis, the SAR statistic maintains the same distribution as established in Chapter 2. Under the alternative hypothesis, the only change to the distribution is the calculation of the probabilities from Equations (2.5), (2.6), and (2.7). To account for the multiplicative treatment effect, they are now replaced with

$$p = \int_0^\infty F(u) \left[(1-\theta)f(u) + \frac{\theta}{\delta} f\left(\frac{u}{\delta}\right)\right] du \tag{4.1}$$

$$p_1 = \int_0^\infty \left[1 - (1-\theta)F(u) - \theta F\left(\frac{u}{\delta}\right)\right]^2 f(u)\ du \tag{4.2}$$

$$p_2 = \int_0^\infty (F(u))^2 \left[(1-\theta)f(u) + \frac{\theta}{\delta} f\left(\frac{u}{\delta}\right)\right] du \tag{4.3}$$

where $F$ is the CDF of the control group and $f$ is its density function. These probabilities are then used in the calculation of the mean, variance, and covariance of the SAR statistic established in Chapter 2. The asymptotic multivariate normal approximation still holds in this setting.

For the remainder of this work in the multiplicative treatment effect setting, we will focus on the gamma family of distributions to illustrate the properties of the SAR test statistic. A random variable with the gamma density from Table 4.1 has shape parameter $\alpha$ and scale parameter $\beta$. Using a different shape parameter for the design alternative will offer various possibilities of distributions to show the corresponding effects on the test.

## 4.3 Arm Size and Power

Tables 4.2 - 4.5 show arm sizes for various design alternatives with data originating from a gamma distribution as determined by the algorithms developed in Chapter 2. The arm sizes from the simulations are the result of running the simulation three times and displaying the largest arm size. This approach is likely the cause of the difference between the simulated arm sizes and those determined by the normal approximation (in parentheses), since the simulated result is always larger if the two methods are not equal.

As in the location-shift mixture setting, the arm size decreases as $\theta$ or $\delta$ increases. This aligns with the intuition that if there are more responders to treatment and they have much greater values than the control group, the treatment group will differ more from the control group. It can be seen from the tables that as the shape parameter is increasing, the arm size is decreasing. This brings about the same type of results seen from the ordering of the location-shift distributions. That is, the experimenter can expect to have higher power if the true shape parameter is greater than the shape parameter used for the design alternative because their arm size will be more than the minimum needed.

This relationship between the value of the shape parameter and the arm sizes is explained by the Kullback-Leibler (K-L) distances between the distributions under the null and alternative hypotheses. Figure 4.1 displays these distances for four gamma distributions with different shape parameters as $\theta$ changes. A $\delta$ of 1.5 was chosen and held fixed for the alternative distribution, although the same type of ordering occurs for any $\delta > 1$.

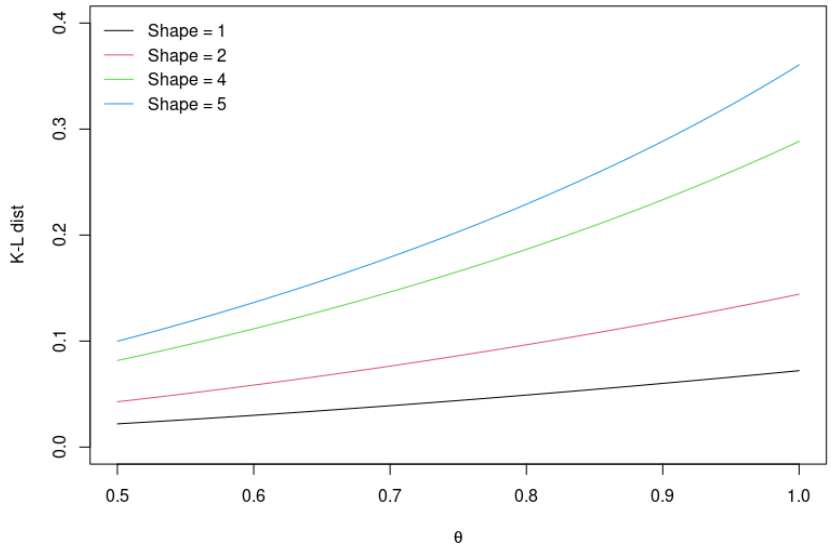


Figure 4.1: Kullback-Leibler distances comparing the null and mixture alternative with $\delta = 1.5$ and a range of $\theta$ for various gamma distributions with $\sigma_F = 1$.

Next, we investigate the power of the SAR procedure in the multiplicative treatment effect setting. Figure 4.2 presents the results of simulated power curves for four different shape parameter settings. Each design alternative is the same except for the shape parameter. Each plot shows power curves for a true $\theta$ less than, equal to, and greater than the design alternative $\theta$. The results match the expectation that the power will be higher if the true $\theta$ is higher than the design alternative $\theta$ and lower if the true $\theta$ is lower.

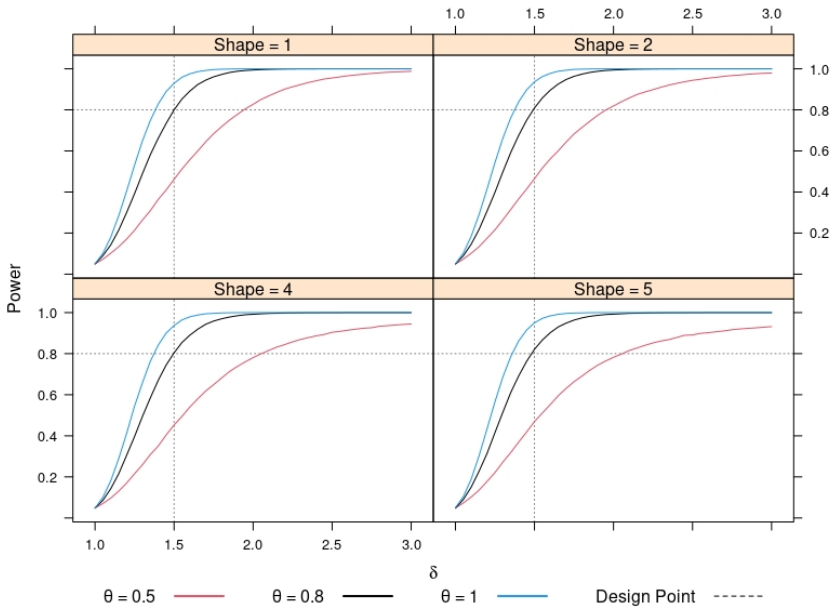


Figure 4.2: Power curves for various gamma distributions with $S = 2$, $\theta = 0.8$, $\delta = 1.5$, $\alpha = 0.05$, $\beta = 0.2$, and $\rho = 2$.

| | | $\delta$ | | | |
|---|---|---|---|---|---|
| $\theta$ | Stage | 1.25 | 1.5 | 1.75 | 2 |
| 0.5 | 1 | 1334 (1333) | 412 (410) | 221 (219) | 148 (146) |
| | 2 | 697 (697) | 216 (215) | 116 (116) | 78 (78) |
| | 3 | 477 (477) | 148 (147) | 80 (79) | 54 (53) |
| | 4 | 364 (363) | 113 (113) | 61 (61) | 41 (41) |
| | 5 | 295 (294) | 91 (91) | 50 (49) | 33 (33) |
| 0.6 | 1 | 928 (925) | 286 (284) | 153 (152) | 102 (101) |
| | 2 | 485 (484) | 149 (149) | 81 (80) | 54 (54) |
| | 3 | 332 (331) | 103 (103) | 56 (55) | 37 (37) |
| | 4 | 253 (252) | 79 (78) | 43 (42) | 29 (29) |
| | 5 | 205 (204) | 64 (63) | 35 (34) | 23 (23) |
| 0.7 | 1 | 681 (679) | 210 (208) | 112 (111) | 75 (73) |
| | 2 | 356 (356) | 110 (110) | 59 (59) | 40 (40) |
| | 3 | 243 (243) | 76 (75) | 41 (41) | 28 (27) |
| | 4 | 186 (186) | 58 (58) | 31 (31) | 21 (21) |
| | 5 | 151 (150) | 47 (47) | 26 (25) | 18 (17) |
| 0.8 | 1 | 521 (519) | 160 (159) | 86 (84) | 58 (56) |
| | 2 | 272 (272) | 84 (84) | 45 (45) | 31 (30) |
| | 3 | 186 (186) | 58 (58) | 31 (31) | 21 (21) |
| | 4 | 142 (142) | 44 (44) | 24 (24) | 17 (16) |
| | 5 | 115 (115) | 36 (36) | 20 (20) | 14 (13) |
| 0.9 | 1 | 412 (410) | 127 (125) | 68 (66) | 45 (44) |
| | 2 | 216 (215) | 67 (66) | 36 (36) | 24 (24) |
| | 3 | 148 (147) | 46 (46) | 25 (25) | 17 (17) |
| | 4 | 113 (113) | 35 (35) | 19 (19) | 13 (13) |
| | 5 | 91 (91) | 29 (29) | 16 (16) | 11 (11) |
| 1 | 1 | 333 (331) | 103 (101) | 55 (53) | 36 (35) |
| | 2 | 175 (174) | 54 (54) | 29 (29) | 20 (19) |
| | 3 | 120 (119) | 37 (37) | 21 (20) | 14 (14) |
| | 4 | 91 (91) | 29 (29) | 16 (16) | 11 (11) |
| | 5 | 74 (74) | 23 (23) | 13 (13) | 9 (9) |

Table 4.2: Arm sizes required to detect the mixture alternative where $\alpha = 0.05$, $\beta = 0.2$, $\rho = 2$, and $F$ is the gamma distribution with shape parameter equal to 1 using 1,000,000 simulations. Arm sizes determined by the normal approximation are in parentheses.

| | | $\delta$ | | | |
|---|---|---|---|---|---|
| $\theta$ | Stage | 1.25 | 1.5 | 1.75 | 2 |
| 0.5 | 1 | 598 (596) | 188 (186) | 103 (101) | 71 (69) |
| | 2 | 312 (312) | 99 (98) | 54 (54) | 38 (37) |
| | 3 | 214 (214) | 68 (68) | 38 (37) | 26 (26) |
| | 4 | 163 (163) | 52 (52) | 29 (29) | 20 (20) |
| | 5 | 132 (132) | 42 (42) | 24 (23) | 17 (16) |
| 0.6 | 1 | 415 (413) | 130 (128) | 72 (70) | 49 (47) |
| | 2 | 217 (217) | 69 (68) | 38 (38) | 26 (26) |
| | 3 | 149 (149) | 47 (47) | 26 (26) | 18 (18) |
| | 4 | 114 (113) | 36 (36) | 20 (20) | 14 (14) |
| | 5 | 92 (92) | 30 (29) | 17 (16) | 12 (11) |
| 0.7 | 1 | 305 (303) | 95 (94) | 52 (51) | 36 (34) |
| | 2 | 160 (159) | 50 (50) | 28 (28) | 19 (19) |
| | 3 | 109 (109) | 35 (35) | 19 (19) | 14 (13) |
| | 4 | 84 (83) | 27 (27) | 15 (15) | 11 (10) |
| | 5 | 68 (68) | 22 (22) | 13 (12) | 9 (9) |
| 0.8 | 1 | 233 (231) | 73 (71) | 40 (38) | 27 (26) |
| | 2 | 122 (122) | 39 (38) | 21 (21) | 15 (15) |
| | 3 | 84 (84) | 27 (27) | 15 (15) | 11 (10) |
| | 4 | 64 (64) | 21 (20) | 12 (12) | 8 (8) |
| | 5 | 52 (52) | 17 (17) | 10 (9) | 7 (7) |
| 0.9 | 1 | 184 (182) | 57 (56) | 31 (30) | 21 (20) |
| | 2 | 97 (96) | 31 (30) | 17 (17) | 12 (12) |
| | 3 | 67 (66) | 21 (21) | 12 (12) | 9 (8) |
| | 4 | 51 (51) | 17 (16) | 10 (9) | 7 (6) |
| | 5 | 41 (41) | 14 (13) | 8 (8) | 6 (5) |
| 1 | 1 | 149 (147) | 46 (45) | 25 (24) | 17 (16) |
| | 2 | 79 (78) | 25 (25) | 14 (14) | 10 (9) |
| | 3 | 54 (54) | 18 (17) | 10 (10) | 7 (7) |
| | 4 | 42 (41) | 14 (13) | 8 (8) | 6 (5) |
| | 5 | 34 (33) | 11 (11) | 7 (6) | 5 (4) |

Table 4.3: Arm sizes required to detect the mixture alternative where $\alpha = 0.05$, $\beta = 0.2$, $\rho = 2$, and $F$ is the gamma distribution with shape parameter equal to 2 using 1,000,000 simulations. Arm sizes determined by the normal approximation are in parentheses.

| | | $\delta$ | | | |
|---|---|---|---|---|---|
| $\theta$ | Stage | 1.25 | 1.5 | 1.75 | 2 |
| 0.5 | 1 | 286 (284) | 93 (91) | 53 (51) | 38 (36) |
| | 2 | 150 (149) | 49 (49) | 28 (28) | 20 (20) |
| | 3 | 103 (102) | 34 (34) | 20 (20) | 15 (14) |
| | 4 | 78 (78) | 26 (26) | 15 (15) | 11 (11) |
| | 5 | 64 (63) | 21 (21) | 13 (12) | 9 (9) |
| 0.6 | 1 | 198 (196) | 64 (62) | 37 (35) | 26 (25) |
| | 2 | 104 (104) | 34 (34) | 20 (20) | 14 (14) |
| | 3 | 71 (71) | 24 (23) | 14 (14) | 10 (10) |
| | 4 | 55 (54) | 18 (18) | 11 (11) | 8 (8) |
| | 5 | 44 (44) | 15 (15) | 9 (9) | 7 (6) |
| 0.7 | 1 | 145 (143) | 47 (45) | 27 (25) | 19 (18) |
| | 2 | 76 (76) | 25 (25) | 15 (14) | 11 (10) |
| | 3 | 53 (52) | 18 (17) | 11 (10) | 8 (7) |
| | 4 | 40 (40) | 14 (13) | 8 (8) | 6 (6) |
| | 5 | 33 (33) | 11 (11) | 7 (7) | 5 (5) |
| 0.8 | 1 | 111 (109) | 36 (34) | 20 (19) | 15 (13) |
| | 2 | 59 (58) | 19 (19) | 11 (11) | 8 (8) |
| | 3 | 40 (40) | 14 (13) | 8 (8) | 6 (6) |
| | 4 | 31 (31) | 11 (10) | 7 (6) | 5 (5) |
| | 5 | 25 (25) | 9 (9) | 6 (5) | 4 (4) |
| 0.9 | 1 | 88 (86) | 28 (26) | 16 (14) | 12 (10) |
| | 2 | 46 (46) | 16 (15) | 9 (9) | 7 (6) |
| | 3 | 32 (32) | 11 (11) | 7 (6) | 5 (5) |
| | 4 | 25 (25) | 9 (8) | 5 (5) | 4 (4) |
| | 5 | 20 (20) | 7 (7) | 5 (4) | 4 (3) |
| 1 | 1 | 71 (69) | 23 (21) | 13 (11) | 9 (8) |
| | 2 | 38 (37) | 12 (12) | 7 (7) | 6 (5) |
| | 3 | 26 (26) | 9 (9) | 5 (5) | 4 (4) |
| | 4 | 20 (20) | 7 (7) | 5 (4) | 4 (3) |
| | 5 | 17 (16) | 6 (6) | 4 (3) | 3 (3) |

Table 4.4: Arm sizes required to detect the mixture alternative where $\alpha = 0.05$, $\beta = 0.2$, $\rho = 2$, and $F$ is the gamma distribution with shape parameter equal to 4 using 1,000,000 simulations. Arm sizes determined by the normal approximation are in parentheses.

| | | $\delta$ | | | |
|---|---|---|---|---|---|
| $\theta$ | Stage | 1.25 | 1.5 | 1.75 | 2 |
| 0.5 | 1 | 227 (225) | 75 (73) | 44 (42) | 32 (31) |
| | 2 | 119 (119) | 40 (40) | 24 (23) | 17 (17) |
| | 3 | 82 (82) | 28 (27) | 17 (16) | 12 (12) |
| | 4 | 63 (62) | 21 (21) | 13 (13) | 10 (9) |
| | 5 | 51 (51) | 17 (17) | 11 (10) | 8 (8) |
| 0.6 | 1 | 158 (156) | 52 (50) | 31 (29) | 22 (21) |
| | 2 | 83 (83) | 28 (28) | 16 (16) | 12 (12) |
| | 3 | 57 (57) | 19 (19) | 12 (11) | 9 (8) |
| | 4 | 44 (44) | 15 (15) | 9 (9) | 7 (7) |
| | 5 | 36 (35) | 13 (12) | 8 (7) | 6 (6) |
| 0.7 | 1 | 116 (114) | 38 (36) | 22 (21) | 16 (15) |
| | 2 | 61 (61) | 20 (20) | 12 (12) | 9 (9) |
| | 3 | 42 (42) | 15 (14) | 9 (8) | 7 (6) |
| | 4 | 32 (32) | 11 (11) | 7 (7) | 5 (5) |
| | 5 | 26 (26) | 9 (9) | 6 (6) | 5 (4) |
| 0.8 | 1 | 88 (87) | 29 (27) | 17 (15) | 12 (11) |
| | 2 | 47 (47) | 16 (15) | 9 (9) | 7 (7) |
| | 3 | 32 (32) | 11 (11) | 7 (7) | 5 (5) |
| | 4 | 25 (25) | 9 (9) | 6 (5) | 4 (4) |
| | 5 | 21 (20) | 7 (7) | 5 (4) | 4 (3) |
| 0.9 | 1 | 70 (68) | 23 (21) | 13 (12) | 9 (8) |
| | 2 | 37 (37) | 12 (12) | 8 (7) | 6 (5) |
| | 3 | 26 (25) | 9 (9) | 6 (5) | 4 (4) |
| | 4 | 20 (20) | 7 (7) | 5 (4) | 4 (3) |
| | 5 | 16 (16) | 6 (6) | 4 (4) | 3 (3) |
| 1 | 1 | 56 (55) | 18 (17) | 11 (9) | 7 (6) |
| | 2 | 30 (30) | 10 (10) | 6 (6) | 5 (4) |
| | 3 | 21 (21) | 7 (7) | 5 (4) | 4 (3) |
| | 4 | 16 (16) | 6 (6) | 4 (3) | 3 (3) |
| | 5 | 14 (13) | 5 (5) | 4 (3) | 3 (2) |

Table 4.5: Arm sizes required to detect the mixture alternative where $\alpha = 0.05$, $\beta = 0.2$, $\rho = 2$, and $F$ is the gamma distribution with shape parameter equal to 5 using 1,000,000 simulations. Arm sizes determined by the normal approximation are in parentheses.

# Chapter 5

# Estimation

Under the traditional pure shift assumption, the treatment effect is $\delta$. However, in the mixture setting the treatment effect consists of the pair $(\theta, \delta)$. In this chapter, we consider several possible estimation techniques in both the location-shift and multiplicative treatment effect mixture settings. Each estimator will be calculated based on observations up to and including stage $s'$, the stage at which the trial ends. They will be evaluated on their bias and square root of their mean squared error.

## 5.1 Location-shift Mixture Alternative

In this section, we consider several possible estimation techniques applied in the location-shift mixture alternative setting.

### 5.1.1 Maximum Likelihood Estimation

Maximum likelihood estimators (MLEs) are a popular technique for their effectiveness in estimation. We obtain the MLEs from the pooled control and treatment groups by maximizing over all four parameters in the joint likelihood in Equation (5.1). Estimates are obtained for $\{\mu_F \in (-\infty, +\infty), \sigma_F \in (0, +\infty), \theta \in (0, 1], \delta \in (0, +\infty)\}$, although we are most interested in the estimates of $(\theta, \delta)$.

$$L(\mu_F, \sigma_F^2, \theta, \delta | \boldsymbol{x}, \boldsymbol{y}) = \left[\prod_{i=1}^{s'm} f(x_i; \mu_F, \sigma_F^2)\right] \left[\prod_{i=1}^{s'm} \left[(1-\theta) f(y_i; \mu_F, \sigma_F^2) + \theta f(y_i; \mu_F + \delta, \sigma_F^2)\right]\right] \tag{5.1}$$

The $F$ used for the likelihood will be the same as the one chosen for the design alternative. These estimates will be primarily considered as a baseline of comparison for the other estimators. This is partially due to the fact that we must choose an $F$ when calculating the MLEs, while the other estimators rely less on this choice or do not require it at all.

### 5.1.2 $k$-means

The second set of estimators for $(\theta, \delta)$ we will investigate are found after applying the popular unsupervised clustering method, $k$-means [13]. This clustering algorithm is designed to separate points into $k$ groups based on their distance from the center of the group. It begins with choosing the number of clusters. Since we are working under the assumption that the treatment observations come from a mixture of an $F$ and a shifted version of $F$, the number of clusters will be two. Then two observations are randomly chosen to start the clusters. They will function as the mean of the cluster on the first

iteration of the algorithm. The rest of the observations are assigned to the cluster of the nearest mean based on some distance metric. In our case, we calculate distance by using euclidean distance. Once all points are assigned, means are calculated from the clusters. These are used as the centers in the second iteration, and potentially new cluster assignments are made. The algorithm proceeds until there is no change in the assignment of the points.

The final solution of clusters should minimize the sum of the within-group sum of squares of the two clusters. However, the algorithm may have different results, depending on the starting values. In order to avoid a local minimum solution, we run the algorithm 50 times with random starting values each time and select the clusters that have the smallest total within-cluster sum of squares.

When running the $k$-means algorithm, we will only use the treatment group observations. Then we calculate the estimators by

$$\hat{\theta}_{k-means} = \frac{\text{number of observations in group with larger mean}}{\text{total number of observations}} \tag{5.2}$$

$$\hat{\delta}_{k-means} = \text{larger mean} - \text{smaller mean} \tag{5.3}$$

### 5.1.3 Constrained $k$-means

Next, we explore estimating the treatment effect using estimators calculated after applying the constrained $k$-means algorithm (CKM) proposed by Wagstaff et al. [23]. This clustering algorithm will run the traditional $k$-means algorithm with an additional set of constraints based on prior knowledge of the data. It takes as input a set of "must-link" observations and a set of "cannot-link" observations. The must-link constraints are

the observations the user identifies as being in the same cluster, while the cannot-link constraints are the observations the user identifies as being in different clusters. In the CKM algorithm, points are assigned to clusters in a way that does not violate the constraints. The only difference between the traditional $k$-means algorithm and the CKM algorithm is that before an observation is assigned to a cluster, the constraints are checked.

With the CKM algorithm, we are no longer limited to using only the treatment group observations as with traditional $k$-means. We input the control observations for the must-link constraints, since we know they are all from $F$. CKM will always keep these observations in the same cluster. We assign the largest observation from the treatment group and the smallest observation from the combined sample as the cannot-link constraint. The largest observation from the treatment group is likely a responder to the treatment that comes from the shifted version of $F$. The smallest observation from either the control treatment group is likely to come from the null distribution. Therefore these two observations should not be grouped into the same cluster. In order to avoid a local minimum solution, we run the algorithm 50 times with random starting values and choose the result with the smallest total within-cluster sum of squares.

Occasionally for two clusters, the cluster that includes the control group observations will also include the largest observations of the treatment group. This contradicts the assumption that the largest observations in the treatment group would come from a shifted version of $F$. However, it may suggest that there is no shift and give evidence of the null hypothesis. If CKM clusters the observations in this way, we will set the estimates for $(\theta, \delta)$ to $(0, 0)$ to represent the null scenario.

Under the mixture alternative, there should be two clusters. Let the cluster with the control group observations be Cluster 1 with mean $\overline{C}_1$ and the cluster without the control group observations be Cluster 2 with mean $\overline{C}_2$. Then we calculate the estimates for $(\theta, \delta)$ as

$$\hat{\theta}_{CKM} = \begin{cases} \frac{\text{\# of observations in Cluster 2}}{s'm} & \text{if } \overline{C}_2 - \overline{C}_1 > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

$$\hat{\delta}_{CKM} = \begin{cases} \overline{C}_2 - \overline{C}_1 & \text{if } \overline{C}_2 - \overline{C}_1 > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

### 5.1.4 Method of Moments

We consider a method of moments (MoM) estimator that builds upon the estimator proposed by Jeske and Yao [9]. These estimates will use the sample mean and sample variance of the control group $(\overline{X}, S_X^2)$, respectively, and treatment group, $(\overline{Y}, S_Y^2)$, respectively, from all observations up to and including the termination stage, $s'$.

Due to the randomness of the sample statistics, it is possible for the estimators to violate the parameter space. In order for the estimators to appropriately represent the null scenario, if either estimator is equal to zero then the estimate of the other will also be set to zero. For our MoM estimators, this situation would occur only if $\overline{Y} - \overline{X} \leq 0$ which would lead to the estimate of $\delta$ to be zero. Therefore, we modify the estimators from Jeske and Yao [9] such that they are calculated if $\overline{Y} - \overline{X} > 0$ and are set to zero otherwise.

Our MoM estimators for $(\theta, \delta)$ can be calculated as

$$\hat{\theta}_{MoM} = \begin{cases} \left[1 + \frac{(S_Y^2 - S_X^2)_+}{(\overline{Y}-\overline{X})^2 + \varepsilon_{s'}}\right]^{-1} & \text{if } \overline{Y} - \overline{X} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

$$\hat{\delta}_{MoM} = \begin{cases} (\overline{Y} - \overline{X}) \left[1 + \frac{(S_Y^2 - S_X^2)_+}{(\overline{Y}-\overline{X})^2 + \varepsilon_{s'}}\right] & \text{if } \overline{Y} - \overline{X} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

where $t_+ = t$ if $t > 0$ and 0 otherwise is used to restrict the estimates to the parameter space. For $\varepsilon_{s'}$, we will use an adaptive approach suggested by Lubich et al. [12]. It is calculated by

$$\varepsilon_{s'} = S_X^2 \frac{\log\left((2ms')^2\right)}{2ms'}. \tag{5.8}$$

Use of $\varepsilon_{s'}$ ensures consistency of the estimators and makes the estimators invariant to scale transformations.

### 5.1.5 Modified MoM

As noted by Jeske and Yao [9], the MoM estimator has worse performance when $F$ has heavy tails. In order to handle these situations, we make an attempt at a more robust version of the MoM estimator. First, we replace the sample means in (5.6) and (5.7) with sample medians. Then we replace the sample variances with functions involving the sample interquartile ranges (IQRs).

For the location-scale family, $X \sim F$ if and only if $X \sim \mu_F + \sigma_F Z$, where $Z \sim \Phi$. Then the $p$th percentile of $X$ is $X_p = \mu_F + \sigma_F Z_p$. Using this, it can be seen that the IQR

of $X$ is

$$\text{IQR}(X) = X_{75} - X_{25} = (\mu_F + \sigma_F Z_{75}) - (\mu_F + \sigma_F Z_{25}) = \sigma_F(Z_{75} - Z_{25}). \tag{5.9}$$

Therefore

$$\hat{\sigma}_F = \frac{\widehat{\text{IQR}}(X)}{Z_{75} - Z_{25}} \tag{5.10}$$

is an unbiased estimate of $\sigma_F$. We can square this to get an estimate of $\sigma_F^2$ for use in the MoM estimators. This will be a biased estimate of $\sigma_F^2$ due to Jensen's inequality, but should suffer less from heavy-tailed distributions.

This adjustment will work as a replacement for the sample variance of the control group, however the same cannot be done for the treatment group. Because of the mixture distribution, we are unable to factor out $\sigma_F$ in the IQR as in Equation (5.9). In our setting, we know that the variance of the treatment group is

$$\sigma_Y^2 = \sigma_F^2 + \theta(1-\theta)\delta^2. \tag{5.11}$$

We suggest using the square of Equation (5.10) as an estimate of the variance of the control group and using $(\hat{\theta}_{k-means}, \hat{\delta}_{k-means})$ as estimates of $(\theta, \delta)$. Even if the results of the $k$-means estimators may be inaccurate, they may only have a small effect on the modified MoM estimators.

| | Normal | Logistic | Laplace | $t_3$ |
|---|---|---|---|---|
| $Z_{75} - Z_{25}$ | 1.3490 | 1.2114 | 0.9802 | 0.8832 |

Table 5.1: IQR values for standardized location-scale distributions.

In our setting, we will have

$$\hat{\sigma}_{\text{IQR}} = \frac{\widehat{\text{IQR}}(X_{s'})}{Z_{75} - Z_{25}} \tag{5.12}$$

where $s'$ indicates we will use observations up to and including the termination stage and $Z \sim \Phi$. In our simulations, we choose $\Phi$ to match the choice of $F$ from the design alternative. Thus, this estimator depends on the choice of $F$. Table 5.1 shows the theoretical IQRs used in the denominator of Equation (5.10) for each $F$.

To accompany the use of a robust estimator for the spread of $F$, we will use the median as a robust estimator as the center of $F$. Let $\tilde{X}$ and $\tilde{Y}$ be the medians of the control and treatment observations up to and including stage $s'$, respectively. To keep the values of the estimators within the parameter space, the modified MoM estimators we use will be calculated only when the $\tilde{Y}$ is less than $\tilde{X}$. We set $(\hat{\theta}_{Mod.\ MoM}, \hat{\delta}_{Mod.\ MoM})$ to $(0, 0)$ if $\tilde{Y} - \tilde{X} < 0$, otherwise

$$\hat{\theta}_{Mod.\ MoM} = \left\{ 1 + \frac{(\hat{\sigma}^2_{\text{IQR}} + \hat{\theta}_{k-means}(1 - \hat{\theta}_{k-means})\hat{\delta}^2_{k-means} - \hat{\sigma}^2_{\text{IQR}})_+}{(\tilde{Y}_{s'} - \tilde{X}_{s'})^2 + \varepsilon_{\text{IQR}}} \right\}^{-1} \tag{5.13}$$

$$\hat{\delta}_{Mod.\ MoM} = (\tilde{Y}_{s'} - \tilde{X}_{s'}) \left\{ 1 + \frac{(\hat{\sigma}^2_{\text{IQR}} + \hat{\theta}_{k-means}(1 - \hat{\theta}_{k-means})\hat{\delta}^2_{k-means} - \hat{\sigma}^2_{\text{IQR}})_+}{(\tilde{Y}_{s'} - \tilde{X}_{s'})^2 + \varepsilon_{\text{IQR}}} \right\} \tag{5.14}$$

Since the adaptive epsilon from Equation (5.8) uses the sample standard deviation from the control group, we replace this with Equation (5.12) to get

$$\varepsilon_{\text{IQR}} = \hat{\sigma}^2_{\text{IQR}} \frac{\log\left((2ms')^2\right)}{2ms'}. \tag{5.15}$$

### 5.1.6 Bootstrap Bias-corrected MoM

By only calculating the estimate upon termination, the estimator will be biased [8, 16]. To adjust for this, we apply the bootstrap bias correction used by Peng et al. [16] to our MoM estimators. In order to avoid relying on a choice of $F$, we will use a nonparametric bootstrap sampling technique in place of parametric sampling.

After the trial ends, we obtain the MoM estimates of $(\theta, \delta)$ and proceed with the bootstrap bias correction if the estimates are not $(0, 0)$. For the bootstrap sampling, we calculate the empirical cumulative distribution function, $\hat{F}$, using only observations from the control group up to and including the stage at which the experiment terminated. Each bootstrap sample will simulate a group sequential trial of its own, resulting in bootstrap samples that may terminate earlier or later than the original sample.

Nonparametric bootstrap sampling is commonly done by resampling from the original sample, or equivalently $\hat{F}$, with replacement. However, resampling using $\hat{F}$ can result in ties in the bootstrap sample. This would be a violation of our assumptions for the test, namely that $F$ is continuous and therefore the probability of one observation being equal to a specific value is zero. We will use a method that can mimic simulating from a continuous distribution. Let $X_{(1)}, X_{(2)}, \ldots, X_{(s'm-1)}, X_{(s'm)}$ be the order statistics of the control group. We use Ogive: linear interpolations between the points $(X_{(i)}, \hat{F}(X_{(i)}))$ and

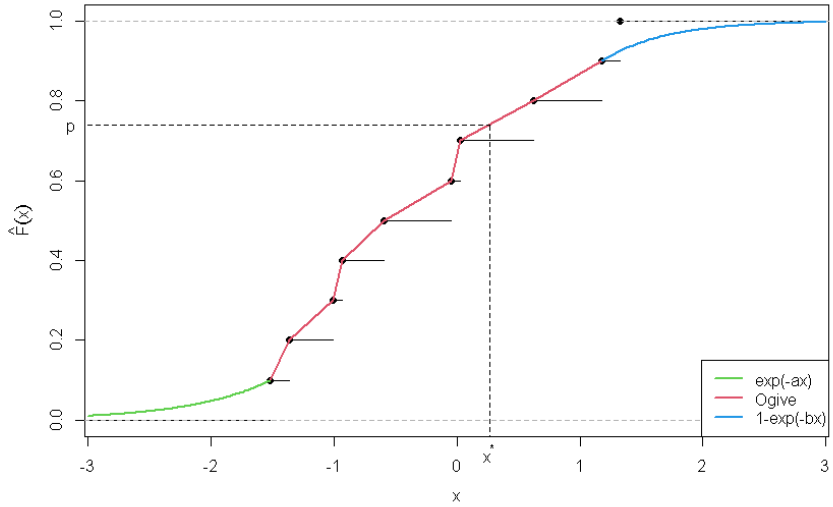


Figure 5.1: $\hat{F}$ with linear interpolation (Ogive) between points and exponential functions in the tails and generation of a bootstrap sample value. An example bootstrap sample value, $x^*$, is shown.

$(X_{(i+1)}, \hat{F}(X_{(i+1)}))$ for $i = 1, \ldots, s'm - 2$. For the tail behavior of $\hat{F}$, we will apply the framework of the location-scale families used previously and assume the range of the data is $(-\infty, +\infty)$. In the lower tail, we plot the function $f(x) = \exp(-ax)$ such that it goes through the point $(X_{(1)}, 1/(s'm))$. Then, this can be used to solve for $a$. Similarly, in the upper tail we plot the function $g(x) = 1 - \exp(-bx)$ such that it goes through the point $(X_{(s'm-1)}, (s'm-1)/(s'm))$. Then solve for $b$ in the same fashion as $a$. Figure 5.1 provides a visual of $\hat{F}$ with the Ogive to connect the points and exponential functions for the tails.

To obtain a value, $x^*$, for a bootstrap sample, we can create a plot as in Figure 5.1 using $X_1, \ldots, X_{s'm}$. Generate $p \in (0,1)$. Then

$$x^* = \begin{cases} -\log(p)/a & \text{if } p < \hat{F}(X_{(1)}) \\ (p - b_{0i})/b_{1i} & \text{if } \hat{F}(X_{(i)}) \leq p \leq \hat{F}(X_{(i+1)}),\ i = 1, \ldots, s'm-2 \\ -\log(1-p)/b & \text{if } p > \hat{F}(X_{(s'm-1)}) \end{cases} \tag{5.16}$$

where $(b_{0i}, b_{1i})$ are the intercept and slope of the line connecting $(X_{(i)}, \hat{F}(X_{(i)}))$ and $(X_{(i+1)}, \hat{F}(X_{(i+1)}))$ for $i = 1, \ldots, s'm-2$. All of the observations for the bootstrap sample for the control group can be generated in this way.

For a bootstrap sample value for the treatment group, $y^*$, we first find a value $x^*$ as above. Then

$$y^* = \begin{cases} x^* & \text{with probability } 1 - \hat{\theta}_{MoM} \\ x^* + \hat{\delta}_{MoM} & \text{with probability } \hat{\theta}_{MoM} \end{cases} \tag{5.17}$$

We use this process to generate all bootstrap treatment observations.

Once we have a terminated bootstrap trial, we calculate bootstrap estimates $(\hat{\theta}^*, \hat{\delta}^*)$ using the MoM estimators. Then we adjust these to correct for their bias. The algorithm for calculating the bootstrap bias corrected estimates of $(\theta, \delta)$ is outlined with the following steps:

1. Generate $B$ bootstrap samples until termination.

2. Calculate method of moments estimates of $(\theta, \delta)$ for each sample path, denoted $(\hat{\theta}^*_j, \hat{\delta}^*_j)$, $j = 1, \ldots, B$.

3. Bootstrap bias corrected MoM estimates of $(\theta, \delta)$ are

$$\hat{\theta}_{BC\ MoM} = 2\hat{\theta}_{MoM} - \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_j^* \tag{5.18}$$

$$\hat{\delta}_{BC\ MoM} = 2\hat{\delta}_{MoM} - \frac{1}{B}\sum_{j=1}^{B}\hat{\delta}_j^* \tag{5.19}$$

### 5.1.7 Results Comparisons

This section will discuss the results of simulating estimates of $(\theta, \delta)$ using the estimators from the previous sections. They will be evaluated on their bias and square root of their mean squared error (RMSE). The process will involve simulating a group sequential trial until termination, then calculating the estimates using the observations up to and including the termination stage. We present the results such that a specific group sequential design alternative is set and the true $\theta$, $\delta$, and/or $F$ may be different than the values chosen for the design alternative.

First, consider the estimation of $\theta$ as presented in Table 5.2. Notably, the $k$-means estimator does well when the true $\theta$ is around 0.5. However, as $\theta$ moves closer to one the $k$-means estimator is no longer preferred. In empirical results that are not shown, the same effect happens when $\theta$ moves closer to zero. For the cases where $\theta = 0.7$ and $\theta = 0.9$, we see some mixed results for the best estimator. Generally, the MoM estimator is either the best or second-best choice. As $\theta$ increases with all else equal, the bias and RMSE of the $k$-means and CKM estimators increase. In the same setting, the RMSE of the Mod. MoM generally increases. The RMSE of the MLE, MoM, and BC MoM decreases as $\theta$ increases. When estimating $\theta$, all estimators generally see their bias and RMSE decrease as the size

of the shift increases with all else equal. The $k$-means-based estimators seem to attempt to split the observations into roughly equal-sized clusters, whereas the other estimators seem to perform better if there is less mixing of the two components of the mixture distribution.

A final consideration is the effect of $F$. The bias of the MLE has mixed results on whether heavy-tailed distributions increase its bias. For large $\theta$, the RMSE of the MLE decreases for heavy-tailed distributions. The $F$ does not appear to have much of an effect on the $k$-means estimator. The bias of the CKM estimator sees an increase for heavy tails only in small shift scenarios, while the RMSE has an overall increase with heavy-tailed $F$. Similarly, the MoM estimator's bias increases with the heaviness of the tails of $F$ and large $\theta$. The RMSE sees and overall increase with heavy-tailed $F$. The Mod. MoM bias decreases with heavier-tailed $F$. The RMSE is generally decreasing, but sometimes has a slightly increase for the $t_3$ distribution which has the heaviest tails. Bias of the BC MoM does not see much change with $F$, but the RMSE is increasing with $F$ and large $\theta$.

Next, we compare the estimators of $\delta$. The results can be seen in Table 5.3. We quickly see that the $k$-means and CKM estimators are generally among the poorest of the six choices by both metrics. The MLE appears as the best choice in many of the scenarios, especially when the shift is large. Once again the MoM offers several instances of being the second best estimator mixed with some situations as the best choice based on either the bias or RMSE.

Similar patterns that are present in the estimation of $\theta$ also appear when estimating $\delta$. We examine the tables as $\theta$ increases with all else equal. For both the MLE and MoM, the bias does not seem to exhibit any particular pattern, but the RMSE decreases as $\theta$

increases. The $k$-means estimator sees its bias decrease. The CKM estimator shows a generally increasing bias. Both the $k$-means and CKM estimators show no pattern in the RMSE with increasing $\theta$. The modified MoM and BC MoM both generally show an increase in bias and decrease in RMSE. When we consider increasing $\delta$ with all else equal, all estimators generally see a decrease in both bias and RMSE. This is unsurprising as we expect the larger shift to make it easier to detect the centers of the null and shifted distributions.

Now, consider the effects of moving from the light-tailed normal distribution to the heavy-tailed $t_3$ distribution. The notable effects on the bias are seen in the $k$-means estimator and BC MoM estimator which both exhibit a general increase. The MoM is unique and seems to show bias decreasing with heavier tails and $\theta = 0.5$ but increasing bias otherwise. All estimators except the Mod. MoM generally show monotone increasing of the RMSE. The Mod. MoM actually appears to be decreasing until it reaches the $t_3$ where it too succumbs to the heaviest tails.

Another situation we wish to consider is when the true $(\theta, \delta)$ are $(0, 0)$ to investigate the performance of the estimators if the null hypothesis were true. Table 5.4 presents the bias and RMSE results from simulations of group sequential trials for some design alternative but the true $(\theta, \delta)$ were $(0, 0)$. In this setting, although we define the hull hypothesis as the point $(0, 0)$, either $\theta = 0$ or $\delta = 0$ is essentially the null scenario. Thus, a pair of estimators only correctly identifying one of the parameters as zero is evidence of the null. With this in mind, it seems that the Mod. MoM has the best performance for estimating $\theta$ as zero, while the MoM has the best performance for estimating $\delta$ as zero.

We present some final thoughts on the estimators in the mixture alternative setting where the responder distribution is a location-shift of the control distribution. First, we note that the MLE is not always the best estimator as we may have expected. Lindsay [11] provides some evidence for this by explaining that when the shift is less than two standard deviations, there is very little information about $\theta$. Furthermore, the MLE relies on the design alternative choice of $F$. Both the $k$-means and CKM estimators often estimate $\theta$ as 0.5 even when the true value is much different. This then leads to the downfall of its estimate of $\delta$. If the experimenter had prior knowledge that $\theta$ may be around 0.5, they may be interested in using the $k$-means or CKM estimators. The Modified MoM estimators exhibited decreasing in both bias and RMSE as the tails of the distribution became increasingly heavy. This suggests that using robust estimates of the center and spread of the control and treatment distributions can help. Unfortunately, it too even struggled with the $t_3$ distribution which had the heaviest tails of all the $F$'s considered. Like the MLE, the Mod. MoM depends on the choice of $F$ for the design alternative. The BC MoM has mixed results on effectively reducing the bias of the MoM estimator. Whether the bias reduction is present or not, use of the BC MoM comes with the cost of increased RMSE and computation time.

Overall, we recommend the use of the MoM estimator for both the pair $(\theta, \delta)$. It is one of the most competitive estimators, often showing as the best or second best choice by either bias or RMSE for $(\theta, \delta)$. It also does not depend on the choice of $F$ in the design alternative.

| | | | Bias | | | | RMSE | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | $\delta$ | Estimator | Normal | Logistic | Laplace | $t_3$ | Normal | Logistic | Laplace | $t_3$ |
| 0.5 | 1 | MLE | 0.1074 | 0.0957 | 0.1147 | 0.1587 | 0.3522 | 0.3510 | 0.3715 | 0.3896 |
| | | k-means | **-0.0054** | **0.0009** | **-0.0015** | **0.0042** | **0.1454** | **0.1596** | **0.1833** | **0.2222** |
| | | CKM | -0.1556 | -0.1734 | -0.1906 | -0.2101 | *0.2344* | *0.2551* | 0.2825 | 0.3057 |
| | | MoM | 0.1204 | 0.0908 | 0.0955 | 0.0907 | 0.3553 | 0.3460 | 0.3574 | 0.3755 |
| | | Mod. MoM | -0.1835 | -0.1763 | -0.1517 | -0.1234 | 0.2761 | 0.2665 | *0.2468* | *0.2351* |
| | | BC MoM | *0.0466* | *0.0216* | *0.0273* | *0.0360* | 0.4175 | 0.3999 | 0.4208 | 0.4282 |
| | 1.5 | MLE | 0.0742 | 0.0997 | 0.1117 | 0.1194 | 0.2829 | 0.2995 | 0.3152 | 0.3119 |
| | | k-means | **-0.0036** | **-0.0037** | **-0.0064** | **-0.0054** | **0.1304** | **0.1345** | **0.1418** | **0.1736** |
| | | CKM | -0.0870 | -0.0938 | -0.1013 | -0.1175 | *0.1875* | *0.2014* | *0.2218* | 0.2393 |
| | | MoM | 0.0962 | 0.1039 | 0.0947 | 0.0824 | 0.2841 | 0.2956 | 0.3017 | 0.3012 |
| | | Mod. MoM | -0.1012 | -0.0830 | -0.0856 | -0.0651 | 0.2283 | 0.2229 | 0.2296 | *0.2143* |
| | | BC MoM | *0.0364* | *0.0500* | *0.0452* | *0.0411* | 0.3377 | 0.3431 | 0.3447 | 0.3394 |
| | 2 | MLE | 0.0530 | 0.0635 | 0.0927 | 0.0896 | 0.2223 | 0.2314 | 0.2562 | 0.2457 |
| | | k-means | **0.0019** | **-0.0046** | **-0.0002** | **0.0017** | **0.1227** | **0.1216** | **0.1203** | **0.1379** |
| | | CKM | -0.0410 | -0.0380 | -0.0442 | -0.0523 | *0.1433* | *0.1501* | *0.1662* | *0.1842* |
| | | MoM | 0.0533 | 0.0701 | 0.0773 | 0.0675 | 0.2252 | 0.2344 | 0.2453 | 0.2527 |
| | | Mod. MoM | -0.0531 | -0.0434 | *-0.0346* | *-0.0353* | 0.2116 | 0.2117 | 0.2131 | 0.2208 |
| | | BC MoM | *0.0085* | *0.0342* | 0.0434 | 0.0415 | 0.2668 | 0.2634 | 0.2710 | 0.2711 |
| 0.7 | 1 | MLE | **0.0015** | *0.0156* | 0.0561 | 0.0862 | 0.2832 | 0.2874 | 0.2872 | *0.2725* |
| | | k-means | -0.1839 | -0.1891 | -0.1718 | -0.1796 | **0.2301** | **0.2476** | **0.2649** | 0.2941 |
| | | CKM | -0.2807 | -0.3034 | -0.3226 | -0.3152 | 0.3290 | 0.3617 | 0.4012 | 0.4037 |
| | | MoM | *0.0168* | **0.0099** | **0.0169** | *0.0146* | *0.2694* | *0.2789* | 0.2907 | 0.2916 |
| | | Mod. MoM | -0.2632 | -0.2429 | -0.2009 | -0.1635 | 0.3307 | 0.3094 | *0.2669* | **0.2427** |
| | | BC MoM | -0.0270 | -0.0284 | *-0.0181* | **-0.0108** | 0.3264 | 0.3329 | 0.3349 | 0.3253 |
| | 1.5 | MLE | *0.0453* | 0.0649 | 0.0802 | 0.0848 | 0.2130 | 0.2226 | 0.2235 | *0.2135* |
| | | k-means | -0.1647 | -0.1519 | -0.1320 | -0.1338 | 0.2161 | *0.2174* | *0.2128* | 0.2447 |
| | | CKM | -0.1816 | -0.1885 | -0.1749 | -0.1904 | 0.2371 | 0.2623 | 0.2666 | 0.3070 |
| | | MoM | 0.0457 | *0.0434* | *0.0498* | *0.0256* | *0.2085* | 0.2279 | 0.2273 | 0.2349 |
| | | Mod. MoM | -0.1109 | -0.0983 | -0.0718 | -0.0480 | **0.2073** | **0.1992** | **0.1718** | **0.1635** |
| | | BC MoM | **0.0252** | **0.0222** | **0.0288** | **0.0094** | 0.2260 | 0.2496 | 0.2509 | 0.2525 |
| | 2 | MLE | 0.0562 | 0.0552 | 0.0632 | 0.0623 | 0.1756 | *0.1698* | *0.1764* | *0.1757* |
| | | k-means | -0.1309 | -0.1114 | -0.0922 | -0.0834 | 0.1949 | 0.1856 | 0.1779 | 0.1922 |
| | | CKM | -0.1118 | -0.1101 | -0.0983 | -0.1062 | *0.1750* | 0.1908 | 0.1980 | 0.2314 |
| | | MoM | 0.0476 | 0.0406 | 0.0495 | 0.0325 | 0.1752 | 0.1768 | 0.1872 | 0.1963 |
| | | Mod. MoM | **-0.0303** | **-0.0182** | **0.0014** | **0.0058** | **0.1437** | **0.1327** | **0.1247** | **0.1349** |
| | | BC MoM | *0.0318* | *0.0266* | *0.0359* | *0.0210* | 0.1860 | 0.1836 | 0.1958 | 0.2043 |
| 0.9 | 1 | MLE | *-0.0725* | **-0.0518** | **-0.0272** | **-0.0046** | *0.2374* | **0.2146** | **0.1802** | **0.1657** |
| | | k-means | -0.3979 | -0.3859 | -0.3871 | -0.3996 | 0.4246 | 0.4251 | 0.4566 | 0.4805 |
| | | CKM | -0.4134 | -0.4363 | -0.4463 | -0.4598 | 0.4461 | 0.4835 | 0.5154 | 0.5424 |
| | | MoM | **-0.0597** | *-0.0690* | *-0.0815* | *-0.0906* | **0.2030** | *0.2197* | *0.2287* | *0.2559* |
| | | Mod. MoM | -0.3252 | -0.3081 | -0.2688 | -0.2466 | 0.3736 | 0.3515 | 0.3028 | 0.2984 |
| | | BC MoM | -0.0787 | -0.0902 | -0.0958 | -0.1035 | 0.2378 | 0.2685 | 0.2507 | 0.2802 |
| | 1.5 | MLE | **-0.0128** | **-0.0050** | **0.0041** | **0.0072** | *0.1387* | **0.1301** | **0.1234** | **0.1191** |
| | | k-means | -0.3710 | -0.3608 | -0.3431 | -0.3374 | 0.4036 | 0.4104 | 0.4191 | 0.4240 |
| | | CKM | -0.2992 | -0.3051 | -0.3039 | -0.2859 | 0.3379 | 0.3625 | 0.3962 | 0.3936 |
| | | MoM | *-0.0179* | *-0.0294* | *-0.0339* | *-0.0495* | **0.1325** | *0.1480* | *0.1614* | 0.1905 |
| | | Mod. MoM | -0.1722 | -0.1609 | -0.1429 | -0.1285 | 0.2105 | 0.1963 | 0.1777 | *0.1827* |
| | | BC MoM | -0.0248 | -0.0368 | -0.0415 | -0.0558 | 0.1423 | 0.1586 | 0.1755 | 0.2036 |
| | 2 | MLE | **-0.0092** | **-0.0029** | **-0.0019** | **0.0060** | *0.1229* | **0.1101** | **0.1022** | **0.0938** |
| | | k-means | -0.3356 | -0.3226 | -0.2940 | -0.2769 | 0.3760 | 0.3772 | 0.3733 | 0.3747 |
| | | CKM | -0.2169 | -0.2031 | -0.1917 | -0.1682 | 0.2568 | 0.2668 | 0.2837 | 0.2923 |
| | | MoM | *-0.0160* | *-0.0189* | *-0.0253* | *-0.0327* | **0.1200** | *0.1243* | 0.1360 | 0.1565 |
| | | Mod. MoM | -0.1028 | -0.0911 | -0.0835 | -0.0731 | 0.1401 | 0.1263 | *0.1173* | *0.1314* |
| | | BC MoM | -0.0215 | -0.0228 | -0.0291 | -0.0372 | 0.1335 | 0.1309 | 0.1427 | 0.1733 |

Table 5.2: Bias and RMSE values for estimating $\theta$ based on 1000 simulations with 1000 bootstrap sample paths where the data come from the $F$, $\theta$, and $\delta$ values indicated in the table with $\sigma_F = 1$. The design alternative uses $S = 2$, $\theta = 0.7$, $K = 1.5$, $F =$ Normal, $\alpha = 0.01$, $\beta = 0.1$, and $\rho = 2$ resulting in an arm size of 17.

| | | | Bias | | | | RMSE | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | $\delta$ | Estimator | Normal | Logistic | Laplace | $t_3$ | Normal | Logistic | Laplace | $t_3$ |
| 0.5 | 1 | MLE | **0.0011** | **0.0334** | 0.1164 | *0.0378* | 0.7938 | 0.9689 | 1.0517 | 1.2303 |
| | | k-means | 0.8667 | 0.8820 | 0.9357 | 1.2552 | 0.9302 | 0.9999 | 1.1393 | 2.2790 |
| | | CKM | 0.5753 | 0.5260 | 0.5100 | 0.6038 | 0.8773 | 0.9853 | 1.1241 | 1.7572 |
| | | MoM | -0.2155 | -0.1377 | *-0.0534* | **0.0181** | **0.5884** | **0.6245** | *0.7094* | **0.9037** |
| | | Mod. MoM | 0.2228 | 0.2550 | 0.2842 | 0.2804 | 0.7498 | 0.7228 | **0.6682** | *1.0353* |
| | | BC MoM | *-0.1364* | *-0.0469* | **0.0498** | 0.1339 | *0.6500* | *0.7186* | 0.8525 | 1.1144 |
| | 1.5 | MLE | **0.0082** | **-0.0141** | *-0.0097* | **-0.0043** | **0.6039** | 0.6814 | 0.7979 | *0.9645* |
| | | k-means | 0.5819 | 0.5599 | 0.5853 | 0.7503 | 0.6709 | 0.6636 | 0.7892 | 1.6906 |
| | | CKM | 0.3705 | 0.3700 | 0.3518 | 0.4050 | 0.7053 | 0.8134 | 0.9586 | 1.5185 |
| | | MoM | -0.1679 | -0.1365 | -0.0916 | *0.0616* | *0.6276* | *0.6602* | *0.7153* | 1.0136 |
| | | Mod. MoM | 0.2251 | 0.2074 | 0.1967 | 0.1926 | 0.6661 | **0.6291** | **0.6602** | **0.8539** |
| | | BC MoM | *-0.0809* | *-0.0436* | **0.0008** | 0.1727 | 0.6644 | 0.7655 | 0.8209 | 1.2956 |
| | 2 | MLE | **-0.0232** | **-0.0183** | *-0.0535* | *-0.0569* | **0.5098** | *0.5473* | 0.6564 | **0.7474** |
| | | k-means | 0.3862 | 0.3653 | 0.3706 | 0.5025 | *0.5200* | **0.5093** | **0.5745** | 1.7330 |
| | | CKM | 0.2598 | 0.3002 | 0.2951 | 0.2928 | 0.5497 | 0.6299 | 0.8387 | 1.3145 |
| | | MoM | -0.1204 | -0.1032 | -0.0877 | **0.0159** | 0.6392 | 0.6419 | 0.7096 | *0.8898* |
| | | Mod. MoM | 0.1601 | 0.1699 | 0.1412 | 0.1922 | 0.6516 | 0.6317 | *0.6349* | 1.6528 |
| | | BC MoM | *-0.0295* | *-0.0355* | **-0.0161** | 0.0879 | 0.7209 | 0.7052 | 0.8171 | 1.0771 |
| 0.7 | 1 | MLE | 0.1249 | 0.1575 | *0.0945* | **0.0824** | 0.5742 | 0.7479 | 0.7524 | **0.8201** |
| | | k-means | 0.8314 | 0.8516 | 0.9163 | 1.1709 | 0.8853 | 0.9620 | 1.1498 | 2.1862 |
| | | CKM | 0.6300 | 0.6331 | 0.6184 | 0.7780 | 0.8390 | 1.0050 | 1.1883 | 2.0327 |
| | | MoM | **0.0191** | **0.0622** | **0.0923** | *0.2718* | **0.4787** | **0.5294** | **0.5887** | 1.1048 |
| | | Mod. MoM | 0.4859 | 0.5144 | 0.4773 | 0.4998 | 0.7085 | 0.6986 | *0.6318* | *1.0227* |
| | | BC MoM | *0.0665* | *0.1074* | 0.1488 | 0.3409 | *0.5176* | *0.5848* | 0.6914 | 1.3874 |
| | 1.5 | MLE | 0.0700 | **0.0218** | **-0.0001** | **0.0027** | **0.4402** | **0.4615** | **0.4860** | **0.5501** |
| | | k-means | 0.5335 | 0.5398 | 0.5677 | 0.8217 | 0.6330 | 0.7154 | 0.8777 | 1.9417 |
| | | CKM | 0.5026 | 0.4671 | 0.4336 | 0.5015 | 0.6678 | 0.8177 | 0.9101 | 1.5170 |
| | | MoM | **0.0487** | *0.0505* | *0.0512* | *0.2179* | *0.4709* | *0.5018* | *0.5176* | 0.9961 |
| | | Mod. MoM | 0.4520 | 0.3964 | 0.3836 | 0.3908 | 0.5820 | 0.5581 | 0.5300 | *0.7217* |
| | | BC MoM | *0.0676* | 0.0758 | 0.0756 | 0.2606 | 0.4821 | 0.5352 | 0.5538 | 1.3181 |
| | 2 | MLE | **-0.0014** | **0.0294** | **0.0165** | **-0.0174** | **0.4023** | **0.4173** | *0.5196* | **0.5132** |
| | | k-means | 0.2783 | 0.3435 | 0.3530 | 0.4815 | 0.5123 | 0.6217 | 0.7033 | 1.3866 |
| | | CKM | 0.3249 | 0.3637 | 0.3649 | 0.3643 | 0.5061 | 0.6982 | 0.8547 | 1.2377 |
| | | MoM | *0.0041* | *0.0642* | *0.0467* | *0.0920* | *0.4322* | *0.4784* | 0.5222 | 0.7468 |
| | | Mod. MoM | 0.2965 | 0.3482 | 0.3521 | 0.3327 | 0.4860 | 0.5023 | **0.4850** | *0.6034* |
| | | BC MoM | 0.0157 | 0.0697 | 0.0519 | 0.0995 | 0.4409 | 0.4824 | 0.5249 | 0.8421 |
| 0.9 | 1 | MLE | 0.2032 | **0.1490** | **0.1804** | **0.1389** | 0.4766 | *0.4905* | **0.4597** | **0.5885** |
| | | k-means | 0.7489 | 0.7787 | 0.9310 | 1.2112 | 0.8118 | 0.9503 | 1.1939 | 2.5463 |
| | | CKM | 0.7365 | 0.6347 | 0.7646 | 0.8734 | 0.8692 | 0.9329 | 1.2168 | 1.9201 |
| | | MoM | **0.1553** | *0.1659* | *0.2710* | *0.3526* | **0.4312** | **0.4811** | *0.5522* | *0.9795* |
| | | Mod. MoM | 0.6021 | 0.5558 | 0.5673 | 0.6048 | 0.7130 | 0.6569 | 0.6470 | 1.8585 |
| | | BC MoM | *0.1694* | 0.1807 | 0.2860 | 0.3901 | *0.4432* | 0.4984 | 0.5755 | 1.2697 |
| | 1.5 | MLE | **0.1187** | **0.1053** | **0.0965** | **0.0819** | **0.3816** | **0.3632** | **0.3884** | **0.5134** |
| | | k-means | 0.3371 | 0.4321 | 0.5461 | 0.6685 | 0.5115 | 0.7619 | 0.9962 | 1.7338 |
| | | CKM | 0.5292 | 0.5588 | 0.5411 | 0.5899 | 0.6446 | 0.8793 | 1.0528 | 1.5916 |
| | | MoM | *0.1260* | *0.1512* | *0.1683* | *0.2484* | *0.3946* | *0.4342* | *0.4884* | 0.8771 |
| | | Mod. MoM | 0.4460 | 0.4365 | 0.4168 | 0.4200 | 0.5549 | 0.5466 | 0.5197 | *0.7679* |
| | | BC MoM | 0.1275 | 0.1530 | 0.1704 | 0.2629 | 0.3972 | 0.4372 | 0.4914 | 1.0648 |
| | 2 | MLE | *0.0529* | **0.0655** | **0.0676** | **0.0607** | **0.3533** | **0.3633** | **0.3420** | **0.3200** |
| | | k-means | **0.0227** | *0.1012* | 0.2371 | 0.3785 | 0.5338 | 0.7002 | 0.8874 | 1.7632 |
| | | CKM | 0.3249 | 0.3379 | 0.3650 | 0.4073 | 0.4671 | 0.6423 | 0.7971 | 1.4140 |
| | | MoM | 0.0663 | 0.1048 | 0.1185 | *0.2277* | 0.3772 | 0.4282 | 0.4535 | 1.1534 |
| | | Mod. MoM | 0.3312 | 0.3625 | 0.3521 | 0.3664 | 0.4993 | 0.5085 | 0.4704 | *0.8198* |
| | | BC MoM | 0.0652 | 0.1014 | *0.1143* | 0.2535 | *0.3736* | *0.4250* | *0.4494* | 1.5974 |

Table 5.3: Bias and RMSE values for estimating $\delta$ based on 1000 simulations with 1000 bootstrap sample paths where the data come from the $F$, $\theta$, and $\delta$ values indicated in the table with $\sigma_F = 1$. The design alternative uses $S = 2$, $\theta = 0.7$, $K = 1.5$, $F =$ Normal, $\alpha = 0.01$, $\beta = 0.1$, and $\rho = 2$ resulting in an arm size of 17.

| Parameter | Statistic | Estimator | Normal | Logistic | Laplace | $t_3$ |
|---|---|---|---|---|---|---|
| $\theta$ | Bias | MLE | 0.6062 | 0.5640 | 0.5418 | 0.5699 |
| | | $k$-means | 0.4898 | 0.4929 | 0.4797 | 0.5037 |
| | | CKM | *0.2021* | *0.1771* | *0.1364* | **0.1394** |
| | | MoM | 0.3351 | 0.3548 | 0.3208 | 0.2895 |
| | | Mod. MoM | **0.1274** | **0.1416** | **0.1360** | *0.1434* |
| | | BC MoM | 0.2582 | 0.2874 | 0.2425 | 0.2198 |
| | RMSE | MLE | 0.7229 | 0.7010 | 0.6891 | 0.7136 |
| | | $k$-means | 0.5135 | 0.5312 | 0.5444 | 0.5734 |
| | | CKM | *0.2647* | *0.2394* | **0.1939** | **0.2034** |
| | | MoM | 0.5368 | 0.5582 | 0.5280 | 0.5016 |
| | | Mod. MoM | **0.2077** | **0.2218** | *0.2241* | *0.2441* |
| | | BC MoM | 0.5214 | 0.5458 | 0.5251 | 0.5016 |
| $\delta$ | Bias | MLE | **0.1089** | **0.0900** | **0.2532** | **0.1421** |
| | | $k$-means | 1.6954 | 1.7806 | 1.9372 | 2.0936 |
| | | CKM | 1.0542 | 1.1745 | 1.3436 | 1.3157 |
| | | MoM | *0.1647* | *0.1861* | *0.2599* | 0.3244 |
| | | Mod. MoM | 0.4321 | 0.3855 | 0.3357 | *0.3072* |
| | | BC MoM | 0.2097 | 0.2370 | 0.3507 | 0.4480 |
| | RMSE | MLE | 1.0161 | 1.3788 | 1.6952 | 1.9066 |
| | | $k$-means | 1.7300 | 1.8785 | 2.1130 | 2.7665 |
| | | CKM | 1.3190 | 1.5349 | 1.8064 | 2.1513 |
| | | MoM | **0.3310** | **0.3969** | **0.6173** | *1.1834* |
| | | Mod. MoM | 0.7306 | 0.6629 | *0.6293* | **0.7685** |
| | | BC MoM | *0.4448* | *0.5456* | 0.8846 | 1.7946 |

Table 5.4: Bias and RMSE values for estimating $\theta$ and $\delta$ based on 1000 simulations and 1000 bootstrap simulations where the true $(\theta, \delta)$ are $(0, 0)$ and the data come from the $F$ indicated in the table with $\sigma_F = 1$. The design alternative uses $S = 2$, $\theta = 0.7$, $K = 1.5$, $F =$ Normal, $\alpha = 0.01$, $\beta = 0.1$, and $\rho = 2$ resulting in an arm size of 17.

## 5.2 Multiplicative Treatment Effect Mixture Alternative

In this section, we examine estimators in the mixture alternative setting where one component of the mixture has a multiplicative treatment effect of the control distribution. We will use a limited set of estimators influenced by the results of the estimators in the location-shift mixture setting. As in the location-shift setting, estimation is treated as being calculated only upon termination of the group sequential clinical trial and all observations up to and including the termination stage are used.

### 5.2.1 MoM

The same MoM estimators defined in Section 5.1.4 cannot be appropriately applied to the multiplicative treatment effect data since it was derived in the location-shift setting. However, we can follow similar steps to solve for method of moments estimators to be used in the multiplicative treatment effect setting. To do this, we find the first and second moments of the treatment distribution. The mean of the treatment is

$$\mathrm{E}[Y] = (1-\theta)\mathrm{E}[X] + \theta\delta\mathrm{E}[X] = \mathrm{E}[X]\left((1-\theta) + \theta\delta\right) = \mathrm{E}[X]\left(1 + \theta(\delta - 1)\right) \tag{5.20}$$

and, if $I \sim \text{Bernoulli}(\theta)$, the variance is

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{Var}\left((1-I)X + I\delta X\right) \\
&= \mathrm{Var}\left(\mathrm{E}[(1-I)X + I\delta X|I]\right) + \mathrm{E}\left[\mathrm{Var}((1-I)X + I\delta X|I)\right] \\
&= (\mathrm{E}[X])^2\theta(1-\theta)(\delta-1)^2 + \mathrm{Var}(X)\left((1-\theta) + \theta\delta^2\right)
\end{aligned} \tag{5.21}$$

The MoM estimator is found by setting the sample moments equal to the theoretical moments and solving for the parameters of interest. We solve the following sets of equations for $\theta$ and $\delta$

$$\bar{X} = \mathrm{E}[X] \qquad \bar{Y} = \mathrm{E}[X]\,(1 + \theta(\delta - 1)) \tag{5.22}$$

$$S_X^2 = \mathrm{Var}(X) \qquad S_Y^2 = (\mathrm{E}[X])^2\theta(1-\theta)(\delta-1)^2 + \mathrm{Var}(X)\left((1-\theta) + \theta\delta^2\right) \tag{5.23}$$

where $(\bar{X}, S_X^2)$ are the sample mean and variance, respectively, of the control group observations and $(\bar{Y}, S_Y^2)$ are the sample mean and variance, respectively, of the treatment group observations. To aid in the solution, define

$$\Delta = \frac{\mathrm{E}[Y] - \mathrm{E}[X]}{\mathrm{E}[X]} = \theta(\delta - 1) \tag{5.24}$$

With plug-in estimator of Equation (5.24),

$$\hat{\Delta} = \frac{\bar{Y} - \bar{X}}{\bar{X}} \tag{5.25}$$

Then, method of moments estimators of $(\theta, \delta)$ can be found to be

$$\hat{\theta} = \left[1 + \frac{S_Y^2 - S_X^2(\hat{\Delta}+1)^2}{\hat{\Delta}^2(\bar{X}^2 + S_X^2)}\right]^{-1} \tag{5.26}$$

$$\hat{\delta} = \frac{1}{\hat{\theta}}\hat{\Delta} + 1 \tag{5.27}$$

As in the location-shift mixture setting, there are quantities within these equations that should always be positive based on the theoretical values. However, due to the randomness of the sample statistics used in place of theoretical values, they may violate the parameter space. Thus, we provide support of restricting certain sample quantities to maintain the quality of the estimators. First, consider Equation (5.26). If the fraction is negative, then the value of $\hat{\theta}$ will be outside of the parameter space in which $\theta$ resides. The fraction could only be negative if $S_Y^2 - S_X^2(\hat{\Delta}+1)^2$ is negative. Therefore, we subject this quantity to the $t_+$ function which is equal to $t$ if $t$ is positive and zero otherwise. Further justification that this quantity must be positive using the true theoretical values in place of the plug-in estimates is shown in Appendix A.

The other quantity of concern is $\hat{\Delta}$. It is easily seen that $\Delta$ must be positive. We also note that if $\hat{\Delta}$ is negative, then Equation (5.27) will be less than one which violates the parameter space of $\delta$. The only time $\hat{\delta}$ is negative is if $\overline{Y} - \overline{X} < 0$. If this occurs, it is evidence of the null scenario. Thus, keep the estimators within the parameter space, we will set the estimates of $\theta$ to zero and the estimate of $\delta$ to one if $\hat{\Delta} < 0$ and calculate them otherwise.

With these considerations, we define the MoM estimators for the multiplicative treatment effect in the mixture setting as

$$\hat{\theta}_{MoM} = \begin{cases} \left[1 + \frac{(S_Y^2 - S_X^2(\hat{\Delta}+1)^2)_+}{\hat{\Delta}^2(\bar{X}^2 + S_X^2) + \varepsilon_{s'}}\right]^{-1} & \text{if } \hat{\Delta} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.28}$$

$$\hat{\delta}_{MoM} = \begin{cases} \frac{1}{\hat{\theta}}\hat{\Delta} + 1 & \text{if } \hat{\Delta} > 0 \\ 1 & \text{otherwise} \end{cases} \tag{5.29}$$

where $t_+ = t$ if $t > 0$ and 0 otherwise, $\varepsilon_{s'}$ is defined in Equation (5.8).

### 5.2.2 CKM

We implement the CKM algorithm discussed in Section 5.1.3. The CKM algorithm can be implemented in the same way as defined previously. In fact, we even use the same estimator for $\theta$. However, with a multiplicative treatment effect it is more appropriate to use the ratio of the means of the clusters to estimate $\delta$ instead of the difference of their means. Like the location-shift setting, if the mean of the cluster with the control group data is greater than the mean of the cluster without the control group data, it suggests that there is no treatment effect. Therefore, if this occurs we set the estimates of $(\theta, \delta)$ to $(0, 1)$.

Let Cluster 1 be the cluster with the control group observations with mean $\overline{C}_1$ and Cluster 2 be the cluster without the control group observations with mean $\overline{C}_2$. Then

the estimators for $(\theta, \delta)$ are

$$\hat{\theta}_{CKM} = \begin{cases} \frac{\text{\# of observations in Cluster 2}}{s'm} & \text{if } \overline{C}_2 - \overline{C}_1 > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.30}$$

$$\hat{\delta}_{CKM} = \begin{cases} \frac{\overline{C}_2}{\overline{C}_1} & \text{if } \overline{C}_2 - \overline{C}_1 > 0 \\ 1 & \text{otherwise} \end{cases} \tag{5.31}$$

### 5.2.3 CKM-log

Due to the skewed nature of the data, the CKM estimators may have poor performance. We can imagine a single responder to the treatment having an observation that is much larger than all other observations. When using euclidean distance from the mean of a cluster to determine the cluster assignment of an observation, this point's distance could skew the results. Then the CKM algorithm may put this in a cluster of its own and group all other observations in a single cluster, giving an estimate of $\theta$ close to zero. This result could be counterproductive to the estimation if there are other responders to the treatment. Therefore, we are interested in transforming the data in such a way that the CKM algorithm using euclidean distance from a cluster mean is an appropriate distance measure. We attempt to do this by taking the log of the entire sample as the data on which we run the CKM algorithm.

Using the log of all of the observations as the input, then the CKM algorithm is implemented as previously described. Once the clusters are found, we check if there is evidence of no treatment effect. Since taking the log of the observations is a monotone transformation, it is again appropriate to check if the mean of the cluster with the control

group observations is greater than the mean of the cluster without the control group observations. If it is, then we set the estimates of $(\theta, \delta)$ to $(0, 1)$. Otherwise, we calculate the estimate of $\theta$ in the same way as before. For the estimate of $\delta$, we will now have cluster means on the log scale. Therefore, we use the exponential function on their difference to get an estimate of the ratio of the means of the raw data.

Let Cluster 1 be the cluster with the control group observations with mean $\overline{C}_1$ and Cluster 2 be the cluster without the control group observations with mean $\overline{C}_2$. Then the CKM-log (CKM-l) estimators for $(\theta, \delta)$ are

$$\hat{\theta}_{CKM-l} = \begin{cases} \frac{\text{\# of observations in cluster 2}}{s'm} & \text{if } \overline{C}_2 - \overline{C}_1 > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.32}$$

$$\hat{\delta}_{CKM-l} = \begin{cases} \exp\left(\overline{C}_2 - \overline{C}_1\right) & \text{if } \overline{C}_2 - \overline{C}_1 > 0 \\ 1 & \text{otherwise} \end{cases} \tag{5.33}$$

### 5.2.4 Results Comparisons

We examine the results of simulating estimates of $(\theta, \delta)$ with the MoM, CKM, and CKM-l estimators in the multiplicative treatment effect mixture setting. They are evaluated on their bias and RMSE. We simulate a group sequential clinical trial until termination, then calculate the estimates using the observations up to and including the termination stage. The results are presented such that a specific group sequential design alternative was chosen and the true parameter values may have been different than the values used in the design alternative.

We begin by investigating the estimation of $\theta$. The bias and RMSE of the $\theta$ estimators can be found in Table 5.5. It can immediately be seen that for a true $\theta$ of 0.7 or 0.9, the MoM estimator performs best in all but two instances where it is a close second. The CKM-l estimator appears to do best when the true $\theta$ is 0.5. This parallels the results from the location-shift mixture setting. A $k$-means based estimator may be preferred if there is reason to believe the true $\theta$ is close to 0.5, otherwise the MoM estimator is preferred.

There are several patterns that can be seen in Table 5.5. First, as $\theta$ increases, the bias and RMSE of the CKM and CKM-l estimators increases. As $\delta$ increases, the bias of the CKM and CKM-l estimators and RMSE of all three estimators decreases. Finally, as the shape parameter increases, the bias and RMSE of all three estimators generally decrease.

For the estimation of $\delta$, the MoM estimator is far and away the best choice over the CKM and CKM-l estimators as presented in Table 5.6. Again, there are patterns apparent in the estimation. As $\theta$ increases the bias of the MoM estimator increases, while the bias and RMSE of the CKM and CKM-l estimators generally decrease. As $\delta$ increases, the bias and RMSE of all three estimators generally increase. As the shape parameter increases, the bias and RMSE of all three estimators decrease.

Overall, the MoM estimator is the preferred estimator for $(\theta, \delta)$. If the experimenter has reason to believe the true $\theta$ is around 0.5, they may wish to use the CKM-l estimator for $\theta$. In a comparison between the CKM and CKM-l estimators of $(\theta, \delta)$, the CKM-l estimator generally offers a slight improvement over the CKM estimators.

| | | | Bias | | | | RMSE | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | $\delta$ | Estimator | Shape 1 | Shape 2 | Shape 4 | Shape 5 | Shape 1 | Shape 2 | Shape 4 | Shape 5 |
| 0.5 | 1.5 | MoM | **-0.0103** | **0.0320** | **0.0859** | **0.0994** | 0.4191 | 0.3894 | 0.3589 | 0.3552 |
| | | CKM | -0.3672 | -0.3322 | -0.2853 | -0.2766 | *0.3819* | *0.3547* | *0.3174* | *0.3131* |
| | | CKM-l | *-0.1743* | *-0.1524* | *-0.1259* | *-0.1179* | **0.2817** | **0.2602** | **0.2213** | **0.2171** |
| | 2 | MoM | **0.0555** | *0.1018* | *0.0899* | *0.0913* | 0.3824 | *0.3317* | *0.2845* | *0.2631* |
| | | CKM | -0.3648 | -0.3287 | -0.2713 | -0.2450 | *0.3802* | 0.3553 | 0.3083 | 0.2857 |
| | | CKM-l | *-0.1211* | **-0.0971** | **-0.0738** | **-0.0555** | **0.2455** | **0.2137** | **0.1848** | **0.1661** |
| | 2.5 | MoM | **0.0900** | **0.0716** | *0.0773* | *0.0692* | *0.3384* | *0.2910* | *0.2320* | *0.2106* |
| | | CKM | -0.3726 | -0.3265 | -0.2513 | -0.2237 | 0.3881 | 0.3509 | 0.2885 | 0.2657 |
| | | CKM-l | *-0.1009* | *-0.0866* | **-0.0403** | **-0.0334** | **0.2325** | **0.2050** | **0.1436** | **0.1371** |
| 0.7 | 1.5 | MoM | **-0.1498** | **-0.0702** | **-0.0133** | **0.0377** | *0.4278* | *0.3773* | **0.3082** | **0.2770** |
| | | CKM | -0.5581 | -0.5168 | -0.4534 | -0.4194 | 0.5689 | 0.5340 | 0.4782 | 0.4465 |
| | | CKM-l | *-0.3461* | *-0.3136* | *-0.2746* | *-0.2468* | **0.4145** | **0.3747** | *0.3325* | *0.3014* |
| | 2 | MoM | **-0.0251** | **0.0185** | **0.0404** | **0.0511** | **0.3369** | **0.2731** | **0.2226** | **0.2131** |
| | | CKM | -0.5491 | -0.4898 | -0.4010 | -0.3723 | 0.5624 | 0.5099 | 0.4318 | 0.4051 |
| | | CKM-l | *-0.2750* | *-0.2356* | *-0.1849* | *-0.1732* | *0.3503* | *0.3026* | *0.2477* | *0.2371* |
| | 2.5 | MoM | **0.0122** | **0.0483** | **0.0315** | **0.0368** | **0.2876** | **0.2361** | **0.1882** | **0.1769** |
| | | CKM | -0.5394 | -0.4633 | -0.3706 | -0.3292 | 0.5531 | 0.4849 | 0.4054 | 0.3668 |
| | | CKM-l | *-0.2382* | *-0.1997* | *-0.1388* | *-0.1258* | *0.3178* | *0.2725* | *0.2001* | *0.1921* |
| 0.9 | 1.5 | MoM | **-0.2659** | **-0.1786** | **-0.0923** | **-0.0761** | **0.4676** | **0.3690** | **0.2599** | **0.2288** |
| | | CKM | -0.7435 | -0.6832 | -0.5993 | -0.5746 | 0.7528 | 0.6985 | 0.6199 | 0.5956 |
| | | CKM-l | *-0.5081* | *-0.4735* | *-0.4126* | *-0.3965* | *0.5554* | *0.5163* | *0.4495* | *0.4318* |
| | 2 | MoM | **-0.1311** | **-0.0737** | **-0.0391** | **-0.0517** | **0.3189** | **0.2317** | **0.1655** | **0.1697** |
| | | CKM | -0.7229 | -0.6355 | -0.5369 | -0.5102 | 0.7344 | 0.6550 | 0.5648 | 0.5421 |
| | | CKM-l | *-0.4395* | *-0.3884* | *-0.3035* | *-0.2896* | *0.4893* | *0.4366* | *0.3388* | *0.3334* |
| | 2.5 | MoM | **-0.0960** | **-0.0613** | **-0.0303** | **-0.0278** | **0.2571** | **0.1917** | **0.1495** | **0.1421** |
| | | CKM | -0.7128 | -0.6218 | -0.4953 | -0.4512 | 0.7242 | 0.6415 | 0.5296 | 0.4883 |
| | | CKM-l | *-0.3940* | *-0.3278* | *-0.2323* | *-0.2051* | *0.4471* | *0.3738* | *0.2728* | *0.2478* |

Table 5.5: Bias and RMSE values for estimating $\theta$ where the data come from the distribution, $\theta$, and $\delta$ values indicated in the table for a design scenario with $S = 2$, $\theta = 0.7$, $\delta = 2$, $F =$ Gamma, $\alpha = 0.05$, $\beta = 0.2$, and $\rho = 2$ resulting in an arm size of 19.

| | | | Bias | | | | RMSE | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | $\delta$ | Estimator | Shape 1 | Shape 2 | Shape 4 | Shape 5 | Shape 1 | Shape 2 | Shape 4 | Shape 5 |
| 0.5 | 1.5 | MoM | **0.0341** | **0.0526** | **-0.0176** | **-0.0268** | **0.7012** | **0.5841** | **0.3693** | **0.3430** |
| | | CKM | 3.3242 | 2.3552 | 1.7341 | 1.6244 | 3.8523 | 2.6653 | 1.8647 | 1.7449 |
| | | CKM-l | *3.1784* | *2.0881* | *1.6010* | *1.4779* | *3.7097* | *2.3339* | *1.6821* | *1.5464* |
| | 2 | MoM | **0.0438** | **0.0436** | **-0.0086** | **0.0022** | **0.9749** | **0.6733** | **0.4991** | **0.4169** |
| | | CKM | *3.8473* | 2.8584 | 2.2113 | 2.1135 | 4.5143 | 3.2176 | 2.3868 | 2.2270 |
| | | CKM-l | 3.9232 | *2.6038* | *1.9835* | *1.9388* | *4.4645* | *2.7978* | *2.0703* | *2.0045* |
| | 2.5 | MoM | **0.1756** | **0.0995** | **0.0237** | **-0.0165** | **1.3124** | **0.9223** | **0.6073** | **0.5018** |
| | | CKM | 4.8281 | 3.4797 | 2.6730 | 2.5220 | 5.7311 | 3.8877 | 2.8610 | 2.6451 |
| | | CKM-l | *4.5790* | *3.0978* | *2.5085* | *2.3628* | *5.1340* | *3.3345* | *2.5755* | *2.4159* |
| 0.7 | 1.5 | MoM | **0.2218** | **0.1044** | **0.0953** | **0.0650** | **0.8151** | **0.5287** | **0.3639** | **0.3000** |
| | | CKM | 3.3439 | 2.1798 | 1.6277 | 1.4628 | 3.9266 | 2.4831 | 1.7560 | 1.5495 |
| | | CKM-l | *3.2517* | *1.9759* | *1.4470* | *1.3498* | *3.8229* | *2.1665* | *1.5322* | *1.4040* |
| | 2 | MoM | **0.2689** | **0.1885** | **0.1070** | **0.0547** | **1.0592** | **0.6730** | **0.4816** | **0.4006** |
| | | CKM | *3.9302* | 2.7999 | 2.0845 | 1.9333 | *4.6418* | 3.1191 | 2.2153 | 2.0272 |
| | | CKM-l | 4.1111 | *2.5602* | *1.9266* | *1.7781* | 4.7068 | *2.7383* | *2.0043* | *1.8381* |
| | 2.5 | MoM | **0.3671** | **0.1790** | **0.1127** | **0.0499** | **1.2155** | **0.8371** | **0.5570** | **0.5169** |
| | | CKM | *4.6514* | 3.3164 | 2.5601 | 2.3612 | *5.3758* | 3.6130 | 2.6814 | 2.4436 |
| | | CKM-l | 4.9679 | *3.1214* | *2.4120* | *2.2376* | 5.6141 | *3.3187* | *2.4763* | *2.2910* |
| 0.9 | 1.5 | MoM | **0.3189** | **0.2429** | **0.1612** | **0.1435** | **0.8349** | **0.5531** | **0.3639** | **0.3196** |
| | | CKM | *3.1906* | 2.0435 | 1.4304 | 1.3232 | *3.7730* | 2.3137 | 1.5398 | 1.4021 |
| | | CKM-l | 3.2679 | *1.8966* | *1.3371* | *1.2012* | 3.7879 | *2.0854* | *1.4103* | *1.2595* |
| | 2 | MoM | **0.4110** | **0.3000** | **0.1458** | **0.1353** | **0.9975** | **0.6720** | **0.4596** | **0.4120** |
| | | CKM | *3.7579* | 2.6826 | 1.8843 | 1.7210 | *4.3301* | 2.9083 | 1.9824 | 1.7961 |
| | | CKM-l | 4.2068 | *2.4766* | *1.7724* | *1.6072* | 4.7409 | *2.6579* | *1.8302* | *1.6579* |
| | 2.5 | MoM | **0.5949** | **0.3052** | **0.1476** | **0.0945** | **1.3008** | **0.8334** | **0.5457** | **0.4688** |
| | | CKM | *4.6359* | 3.1636 | 2.3042 | 2.0765 | *5.2000* | 3.3798 | 2.3899 | 2.1434 |
| | | CKM-l | 4.9921 | *3.0445* | *2.2137* | *2.0294* | 5.5437 | *3.2271* | *2.2715* | *2.0715* |

Table 5.6: Bias and RMSE values for estimating $\delta$ where the data come from the distribution, $\theta$, and $\delta$ values indicated in the table for a design scenario with $S = 2$, $\theta = 0.7$, $\delta = 2$, $F =$ Gamma, $\alpha = 0.05$, $\beta = 0.2$, and $\rho = 2$ resulting in an arm size of 19.

# Chapter 6

# Conclusions

In this work, we argue that the traditional assumption of the pure shift model to represent the treatment distribution may not always be correct. A mixture model is a more appropriate representation of the treatment group as there may be individuals who are unaffected by the treatment. If nonresponders exist but the experimenter erroneously works under the pure shift assumption, a design will be underpowered. Both a location-shift and multiplicative treatment effect mixture are considered.

We present a novel method of using the Wilcoxon Rank Sum statistic in group sequential clinical trials with the SAR test procedure. It's usefulness over existing methods if the experimenter has some idea about $F$ is shown in the reduction of the arm size needed to detect a given alternative. A similar conclusion is reached in the comparison with fixed sample methods by the average sample numbers. The SAR is compared to the SR, an alternative application of the Wilcoxon Rank Sum statistic in the group sequential setting, and their asymptotic equivalence is shown. Both of these test statistics maintain the Type

I error rate irrespective of $F$. The SAR is presented as the preferred method due to its slightly simpler computation.

Several estimators are evaluated for the treatment effect $(\theta, \delta)$ in both the location-shift and multiplicative treatment effect setting. Based on their bias and RMSE, the method of moments estimators for $(\theta, \delta)$ are preferred overall. Prior knowledge that roughly half of the individuals may respond to the treatment could warrant the use of a $k$-means-based estimator for $\theta$.

Future work could investigate several paths. First, discrete data could be considered. This would result in ties in the ranks. Secondly, two-sided tests could be considered where we may reject because the treatment is effective over the control or could be actively worse than the control. A related idea would be to explore the use of a third component in the mixture distribution. The first two components would be those we have explored in this work and the third component could represent individuals that experience negative effects from the treatment. One could explore a discrete mixture with an arbitrary number of components or a continuous mixture. It may be worthwhile to investigate the combination of the location-shift mixture and the multiplicative treatment effect mixture to represent the treatment distribution. Further exploration into the amount of information lost by using the SAR instead of the SR could be a useful avenue.

In regards to estimation, properties of the estimators could be explored. Then confidence intervals could be established for the treatment effect. Other estimators, such as one based on hierarchical clustering instead of $k$-means, may also be investigated.

# Bibliography


[1] Boos, D., and Brownie, C. Testing for a treatment effect in the presence of nonresponders. *Biometrics 42 1* (1986), 191–7.

[2] Casella, G., and Berger, R. L. *Statistical Inference*, 2 ed. Duxbury, 2002.

[3] Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. *mvtnorm: Multivariate Normal and t Distributions*, 2020. R package version 1.1-1.

[4] Gibbons, J. D., and Chakraborti, S. *Nonparametric Statistical Inference*, 4 ed. Marcel Dekker, Inc., New York, NY, 2003.

[5] Good, P. I. Detection of a treatment effect when not all experimental subjects will respond to treatment. *Biometrics 35*, 2 (1979), 483–489.

[6] Hewett, J. E., and Spurrier, J. D. A survey of two stage tests of hypotheses: theory and application. *Communications in Statistics - Theory and Methods 12*, 20 (1983), 2307–2425.

[7] Huang, P., and Tan, M. Multistage nonparametric tests for treatment comparisons in clinical trials with multiple primary endpoints. *Statistics and its Interface 9*, 3 (2016), 343–354.

[8] Jennison, C., and Turnbull, B. W. *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC, 2000.

[9] Jeske, D. R., and Yao, W. Sample size calculations for mixture alternatives in a control group vs. treatment group design. *Statistics 54* (01 2020), 1–17.

[10] Lee, J. W., and DeMets, D. L. Sequential rank tests with repeated measurements in clinical trials. *Journal of the American Statistical Association 87*, 417 (1992), 136–142.

[11] Lindsay, B. G. *Mixture Models: Theory, Geometry, and Applications*. Institute of Mathematical Sciences, Hayward, CA, 1995.

[12] Lubich, B., Jeske, D. R., and Yao, W. Method of moments confidence intervals for a semi-supervised two-component mixture model, Epub ahead of print 2021.

[13] MacQueen, J. B. Some methods for classification and analysis of multivariate observations. In *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability* (1967), L. M. L. Cam and J. Neyman, Eds., vol. 1, University of California Press, pp. 281–297.

[14] Madsen, R. W., and Hewett, J. E. Multi-stage tests based on sequential ranks. *Journal of Statistical Computation and Simulation 7*, 2 (1978), 93–105.

[15] Mardia, K. V., Kent, J. T., and Bibby, J. M. *Multivariate analysis*. Academic Press, 1979.

[16] Peng, H., Jeske, D. R., Sengupta, A., and Yao, W. Designing one-sided group sequential clinical trials to detect a mixture alternative. *Sequential Analysis 37* (04 2018), 268–291.

[17] Rohatgi, V. K. *An introduction to probability and statistics*, 2nd ed. / vijay k. rohatgi, a.k. md. ehsanes saleh. ed. Wiley series in probability and statistics. Texts and references section. Wiley, New York, 2001.

[18] Sen, P. K. Weak convergence of generalized $u$-statistics. *The Annals of Probability 2*, 1 (1974), 90–102.

[19] Shuster, J., Chang, M., and Tian, L. Design of group sequential clinical trials with ordinal categorical data based on the mann–whitney–wilcoxon test. *SEQUENTIAL ANALYSIS 23* (12 2004), 413–426.

[20] Simpson, J., Olsen, A., and Eden, J. C. A bayesian analysis of a multiplicative treatment effect in weather modification. *Technometrics 17*, 2 (1975), 161–166.

[21] Spurrier, J. D., and Hewett, J. E. Two-stage wilcoxon tests of hypotheses. *Journal of the American Statistical Association 71*, 356 (1976), 982–987.

[22] Su, J. Q., and Lachin, J. M. Group sequential distribution-free methods for the analysis of multivariate observations. *Biometrics 48*, 4 (1992), 1033–1042.

[23] Wagstaff, K., Cardie, C., Rogers, S., and Schrödl, S. Constrained k-means clustering with background knowledge. In *Proceedings of 18th International Conference on Machine Learning* (01 2001), pp. 577–584.

[24] Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin 1*, 6 (1945), 80–83.

[25] Wilcoxon, F., Rhodes, L. J., and Bradley, R. A. Two sequential two-sample grouped rank tests with applications to screening experiments. *Biometrics 19*, 1 (1963), 58–84.

[26] Yuan, A., Zheng, Y., Huang, P., and Tan, M. T. A nonparametric test for the evaluation of group sequential clinical trials with covariate information. *Journal of Multivariate Analysis 152* (2016), 82–99.

# Appendix A

# Multiplicative Treatment Effect MoM Restriction

In the MoM estimators used in the multiplicative treatment effect mixture alternative setting, we utilize a function $t_+$ that is equal to $t$ if $t$ is positive and zero otherwise. This function keeps the estimates within the parameter space. We provide justification that the value of the function should be positive in the context of the MoM estimators by using the true theoretical values of the sample estimates used in Equations (5.28) and (5.29). Let $(\mu_X, \sigma_X^2)$ be the mean and variance of the control group, respectively, and let $(\mu_Y, \sigma_Y^2)$ be the mean and variance of the treatment group, respectively. Then $\Delta = (\mu_Y - \mu_X)/\mu_X$ and

we solve the following equation

$$
\begin{aligned}
\sigma_Y^2 - \sigma_X^2(\Delta+1)^2 &\overset{?}{>} 0 \\
\sigma_Y^2 &\overset{?}{>} \sigma_X^2(\Delta+1)^2 \\
\frac{\sigma_Y^2}{\sigma_X^2} &\overset{?}{>} \left(\frac{\mu_Y-\mu_X}{\mu_X}+1\right)^2 \\
\frac{\sigma_Y^2}{\sigma_X^2} &\overset{?}{>} \left(\frac{\mu_Y}{\mu_X}\right)^2 \\
\frac{(\mu_X)^2\theta(1-\theta)(\delta-1)^2+\sigma_X^2\left((1-\theta)+\theta\delta^2\right)}{\sigma_X^2} &\overset{?}{>} \left(\frac{\mu_X\left(1+\theta(\delta-1)\right)}{\mu_X}\right)^2 \\
\frac{\mu_X^2}{\sigma_X^2}\theta(1-\theta)(\delta-1)^2+(1+\theta(\delta^2-1)) &\overset{?}{>} (1+\theta(\delta-1))^2 \\
\frac{\mu_X^2}{\sigma_X^2}\theta(1-\theta)(\delta-1)^2+(1+\theta(\delta-1)(\delta+1)) &\overset{?}{>} 1+2\theta(\delta-1)+\theta^2(\delta-1)^2 \\
\frac{\mu_X^2}{\sigma_X^2}\theta(1-\theta)(\delta-1)^2+\theta(\delta-1)(\delta+1) &\overset{?}{>} 2\theta(\delta-1)+\theta^2(\delta-1)^2 \\
\frac{\mu_X^2}{\sigma_X^2}(1-\theta)(\delta-1)+(\delta+1) &\overset{?}{>} 2+\theta(\delta-1) \\
\frac{\mu_X^2}{\sigma_X^2}(1-\theta)(\delta-1) &\overset{?}{>} \theta(\delta-1)-\delta+1 \\
\frac{\mu_X^2}{\sigma_X^2}(1-\theta) &\overset{?}{>} \theta-\frac{(\delta-1)}{\delta-1} \\
\frac{\mu_X^2}{\sigma_X^2}(1-\theta) &\overset{?}{>} -(1-\theta) \\
\frac{\mu_X^2}{\sigma_X^2} &> -1
\end{aligned}
$$

which is always true. Therefore, we use $(S_Y^2 - S_X^2(\hat{\Delta}+1)^2)_+$ in our MoM estimators in the multiplicative treatment effect mixture setting.

# Appendix B

# R Function Documentation and Code

## B.1 Code Manual

`grp.seq.mix` *Calculate critical values and arm size for a mixture alternative*

**Description**

`grp.seq.mix()` is used for the design of one-sided group sequential clinical trials with mixture alternative using the Sequential Average Rank (SAR) or Sequential Rerank (SR) test statistic. Error spending functions are implemented such that the test may reject early for efficacy or accept early for futility. The function handles fixed sample (single stage) design as well as the pure shift (`theta`=1). Requires packages `rmutil` if `f = "laplace"` and `mvtnorm` if `norm.approx = TRUE`.

**Usage**

```
grp.seq.mix(num.stages, alpha, beta, rho, theta, K, delta,
        f = c("normal", "logistic", "laplace", "t3", "gamma"),
        mu = 0, sigma = 1, shape = 1, method = c("sar","sr"),
        B = 1e5, norm.approx = TRUE)
```

**Arguments**

| | |
|---|---|
| `num.stages` | Maximum number of planned stages. |
| `alpha` | Overall Type I error for the design. |
| `beta` | Overall Type II error for the design. |
| `rho` | Exponent in error spending functions. |
| `theta` | Specified proportion of nonresponders in the treatment distribution. |
| `K` | Size of the location-shift in terms of standard deviations. |
| `delta` | Multiplicative treatment effect. |
| `f` | Distribution of the data. |
| `mu` | Mean of `f` if it is from the location-scale family. |
| `sigma` | Scale parameter. Standard deviation of `f` if it is from the location-scale family. |
| `shape` | Shape parameter of the distribution of the data if it is from the scale family. |
| `method` | Choice of test statistic. |
| `B` | Number of simulations when approximating the exact distribution by simulation. |
| `norm.approx` | If `TRUE`, the normal approximation of the distribution of the test statistic is used instead of simulation. |

**Value**

| | |
|---|---|
| `Design Alternative` | Prints the number of stages, `f`, and `theta` values inputted for the design alternative. Also prints `K` if `f` is in the location-scale family and `delta` if `f = "gamma"`. |
| `Arm Size` | Arm size per stage per group needed to detect the alternative. |
| `Upper Critical Values` | Vector of the upper critical values. In the case of a single stage trial, this is replaced with `Critical Value` showing the only critical value used in that case. |
| `Lower Critical Values` | Vector of the lower critical values. Omitted if `num.stages = 1`. |
| `Type I Error` | Vector of the realized Type I error per stage and overall. |
| `Type II Error` | Vector of the realized Type II error per stage and overall. |

**Example**

```
grp.seq.mix(num.stages=3, alpha=0.05, beta=0.2, rho=2, theta=0.8,
K=0.5, sigma=1, f="logistic", B=10000, method="sar", norm.approx=TRUE)
```

**Example Output**

```
$`Design Alternative`
 Stages         f  theta    K
      3  logistic    0.8  0.5
$`Arm Size`
[1] 27
$`Upper Critical Values`
[1] 2.5392 2.0680 1.6965
$`Lower Critical Values`
[1] -0.4587 0.7480 1.6965
$`Type I Error`
 Stage 1  Stage 2  Stage 3  Overall
  0.0056   0.0167   0.0278   0.0500
$`Type II Error`
 Stage 1  Stage 2  Stage 3  Overall
  0.0222   0.0667   0.1042   0.1931
```

## B.2 R Code

```
# packages: rmutil, mvtnorm
grp.seq.mix <- function(num.stages=2,alpha,beta,rho,theta,delta,K=1,mu=0,
                        sigma=1,shape=2,f="normal",B=1e5,method="sar",
                        norm.approx=FALSE){
  m.start = 1
  k <- num.stages
  method <- ifelse(method=="sr","rerank",method)
  f <- ifelse(f=="t3","t",f)
  std=TRUE
  pi1 <- c(alpha*(1/k)^rho,diff(alpha*((1:k)/k)^rho))
  pi2 <- c(beta*(1/k)^rho,diff(beta*((1:k)/k)^rho))
  d <- f
  # if(d=="gamma" & mu==0){mu <- arg1 <- 1; arg2 <- shape
  # } else if(d=="gamma" & mu>0){arg1 <- mu; arg2 <- shape
  # } else {arg1 <- mu; arg2 <- sigma}
  if(d=="gamma"){arg1 <- mu <- sigma; arg2 <- shape
  } else {arg1 <- mu; arg2 <- sigma}
  if(f=="normal"){
    f <- function(n,arg1=0,arg2=1){rnorm(n=n,mean=arg1,sd=arg2)}
    ppsi <- function(x,arg1=0,arg2=1){pnorm(x,mean=arg1,sd=arg2)}
```

```
  dpsi <- function(x,arg1=0,arg2=1,lg=FALSE){
                  dnorm(x,mean=arg1,sd=arg2,log=lg)}
  qpsi <- function(q,arg1=0,arg2=1) qnorm(q)*arg2 + arg1
} else if(f=="logistic"){
  f <- function(n,arg1=0,arg2){rlogis(n=n,location=arg1,arg2*sqrt(3)/pi)}
  ppsi <- function(x,arg1=0,arg2=1){
                  plogis(x,location=arg1,scale=arg2*sqrt(3)/pi)}
  dpsi <- function(x,arg1=0,arg2=1,lg=FALSE){
                  dlogis(x,location=arg1,scale=arg2*sqrt(3)/pi,log=lg)}
  qpsi <- function(q,arg1=0,arg2=1){
                  qlogis(q,location=arg1,scale=arg2*sqrt(3)/pi) }
} else if(f=="laplace"){
  # uses package rmutil
  f <- function(n,arg1=0,arg2){
               rmutil::rlaplace(n=n,m=arg1,s=arg2/sqrt(2))}
  ppsi <- function(x,arg1=0,arg2=1){
                  rmutil::plaplace(x,m=arg1,s=arg2/sqrt(2))}
  dpsi <- function(x,arg1=0,arg2=1,lg=FALSE){
                  rmutil::dlaplace(x,m=arg1,s=arg2/sqrt(2),log=lg)}
  qpsi <- function(q,arg1=0,arg2=1){
                  rmutil::qlaplace(q,m=arg1,s=arg2/sqrt(2)) }
} else if(f=="t"){
  f <- function(n,arg1=0,arg2){arg1 + arg2*rt(n=n,df=3,ncp=0)/sqrt(3)}
  ppsi <- function(x,arg1=0,arg2=1){pt((x-arg1)*sqrt(3)/arg2,df=3,ncp=0)}
  dpsi <- function(x,arg1=0,arg2=1,lg=FALSE){
    if(lg==FALSE){
       return(dt((x-arg1)*sqrt(3)/arg2,df=3,ncp=0)*sqrt(3)/arg2)
    } else {
       dt((x-arg1)*sqrt(3)/arg2,df=3,ncp=0,log=T)+log(sqrt(3))-log(arg2) }
  }
  qpsi <- function(q,arg1=0,arg2=1) qt(q,df=3)/sqrt(3)*arg2 + arg1
} else if(f=="gamma"){
  # f <- function(n,arg1=mu,arg2=shape){
  #               rgamma(n=n,shape=arg2,scale=arg1/arg2)}
  # ppsi <- function(x,arg1=mu,arg2=shape){
  #                  pgamma(x,shape=arg2,scale=arg1/arg2)}
  # dpsi <- function(x,arg1=mu,arg2=shape,lg=FALSE){
  #                  dgamma(x,shape=arg2,scale=arg1/arg2,log=lg)}
  f <- function(n,arg1=mu,arg2=shape){
               rgamma(n=n,shape=arg2,scale=arg1)}
  ppsi <- function(x,arg1=mu,arg2=shape){
                  pgamma(x,shape=arg2,scale=arg1)}
  dpsi <- function(x,arg1=mu,arg2=shape,lg=FALSE){
                  dgamma(x,shape=arg2,scale=arg1,log=lg)}
}
# normal approx for one stage
if(d!="gamma"){ # probs for location shift
  delta <- K*sigma
  h1<-function(y,theta,K) {ppsi(y)*((1-theta)*dpsi(y) + theta*dpsi(y-K))}
  gamval<-integrate(h1,-Inf,Inf,theta,K)$value
```

```
  h2<-function(y,theta,K){
              ((1-(1-theta)*ppsi(y)-theta*ppsi(y-K))^2)*dpsi(y)}
  xi1<-integrate(h2,-Inf,Inf,theta,K)$value-gamval^2
  h3<-function(y,theta,K) {
              (ppsi(y))^2*((1-theta)*dpsi(y) + theta*dpsi(y-K))}
  xi2<-integrate(h3,-Inf,Inf,theta,K)$value-gamval^2
  lambda <- 1/2
  # P(X<Y)
  pxy <- integrate(function(y){
          ppsi(y)*((1-theta)*dpsi(y)+theta*dpsi(y-K))},-Inf,Inf)$value
  # P(X1 < Y1, X1 < Y2)
  pxyy <- integrate(function(y){
          (1-(1-theta)*ppsi(y)-theta*ppsi(y-K))^2*dpsi(y)},-Inf,Inf)$value
  # P(X1 < Y1, X2 < Y2)
  pxxy <- integrate(function(y){
          (ppsi(y))^2*((1-theta)*dpsi(y)+theta*dpsi(y-K))},-Inf,Inf)$value
} else { # probs for gamma distr multiplicative shift
  h1<-function(y,theta,K) {
              ppsi(y)*((1-theta)*dpsi(y) + theta*dpsi(y/delta)/delta)}
  gamval<-integrate(h1,-Inf,Inf,theta,K)$value
  h2<-function(y,theta,K) {
              ((1-(1-theta)*ppsi(y)-theta*ppsi(y/delta))^2)*dpsi(y)}
  xi1<-integrate(h2,-Inf,Inf,theta,K)$value-gamval^2
  h3<-function(y,theta,K) {
             (ppsi(y))^2*((1-theta)*dpsi(y) + theta*dpsi(y/delta)/delta)}
  xi2<-integrate(h3,-Inf,Inf,theta,K)$value-gamval^2
  lambda <- 1/2
  # P(X<Y)
  pxy <- integrate(function(y){
          ppsi(y)*((1-theta)*dpsi(y)+theta*dpsi(y/delta)/delta)},
          -Inf,Inf)$value
  # P(X1 < Y1, X1 < Y2)
  pxyy <- integrate(function(y){
          (1-(1-theta)*ppsi(y)-theta*ppsi(y/delta))^2*dpsi(y)},
          -Inf,Inf)$value
  # P(X1 < Y1, X2 < Y2)
  pxxy <- integrate(function(y){
          (ppsi(y))^2*((1-theta)*dpsi(y)+theta*dpsi(y/delta)/delta)},
          -Inf,Inf)$value
}
# jeske yao 2020 fixed sample formula for arm size
m.approx <- ((qnorm(1-alpha)/sqrt(6) +
              qnorm(1-beta)*sqrt(xi1+xi2))/(gamval-1/2))^2

if(norm.approx==FALSE){
  if(k==1){ # simulate just one stage
    if(d!="gamma"){delta <- K*sigma}
    p <- 1
    if((m.approx-10)>m.start){m.start <- floor(m.approx - 10)}
    inc <- function(incr = 1, m = m.start){
      while(p>=pi2){
        ifelse(m<15,m<-m+1,m <- m+incr)
        w <- vector("numeric",B)
```

```
    w.alt <- vector("numeric",B)
    for(i in 1:B){
      if(d!="gamma"){
        obs <- matrix(f(n=k*2*m,arg2=arg2),ncol=2,nrow=(k*m),byrow=F)
        obs.alt <- NULL
        y <- sapply(runif(m),FUN = function(i){
              if(i<theta){
                  f(n=1,arg1=delta,arg2=arg2)}else{f(n=1,arg2=arg2)}})
        obs.alt <- rbind(obs.alt,matrix(c(f(n=m,arg2=sigma),y),
                         ncol=2,nrow=m,byrow=F))
      } else {
        obs <- matrix(f(n=k*2*m,arg1=mu,arg2=shape),
                      ncol=2,nrow=(k*m),byrow=F)
        obs.alt <- NULL
        y <- sapply(runif(m),FUN = function(i){if(i<theta){
              f(n=1,arg1=delta*mu,arg2=shape)}else{
              f(n=1,arg1=mu,arg2=shape)}})
        obs.alt <- rbind(obs.alt,
                         matrix(c(f(n=m,arg1=mu,arg2=shape),y),
                         ncol=2,nrow=m,byrow=F))
      }
      # calculate rank sum value
      w[i] <- sum(rank(obs)[(m+1):(m+m)])
      w.alt[i] <- sum(rank(obs.alt)[(m+1):(m+m)])
    }
    p.r <- 0
    index.max <- m*m + m*(m+1)/2
    u <- sort(unique(w),decreasing = TRUE)
    index <- index.max + 1
    while(p.r <= pi1){
      p.rej <- p.r
      index <- index - 1
      p.r <- sum(w>=(index))/B
      if(index==1){break}
    }
    if(index==index.max & sum(w>=index)/B>pi1[1]){
      r <- Inf
    } else if(sum(w>=index)/B <= pi1[1]) {
      r <- index
    } else {
      r <- index+1
    }
    if(std==TRUE){mu <- m*((m+m)+1)/2; sig <- sqrt(m*m*((m+m)+1)/12)
    r <- (r-mu)/sig
    w.alt <- (w.alt - mu)/sig; w <- (w-mu)/sig}
    p <- sum(w.alt<r)/B
  }
  p.rej <- sum(w>=r)/B
  p.acc <- sum(w.alt<r)/B
  # out <- list(r=r,m=m,theta=theta,K=K,B=B,
  #             distribution=d,p.rej=p.rej,p.acc=p.acc)
  if(d!="gamma"){
    out.df <- data.frame(Stages=k,f=d,theta=theta,K=K,row.names = "")
  } else { out.df <- data.frame(Stages=k,f=d,theta=theta,delta=delta,
```

```
                                        row.names = "")}
    out <- list("Design Alternative"=out.df,
                "Arm Size"=m,
                "Critical Value"=round(r,4),
                "Type I Error"=round(p.rej,4),
                "Type II Error"=round(p.acc,4))
    return(out)
  }
  incr.size <- 10
  res <- inc(incr=incr.size)
  # if(res£m>=15){
  #   res <- inc(incr = 1, m = res£m-incr.size-1)
  if(res$`Arm Size`>=15){
    res <- inc(incr = 1, m = res$`Arm Size`-incr.size-1)
    return(res)
  } else {return(res)}
} else { # simulate 2+ stages
  if(d!="gamma"){delta <- K*sigma}
  p.start <- 1
  # use normal approx for one stage to get close initially
  if((m.approx/k-10)>m.start){m.start <- floor(m.approx/k - 10)}
  incr.size <- 10
  incr.arm.size <- function(incr = 1, p = 1, m = m.start){
    while(p>=pi2[k]){
      p.previous <- p
      ifelse(m<15,m<-m+1,m <- m+incr)
      n <- m
      r <- vector("numeric",k)
      a <- vector("numeric",k)
      if(method=="rerank"){w <- matrix(0,ncol = k,nrow=B)
                           w.alt <- matrix(0,ncol = k,nrow=B)}
      else{srs <- matrix(0,ncol = k,nrow=B)
           srs.alt <- matrix(0,ncol = k,nrow=B)}
      ### faster code
      for(i in 1:B){
        if(d!="gamma"){
          obs <- matrix(f(n=k*2*m,arg2=sigma),ncol=2,nrow=(k*m),byrow=F)
          ind <- sample(1:2,prob=c(1-theta,theta),
                        size=k*m,replace = TRUE)
          mus <- c(0,delta)
          y <- f(n=k*m,arg1=mus[ind],arg2=sigma)
          obs.alt <- matrix(c(f(n=k*m,arg2=sigma),y),
                            ncol=2,nrow=(k*m),byrow=F)
        } else {
          obs <- matrix(f(n=k*2*m,arg1=mu,arg2=shape),
                        ncol=2,nrow=(k*m),byrow=F)
          ind <- sample(1:2,prob=c(1-theta,theta),
                        size=k*m,replace = TRUE)
          mus <- c(mu,delta*mu)
          y <- f(n=k*m,arg1=mus[ind],arg2=shape)
          obs.alt <- matrix(c(f(n=k*m,arg1=mu,arg2=shape),y),
                            ncol=2,nrow=(k*m),byrow=F)
        }
        for(i1 in 1:k){
```

```
      if(method=="rerank"){
        w[i,i1] <- sum(rank(obs[1:(i1*m),])[(i1*m+1):(i1*m+i1*m)])
        w.alt[i,i1] <- sum(
                    rank(obs.alt[1:(i1*m),])[(i1*m+1):(i1*m+i1*m)])
      }
      else{
        srs[i,i1] <- sum(
                    rank(obs[((i1-1)*m+1):(i1*m),])[(n+1):(n+m)])
        srs.alt[i,i1] <- sum(
                rank(obs.alt[((i1-1)*m+1):(i1*m),])[(n+1):(n+m)])
      }
    }
  }
  # ts = test statistic
  if(method=="rerank" & std==TRUE){
    mu <- (1:k)*m*((1:k)*(m+m)+1)/2
    sig <- sqrt((1:k)^2*m*m*((1:k)*(m+m)+1)/12)
    ts <- t(apply(w,1,FUN=function(x) (x - mu)/(sig)))
    ts.alt <- t(apply(w.alt,1,FUN=function(x) (x - mu)/(sig)))
  }
  else if(method=="rerank" & std==FALSE){
    ts <- w; ts.alt <- w.alt
  }
  if(method!="rerank" & std==TRUE){ # sar standardized
    mu <- m*((m+m)+1)/2
    sig <- sqrt(m*m*((m+m)+1)/12)
    ts <- t(apply(srs,1,FUN=function(x){
            (cumsum(x)/seq_along(x) -
                mu)/(sig/sqrt(seq_along(x)))}))
    ts.alt <- t(apply(srs.alt,1,FUN=function(x){
            (cumsum(x)/seq_along(x) -
                mu)/(sig/sqrt(seq_along(x)))}))
  }
  if(method!="rerank" & std==FALSE){ # sar not standardized
    ts <- t(apply(srs,1,FUN=function(x) (cumsum(x)/seq_along(x))))
    ts.alt <- t(apply(srs.alt,1,FUN=function(x){
                (cumsum(x)/seq_along(x))}))
  }

  p.r <- vector("numeric",length=k)
  p.a <- vector("numeric",length=k)
  p <- 0
  index <- 0
  u <- sort(unique(ts[,1]),decreasing = T)
  while(p <= pi1[1]){
    p.rej <- p
    index <- index + 1
    p <- sum(ts[,1]>=u[index])/B
    if(index==length(u)){break}
  }
  if(index==1 & sum(ts[,1]>=u[index])/B>pi1[1]){
    r[1] <- Inf
  } else if(sum(ts[,1]>=u[index])/B <= pi1[1]) {
    r[1] <- u[index]
```

```
} else {
  r[1] <- u[index-1]
}
p.r[1] <- sum(ts[,1]>=r[1])/B
p.acc <- 0
u.alt <- sort(unique(ts.alt[,1]))
index <- 0
while(p.acc <= pi2[1]){
  index <- index + 1
  p.acc <- sum(ts.alt[,1]<=u.alt[index])/B
}
if(index==1 & sum(ts.alt[,1]<=u.alt[index])/B>pi2[1]){
  a[1] <- -Inf
} else if(sum(ts.alt[,1]<=u.alt[index])/B <= pi2[1]) {
  a[1] <- u.alt[index]
} else {
  a[1] <- u.alt[index-1]
}
p.a[1] <- sum(ts.alt[,1]<=a[1])/B
s <- ts[ts[,1]<r[1] & ts[,1]>a[1],]
s.alt <- ts.alt[ts.alt[,1]<r[1] & ts.alt[,1]>a[1],]
for(i in 2:k){
  if(dim(s)[1]==0 || dim(s.alt)[1]==0){break}
  p.rej <- 0
  u <- sort(unique(s[,i]),decreasing = TRUE)
  index <- 0
  while(p.rej <= pi1[i]){
    index <- index + 1
    p.rej <- sum(s[,i]>=(u[index]))/B
    if(index==length(u)){break}
  }
  if(index==1 & sum(s[,i]>=u[index])/B>pi1[i]){
    r[i] <- Inf
  } else if(sum(s[,i]>=u[index])/B <= pi1[i]) {
    r[i] <- u[index]
  } else {
    r[i] <- u[index-1]
  }
  p.r[i] <- sum(s[,i]>=r[i])/B
  p.acc <- 0
  u.alt <- sort(unique(s.alt[,i]))
  index <- 0
  while(p.acc <= pi2[i]){
    index <- index + 1
    p.acc <- sum(s.alt[,i]<=u.alt[index])/B
    if(index==length(u.alt)){break}
  }
  if(index==1 & sum(s.alt[,i]<=u.alt[index])/B>pi2[i]){
    a[i] <- -Inf
  } else if(sum(s.alt[,i]<=u.alt[index])/B <= pi2[i]) {
    a[i] <- u.alt[index]
  } else {
    a[i] <- u.alt[index-1]
  }
```

```
          p.a[i] <- sum(s.alt[,i]<=a[i])/B
          if(i!=k){
            s <- s[s[,i]<r[i] & s[,i]>a[i],]
            s.alt <- s.alt[s.alt[,i]<r[i] & s.alt[,i]>a[i],]
          }
        }
        p <- sum(s.alt[,k]<r[k])/B
        p.out <- p
        p.a[k] <- p
      }
      if(dim(s)[1]==0 || dim(s.alt)[1]==0){
          return("Unable to achieve design.")
      } else{
        # return(list(r=r,a=c(a[1:(k-1)],r[k]),m=m,p.r=p.r,
        #        p.a=p.a,p.out=p.out,K=K,theta=theta))
        if(d!="gamma"){
          out.df <- data.frame(Stages=k,f=d,theta=theta,K=K,
                               row.names = "")
        } else { out.df <- data.frame(Stages=k,f=d,theta=theta,
                              delta=delta,row.names = "")}
        out <- list("Design Alternative"=out.df,
                    "Arm Size"=m,
                    "Upper Critical Values"=round(r,4),
                    "Lower Critical Values"=round(c(a[1:(k-1)],r[k]),4),
                    "Type I Error"=round(c(p.r,sum(p.r)),4),
                    "Type II Error"=round(c(c(p.a[1:(k-1)],p.out),
                                            sum(c(p.a[1:(k-1)],p.out))),4))
        names(out$`Type I Error`) <- c(paste("Stage",1:k),"Overall")
        names(out$`Type II Error`) <- c(paste("Stage",1:k),"Overall")
        return(out)
      }
    }
    res1 <- incr.arm.size(incr=incr.size,p=p.start,m=m.start)
    if(is.list(res1)==FALSE){return(res1)
    # } else if(res1£m>=15){
      } else if(res1$`Arm Size`>=15){
      # res2 <- incr.arm.size(incr=1,p=1, m = res1£m-incr.size-1)
      res2 <- incr.arm.size(incr=1,p=1, m = res1$`Arm Size`-incr.size-1)
      return(res2)
    } else{return(res1)}
  }
} else { # norm.approx == TRUE
  if(k==1){ # fixed sample
    if(d!="gamma"){delta <- K*sigma}
    p <- 1
    if((m.approx-10)>m.start){m <- floor(m.approx - 10)}else{m <- 0}
    while(p>=pi2){
      m <- m+1
      mu <- m*(m+m+1)/2
      v <- m*m*(m+m+1)/12
      r <- qnorm(1-pi1)
      p <- pnorm(r,mean=(m*(m*pxy + (m+1)/2) - mu)/sqrt(v),
                 sd=m^2*(pxy*(1-pxy)+(m-1)*(pxyy-pxy^2+pxxy-pxy^2))/v)
    }
```

```
    # return(list(r=r,m=m,p=p,K=K,theta=theta))
    if(d!="gamma"){
      out.df <- data.frame(Stages=k,f=d,theta=theta,K=K,row.names = "")
    } else { out.df <- data.frame(Stages=k,f=d,theta=theta,delta=delta,
                                  row.names = "")}
    out <- list("Design Alternative"=out.df,
                "Arm Size"=m,
                "Critical Value"=round(r,4),
                "Type I Error"=round(pnorm(r,lower.tail = FALSE),4),
                "Type II Error"=round(p,4))
    return(out)
  }else{ # multiple stages
    if(d!="gamma"){delta <- K*sigma}
    p <- 1
    if((m.approx/k-10)>m.start){m <- floor(m.approx/k - 10)}else{m <- 0}
    while(p>=pi2[k]){
      m <- m+1
      r <- vector("numeric",k)
      a <- vector("numeric",k)

      if(method=="sar" & std=="TRUE"){
        mu <- m*(m+m+1)/2
        v <- m*m*(m+m+1)/12
        null.mu <- rep(0,k)
        null.var <- v/(1:k)
        null.cov <- outer(1:k,1:k,
                          Vectorize(function(x,y){sqrt(x*y)/max(x,y)}))
        alt.mu <- (m*(m*gamval + (m+1)/2) - mu)/(sqrt(v)/sqrt(1:k))
        # first one is from jeske/yao
        # alt.v <- ((m*m/sqrt(m+m))^2*(xi1/lambda + xi2/(1-lambda)))/v
        alt.v <- m^2*(pxy*(1-pxy)+(m-1)*(pxyy-pxy^2+pxxy-pxy^2))/v
        # alt.var <- ((m*m/sqrt(m+m))^2*(xi1/lambda +
        #               xi2/(1-lambda)))/((1:k)*v)
        alt.cov <- outer(1:k,1:k,
                         Vectorize(function(x,y){sqrt(x*y)/max(x,y)}))*alt.v
      }
      if(method=="rerank" & std=="TRUE"){
        mu <- (1:k)*m*((1:k)*(m+m)+1)/2
        v <- (1:k)^2*m*m*((1:k)*(m+m)+1)/12
        null.mu <- rep(0,k)
        null.var <- rep(1,k)
        null.cov <- outer(1:k,1:k,Vectorize(function(x,y){
          (((min(x,y)*m)^2*(2*max(x,y)*m+1))/12)/(sqrt(v[x]*v[y]))}))
        alt.mu <- ((1:k)*m*((1:k)*m*gamval + ((1:k)*m+1)/2) - mu)/sqrt(v)
        # don't even use these alt.var variables
        alt.var <- (((1:k)*m)^2*(pxy*(1-pxy) + ((1:k)*m-1)*(pxyy-pxy^2) +
                       ((1:k)*m-1)*(pxxy-pxy^2)))/v
        # from jeske/yao
        # alt.var2 <- (((1:k)^2*m*m/sqrt((1:k)*m+
        #               (1:k)*m))^2*(xi1/lambda + xi2/(1-lambda)))/(v)
        alt.cov <- outer(1:k,1:k,Vectorize(function(x,y){
          # ((min(x,y)*m)^2*(gamval*(1-gamval) + (max(x,y)*m-1)*(xi1) +
          #   (max(x,y)*m-1)*(xi2)))/sqrt(v[x]*v[y])
          ((min(x,y)*m)^2*(pxy*(1-pxy) + (max(x,y)*m-1)*(pxyy-pxy^2 +
              pxxy-pxy^2)))/sqrt(v[x]*v[y])
```

```
  }))
}

# critical values
r[1] <- qnorm(1-pi1[1],mean = null.mu[1],sd=sqrt(null.cov[1,1]))
a[1] <- qnorm(pi2[1],mean=alt.mu[1],sd=sqrt(alt.cov[1,1]))
for(i in 2:k){ # uses mvtnorm package
  if(m>3){
    r[i] <- uniroot(f=function(x){as.numeric(mvtnorm::pmvnorm(
              lower=c(a[1:(i-1)],x),upper=c(r[1:(i-1)],Inf),
              mean=null.mu[1:i],
              sigma=null.cov[1:i,1:i])) - pi1[i]},
                    lower=0,upper=4)$root
  } else {
    r[i] <- optimize(f=function(x){(as.numeric(mvtnorm::pmvnorm(
              lower=c(a[1:(i-1)],x),upper=c(r[1:(i-1)],Inf),
              mean=null.mu[1:i],
              sigma=null.cov[1:i,1:i])) - pi1[i])^2},
                     lower=0,upper=4)$minimum
  }
  if(i!=k){
    if(m>3){
      a[i] <- uniroot(f=function(x){as.numeric(mvtnorm::pmvnorm(
                lower=c(a[1:(i-1)],-Inf),upper=c(r[1:(i-1)],x),
                mean=alt.mu[1:i],
                sigma=alt.cov[1:i,1:i])) - pi2[i]},
                      lower=-4,upper=4)$root
    } else {
      a[i] <- optimize(f=function(x){(as.numeric(mvtnorm::pmvnorm(
                lower=c(a[1:(i-1)],-Inf),upper=c(r[1:(i-1)],x),
                mean=alt.mu[1:i],
                sigma=alt.cov[1:i,1:i])) - pi2[i])^2},
                       lower=-4,upper=4)$minimum
    }
  }
  rm(i)
}
# uses mvtnorm package
p <- as.numeric(mvtnorm::pmvnorm(lower=c(a[1:(k-1)],-Inf),
       upper=c(r[1:(k-1)],r[k]),mean=alt.mu[1:k],
       sigma=alt.cov[1:k,1:k]))

p.r <- p.a <- vector("numeric",k)
p.r[1] <- pnorm(r[1],mean=null.mu[1],sd=sqrt(null.cov[1,1]),
                lower.tail = FALSE)
p.a[1] <- pnorm(a[1],mean=alt.mu[1],sd=sqrt(alt.cov[1,1]))
for(i in 2:k){
  if(i!=k){ # uses mvtnorm package
    p.r[i] <- mvtnorm::pmvnorm(lower=c(a[1:(i-1)],r[i]),
                upper=c(r[1:(i-1)],Inf),mean=null.mu[1:i],
                sigma=null.cov[1:i,1:i])
    p.a[i] <- mvtnorm::pmvnorm(lower=c(a[1:(i-1)],-Inf),
```

```
                        upper=c(r[1:(i-1)],a[i]),mean=alt.mu[1:i],
                        sigma=alt.cov[1:i,1:i])
          } else {
            p.r[i] <- mvtnorm::pmvnorm(lower=c(a[1:(i-1)],r[i]),
                        upper=c(r[1:(i-1)],Inf),mean=null.mu[1:i],
                        sigma=null.cov[1:i,1:i])
            p.a[i] <- mvtnorm::pmvnorm(lower=c(a[1:(i-1)],-Inf),
                        upper=c(r[1:(i-1)],r[i]),mean=alt.mu[1:i],
                        sigma=alt.cov[1:i,1:i])
          }
          rm(i)
        }
      }
      # return(list(r=r,a=c(a[1:(k-1)],r[k]),m=m,p.r=p.r,
      #          p.a=p.a,K=K,theta=theta))
      if(d!="gamma"){
        out.df <- data.frame(Stages=k,f=d,theta=theta,K=K,row.names = "")
      } else { out.df <- data.frame(Stages=k,f=d,theta=theta,delta=delta,
                                    row.names = "")}
      out <- list("Design Alternative"=out.df,
                  "Arm Size"=m,
                  "Upper Critical Values"=round(r,4),
                  "Lower Critical Values"=round(c(a[1:(k-1)],r[k]),4),
                  "Type I Error"=round(c(p.r,sum(p.r)),4),
                  "Type II Error"=round(c(p.a,sum(p.a)),4))
      names(out$`Type I Error`) <- c(paste("Stage",1:k),"Overall")
      names(out$`Type II Error`) <- c(paste("Stage",1:k),"Overall")
      return(out)
    }
  }
}
```