Control Group versus Treatment Group Designs with Mixture Distributions

Daniel R. Jeske

Professor and Vice Provost of Academic Personnel

Department of Statistics, University of California Riverside

ABSTRACT

I will discuss sample size calculations and treatment effect estimation for randomized clinical trials under a model where the responses from the treatment group follow a mixture distribution. The mixture distribution is aimed at capturing the reality that not all treated patients respond to the treatment. Both fixed sample trials and group sequential trials will be discussed.

This is a joint work that spans collaborations with my former student, Hua Peng, my current students Dylan Friel and Bradley Lubich, and my UCR Department of Statistics colleague, Weixin Yao.

Taylor & Francis
Taylor & Francis Group

Statistics
A Journal of Theoretical and Applied Statistics

ISSN: 0233-1888 (Print) 1029-4910 (Online) Journal homepage: https://www.tandfonline.com/loi/gsta20

Sample size calculations for mixture alternatives in a control group vs. treatment group design

Daniel R. Jeske & Weixin Yao

To cite this article: Daniel R. Jeske & Weixin Yao (2020) Sample size calculations for mixture alternatives in a control group vs. treatment group design, Statistics, 54:1, 97-113, DOI: 10.1080/02331888.2020.1715407

To link to this article: https://doi.org/10.1080/02331888.2020.1715407

STATISTICS
2020, VOL. 54, NO. 1, 97–113
https://doi.org/10.1080/02331888.2020.1715407

Taylor & Francis
Taylor & Francis Group

Sample size calculations for mixture alternatives in a control group vs. treatment group design

Daniel R. Jeske and Weixin Yao

Department of Statistics, University of California, Riverside, CA, USA

ABSTRACT

We consider a control group versus treatment group experimental design and assert that powering the design for a potential treatment effect that is represented by a pure shift of the control group distribution is usually unrealistic. Instead, we propose the use of a mixture model as the design alternative in anticipation that there might be a sub-population in the treated population whose responses come from the same control group distribution. When the responses in the treatment group follow a mixture model, the sample size found by the traditional pure shift alternative based method is demonstrated to be under-powered. We develop a new sample size formula for the Wilcoxon test statistic and propose a more general definition of the treatment effect. Method of moment estimators of the treatment effect are proposed and their bias and mean squared error properties are evaluated.

ARTICLE HISTORY
Received 4 March 2018
Accepted 3 January 2020

KEYWORDS
Wilcoxon test; power; mixture models; treatment effect; sample size

1. Introduction

Sample size calculation has long been an important aspect of experimental design. In a control group vs. treatment group design, where the treatments are allocated at random, the method employed for addressing this question is most often based on the assumption that the potential treatment effect can be represented by a shift in the distribution of the control group responses. In this paper, we assert that powering the design based on a pure shift alternative is usually unrealistic. An alternative assumption that has empirical justification is that the shifted distribution only applies to a subpopulation of treated subjects, and the subpopulation is not yet identifiable [1,2]. To this end, a mixture model [3,4] might be used as the design alternative for the treatment group response distribution in anticipation that there might be a sub-population of the treated population whose responses still come from the control group distribution. When such a sub-population exists, we will demonstrate that the sample sizes needed to achieve the desired power for the design can be appreciably larger.

It is well known that sub-populations of treated patients exist in oncology trials [5–7]. A group fMRI example motivated a recent call for more attention to be given to mixture

CONTACT Daniel R. Jeske daniel.jeske@ucr.edu

alternatives for comparing two (alternative) treatments (by a hypothesis test), stating that medical applications, psychiatric-genetics and personalized medicine are important applications where mixtures are plausible alternatives [8]. When treatment effects have the potential to be subpopulation specific, the use of mixture models to represent the response distribution within the treatment group is compelling. A paper that is related to our research addresses simultaneous testing for an overall population treatment effect and a specified sub-population treatment effect, where the sub-population is a-priori identified through a biomarker [9]. The overall treatment effect includes the sub-population whose responses come from the control group distribution. However, we consider a setting in which the relevant sub-populations cannot be identified by a biomarker or another background variable.

Denote the cumulative distribution functions associated with a response from the control group and the treatment group by F and G, respectively. A popular nonparametric test of $H_0 : F = G$ is the Wilcoxon rank-sum test [10]. Analyses of clinical trials frequently utilize this test [11]. The test statistic is asymptotically normally distributed under both the null hypothesis and the alternative hypothesis [12]. These results can be used to design a test of H_0 that has a size α and has a desired level of power for a particular design alternative. Shift alternatives of the form $G(u) = F(u - \delta)$, for a specified $\delta = K\sigma$, are frequently used, where σ is the standard deviation of F. In this paper we assume, without loss of generality, that $K > 0$ and we use a mixture model for the design alternative of the form $G(u) = (1 - \theta)\, F(u) + \theta\, F(u - \delta)$, where $\theta \in (0, 1)$ is also specified. The traditional sample size calculation method relies on the assumption that $\theta = 1$, which is a special case of the assumption considered here. With the mixture model, the treatment effect is represented by the pair (δ, θ) and the average treatment effect in the population is $\Delta = \delta\theta$.

Sample sizes that are determined based on the mixture design alternative can be appreciably larger compared to when the design alternative is a pure shift. The following example illustrates this point. Suppose a researcher wishes to use a Wilcoxon rank-sum test to detect a treatment effect that is hypothesized to be a pure 0.5σ shift in the mean. A size 5% test requires a sample size of 53 subjects in each group when F is a normal distribution. Suppose the treatment effect is hypothesized to be a mixture alternative whose components have shift sizes of 0 with probability $1 - \theta$ and $K\sigma$ with probability θ. The average shift size in the population is $\Delta = (K\sigma)\,\theta = (K\theta)\sigma$. Suppose Δ is held fixed at 0.5σ by keeping $K\theta = 0.5$ as θ varies (implying that K varies accordingly). Then the sample size required to attain 80% power is shown in Figure 1 as a function of θ.

For example, Figure 1 shows that the sample size required to attain 80% power is about 60 instead of 53 when $\theta = .5$ and 125 when θ is as low as .2.

Our premise is that mixture models are realistic alternatives for the distribution of responses from the treatment group when determining the sample size for control group versus treatment comparisons. We highlight this point of view and develop methods for sample size calculation with this assumption. We also demonstrate that the traditional sample size calculation based on a pure shift alternative can be much smaller than what is needed to detect a mixture alternative. For some intuition, if the treatment group includes additional subjects for whom the treatment is ineffective, they are bound to reduce the power in detecting the treatment effect even if the average treatment effect in the population stays fixed. To our knowledge, there is no previous work that has addressed the sample

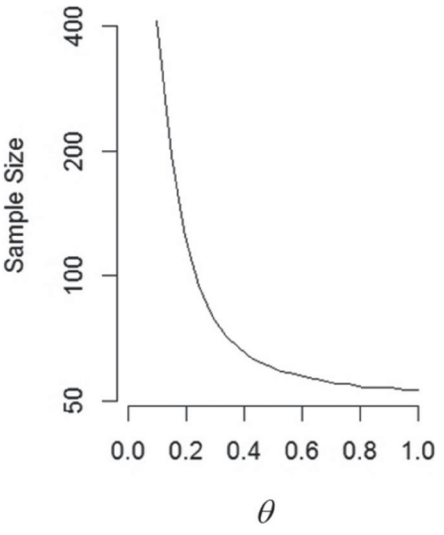

Figure 1. Required sample size (per group), as a function of θ, so that the power of a size 5% Wilcoxon rank-sum test is 80% when Δ is held fixed at 0.5σ (F is normal).

size questions we answer. In the mixture model setting, the treatment effect is described by both θ and δ.

While use of δ is a traditional choice in sample size calculations, θ is a new parameter that is chosen as part of the expanded treatment effect. The value of θ could be chosen in one of two ways. First, it could be subjectively chosen by its clear interpretation as the fraction of the treated population whose responses come from the shifted distribution $F(u-\delta)$. Second, just as the variance parameter σ^2 in the simple one sample normal-theory sample size calculation is often estimated from a pilot experiment, the same approach can be used here. Specifically, a pilot experiment can provide a maximum likelihood estimate (MLE) of θ. Since the pilot experiment has data from both the control group and treatment group, the estimation of the mixture parameter θ is more stable compared to when data is only available from the treatment group. To illustrate the usefulness of an estimate of θ obtained this way, Table 1 shows simulation estimates of the bias and the mean squared error (MSE) of the MLE of θ when F is a standard normal distribution and a standard Laplace distribution. The MLE of θ for each of 1,000 simulated data sets was found using the EM algorithm. It can be seen from Table 1 that even relatively small pilot experiments provide informative ranges for the value of θ that could guide a formal sample size calculation.

For additional insight, the impact of using an estimated θ on the power obtained by using the formulae derived under the assumption of a known value of θ was also investigated. The experiment was done for by simulating data sets from four selected combinations of $(\theta, \delta) \in \{(.7,1),(.7,2),(.9,1),(.9,2)\}$. Based on each combination, and for pilot sample sizes ranging from 15 to 500, we simulated 1,000 pilot samples for initial studies and estimated θ based on those generated pilot samples. From the 1,000 estimates of θ, we then calculated the sample size needed to achieve the desired 80% power for fixed Δ, and the corresponding power that would actually be achieved.

Table 1. Bias and MSE of MLE of θ when F is a standard normal distribution or a standard Laplace distribution.

		bias $(\hat{\theta}_{MLE})$		MSE $(\hat{\theta}_{MLE})$	
(θ, δ)	$m = n$	Normal	Laplace	Normal	Laplace
(.7, 1.0)	15	−.021	.010	.079	.045
	25	−.002	.001	.075	.036
	50	.009	.005	.062	.022
(.7, 2.0)	15	.002	.013	.039	.024
	25	.016	.006	.030	.016
	50	.011	.001	.016	.008
(.9, 1.0)	15	−.117	−.049	.076	.029
	25	−.085	−.037	.060	.020
	50	−.048	−.020	.036	.014
(.9, 2.0)	15	−.054	−.006	.037	.011
	25	−.003	−.003	.013	.008
	50	.003	.001	.006	.005

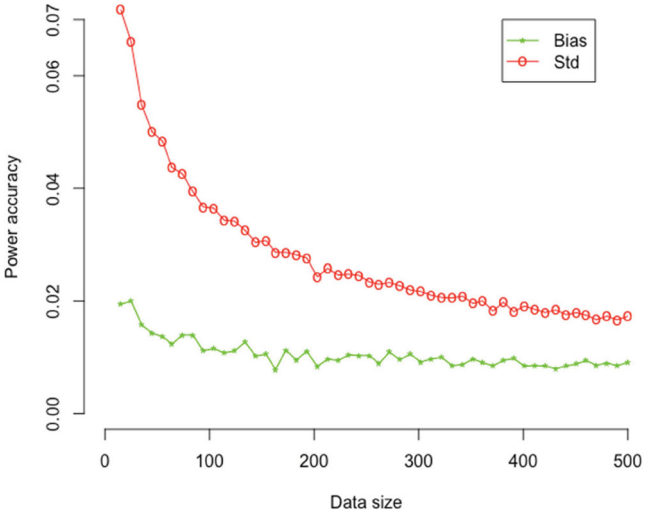

Figure 2. The plot of power accuracy, in terms of Bias and Standard devitations (std) versus the pilot sample size, from which θ is estimated, where F is a standard Laplace distribution and $(\theta, \delta) = (0.7,1)$. The sample sizes are calculated based on estimated θ similar to Tables 2–5.

Therefore, for each of the selected combinations of (θ, δ) we have 1,000 estimated θ values and corresponding achieved powers if those estimates are used as if they were the true values of θ. Figures 2 and 3 report the bias and standard deviation (std) of those achieved powers as compared to the target power of 0.8. To save space, we only show the results for (θ, δ) equal to (0.7,1) and (0.9,1). Based on Figures 2 and 3, even with the small sample sizes, the bias and std of the achieved power based on estimated θ are manageable from a practical point of view. In addition, as expected, both bias and std decrease when the data size increases.

We develop our new approach to sample size calculation using the Wilcoxon rank-sum test, but it can be adapted to other test statistics. The Wilcoxon test is robust with respect to departures from normality, while being only slightly less efficient when the data are

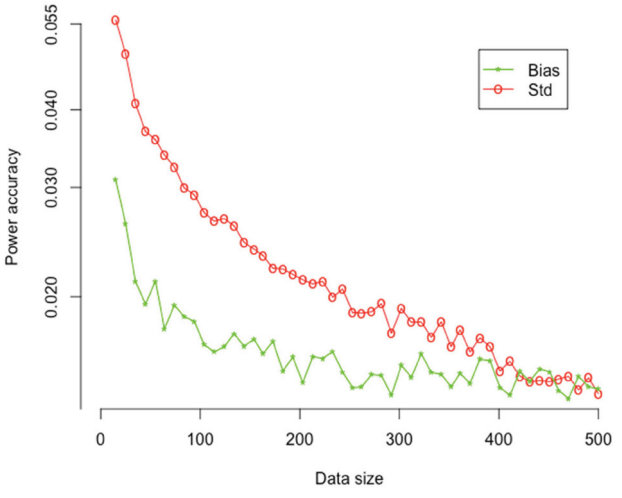

Figure 3. The plot of power accuracy, in terms of Bias and Standard devitations (std) versus the pilot sample size, from which θ is estimated, where F is a standard Laplace distribution and $(\theta, \delta) = (0.9,1)$. The sample sizes are calculated based on estimated θ similar to Tables 2–5.

normally distributed [13,14]. We will also consider the two-sample Kolmogorov–Smirnov (KS) test [15] for additional illustrations.

The rest of this paper is organized as follows. In Section 2 we develop the pertinent technical results. Specifically, a sample size formula for the mixture alternative is provided if using the Wilcoxon rank-sum test, some simplifications for location-scale families are discussed, and benchmark optimal linear rank test statistics for mixture alternatives are developed. Section 3 illustrates specific sample size calculations for the cases when F is from the Normal, Logistic, Laplace, and $t(3)$ location-scale families. In Section 4 we propose and evaluate nonparametric estimators of the mixture parameters (θ, δ) that characterize the treatment effect when the distribution of the responses from the treatment group is $G(u) = (1-\theta)\,F(u) + \theta\,F(u-\delta)$. In Section 5, a real example is given to illustrate the new sample size calculations. We close the paper with a summary and discussion in Section 6.

2. Technical developments

2.1. Wilcoxon rank-sum test

Suppose $X_1, \ldots, X_m$ are iid from a continuous F and $Y_1, \ldots, Y_n$ are iid from a continuous G. Let $R = (R_1, \ldots, R_n)$ denote the vector of ranks of the $Y_1, \ldots, Y_n$ observations. The Wilcoxon rank-sum test statistic is $W = \sum_{i=1}^{n} R_i$. Let $N = m + n$ and assume that $\lambda = \lim_{N\to\infty} (m/N) < 1$, the asymptotic ratio of sample sizes. It is well known [12] that

$$\frac{W - n(m+n+1)/2}{\sqrt{(mn(m+n+1)/12)}} \tag{1}$$

has a limiting (as $N \to \infty$) standard normal distribution under $H_0 : F = G$. This result is often used to determine critical values for testing H_0. For example, if larger responses

are anticipated from the treatment group a one-sided size α test would reject when $W > n(m+n+1)/2 + z_\alpha\sqrt{mn(m+n+1)/12}$.

The limiting distribution of W under general alternatives G is also known [12]. Define,

$$\begin{aligned}\gamma(F,G) &= P(X_1 < Y_1)\\ \xi_1(F,G) &= P(X_1 < Y_1, X_1 < Y_2) - \gamma^2(F,G),\\ \xi_2(F,G) &= P(X_1 < Y_1, X_2 < Y_1) - \gamma^2(F,G).\end{aligned} \tag{2}$$

Then $(\sqrt{N}/mn)(W - \mu_W)$ has a limiting normal distribution with mean zero and variance $\xi_1(F,G)/\lambda + \xi_2(F,G)/(1-\lambda)$, where $\mu_W = n(m\gamma(F,G) + (n+1)/2)$. It follows that the asymptotic power of the Wilcoxon test is

$$\pi(G) = \Phi\left(\frac{\gamma(F,G) - (1/2) - z_\alpha\sqrt{(m+n+1/12mn)}}{\sqrt{(\xi_1(F,G)/m) + (\xi_2(F,G)/n)}}\right).$$

2.2. Sample size formula for detecting mixtures

In the mixture setting that has $G(u) = (1-\theta)F(u) + \theta F(u-\delta)$, we have $\gamma(F,G) \equiv \gamma(\theta,\delta,F)$, $\xi_1(F,G) \equiv \xi_1(\theta,\delta,F)$ and $\xi_2(F,G) \equiv \xi_2(\theta,\delta,F)$.

Define $\rho = n/m$. Asymptotic group sizes that detect the alternative G with power $1-\beta$ are approximate integer solutions to the equation

$$\begin{aligned}\gamma(\theta,\delta,F) - \frac{1}{2} &= z_\alpha\sqrt{\frac{m+n+1}{12mn}} + z_\beta\sqrt{\frac{\xi_1(\theta,\delta,F)}{m} + \frac{\xi_2(\theta,\delta,F)}{n}}\\ &\approx z_\alpha\sqrt{\frac{\rho+1}{12\rho}}\sqrt{\frac{1}{m}} + z_\beta\sqrt{\xi_1(\theta,\delta,F) + \frac{\xi_2(\theta,\delta,F)}{\rho}}\sqrt{\frac{1}{m}}.\end{aligned}$$

Based on straightforward calculations we have the following sample size result.

Proposition 2.1: *Asymptotic group sizes that detect the alternative* $G(u) = (1-\theta)F(u) + \theta F(u-\delta)$ *with power* $1-\beta$ *are*

$$m = \left(\frac{z_\alpha\sqrt{(\rho+1/12\rho)} + z_\beta\sqrt{\xi_1(\theta,\delta,F) + (\xi_2(\theta,\delta,F)/\rho)}}{\gamma(\theta,\delta,F) - (1/2)}\right)^2, \quad n = \rho m. \tag{3}$$

For the case where $\rho = 1$, *so that* $m = n$, *the sample size in (3) becomes*

$$m = \left(\frac{z_\alpha/\sqrt{6} + z_\beta\sqrt{\xi_1(\theta,\delta,F) + \xi_2(\theta,\delta,F)}}{\gamma(\theta,\delta,F) - 1/2}\right)^2. \tag{4}$$

2.3. Minimum cost sample sizes

If the cost of obtaining an observation from the control group and the treatment group is different, it is natural to consider drawing a smaller sample from the group that is associated with the higher cost. Suppose C_1 is the cost per observation from the control group and

C_2 is the cost per observation from the treatment group. We find m and n that minimize the total cost of all of the samples, $mC_1 + nC_2$ subject to $\pi(G) = 1 - \beta$. Equivalently, the constraint can be expressed as

$$\gamma(\theta, \delta, F) - \frac{1}{2} = z_\alpha \sqrt{\frac{m+n+1}{12mn}} + z_\beta \sqrt{\frac{\xi_1(\theta, \delta, F)}{m} + \frac{\xi_2(\theta, \delta, F)}{n}}.$$

Define a Lagrange function as

$$L(m, n, \phi) = mC_1 + nC_2 + \phi \left(\gamma(\theta, \delta, F) - \frac{1}{2} - z_\alpha \sqrt{\frac{m+n+1}{12mn}} - z_\beta \sqrt{\frac{\xi_1(\theta, \delta, F)}{m} + \frac{\xi_2(\theta, \delta, F)}{n}} \right),$$

where ϕ is the Lagrange multiplier. Then based on the two equations $(\partial L(m, n, \phi)/\partial m) = 0$ and $(\partial L(m, n, \phi)/\partial n) = 0$, we have the following result.

Proposition 2.2: *The optimal ratio ρ to minimize the cost $mC_1 + nC_2$ subject to $\pi(G) = 1 - \beta$ is*

$$\rho^2 = \frac{C_1}{C_2} \frac{z_\alpha \sqrt{(1/12)}(m+n+1)^{-1/2} + z_\beta (n\,\xi_1(\theta, \delta, F) + m\,\xi_2(\theta, \delta, F))^{-1/2} \xi_2(\theta, \delta, F)}{z_\alpha \sqrt{(1/12)}(m+n+1)^{-1/2} + z_\beta (n\,\xi_1(\theta, \delta, F) + m\,\xi_2(\theta, \delta, F))^{-1/2} \xi_1(\theta, \delta, F)}.$$

Therefore, the optimal ratio of group sizes depends on the unknown values $\xi_1(\theta, \delta, F)$ and $\xi_2(\theta, \delta, F)$. Our experience suggests that $\xi_1(\theta, \delta, F)$ and $\xi_2(\theta, \delta, F)$ are approximately equal, in which case we have $\rho \approx \sqrt{(C_1/C_2)}$. Equation (3) *then provides m and n. Specifically, if $C_1 = C_2$, then $\rho = 1$ is recommended to minimize the total cost.*

2.4. Location-scale families

We investigate the sample size formula in (3) under the case where F belongs to a location-scale family. That is, $F(u) = \Psi((u - \mu)/\sigma)$, where μ is the location parameter, σ is the scale parameter, and $\Psi(\cdot)$ is a known cumulative distribution function involving no parameters and having probability density function $\psi(\cdot)$. Within the location-scale family, $X \sim F$ if and only if $X \sim \mu + \sigma Z$, where $Z \sim \Psi$. No generality is lost by assuming that $\Psi(\cdot)$ has zero mean and unit variance so that μ and σ can be interpreted as the mean and standard deviation of F. Under the mixture alternative $Y \sim \mu + \sigma Z$ with probability $(1 - \theta)$ and $Y \sim \delta + \mu + \sigma Z$ with probability θ. Proposition 2.3 is proved in Appendix A.

Proposition 2.3: *If F is a location-scale distribution, then the power of the Wilcoxon rank-sum test for mixture alternatives depends on Ψ and (θ, K), where $K = \delta/\sigma$.*

One-dimensional integral representations of (2) *are,*

$$\gamma(\theta, \delta, F) = \int_{-\infty}^{\infty} \Psi(y)\,[(1-\theta)\,\psi(y) + \theta\,\psi(y-K)]\,dy$$

$$\xi_1(\theta, \delta, F) = \int_{-\infty}^{\infty} [1 - (1-\theta)\,\Psi(x) - \theta\,\Psi(x-K)]^2 \psi(x)\,dx - \gamma^2(\theta, \delta, F)$$

$$\xi_2(\theta, \delta, F) = \int_{-\infty}^{\infty} \Psi^2(y)[(1-\theta)\,\psi(y) + \theta\,\psi(y-K)]\,dy - \gamma^2(\theta, \delta, F). \quad (5)$$

2.5. Optimal scores

The Wilcoxon test statistic is a special case of a more general class of linear rank statistics defined by $S = \sum_{i=1}^{n} a(R_i)$, where $a(1) < \cdots < a(N)$ is a sequence of rank scores. We would reject H_0 for large values of S. The Wilcoxon test statistic chooses $a(i) = i$. Choosing scores is driven by what type of alternative is important to detect. In the case of a pure shift alternative $G(u) = F(u - \delta)$, optimal scores $a_O(i)$ can be derived in the sense that there exists an $\varepsilon > 0$ such that $S = \sum_{i=1}^{n} a_O(R_i)$ provides the most powerful test for $0 < \delta < \varepsilon$. The optimal scores are

$$a_O(i) = E\left[\frac{-f'(F^{-1}(U_{i:N}))}{f(F^{-1}(U_{i:N}))}\right] \tag{6}$$

where $\{U_{i:N}\}_{i=1}^{N}$ are the order statistics of a random sample of size N from Uniform(0,1) distribution (see equation 9.1.23 in reference [12]). Proposition 2.4 states the optimal scores for pure shift alternatives also apply to mixture alternatives, and is proved in Appendix A.

Proposition 2.4: *The optimal scores for a linear rank statistic for a testing H_0 against mixture alternatives $G(u) = (1 - \theta)\, F(u) + \theta\, F(u - \delta)$ are given by Equation* (6).

For location-scale distributions $F(u) = \Psi((u - \mu)/\sigma)$, the optimal scores simplify to $a_O(i) = E[(-\psi'(\Psi^{-1}(U_{i:N}))/\psi(\Psi^{-1}(U_{i:N})))]$, implying they only depend on F through Ψ. For a given class of distributions the relative efficiency of the Wilcoxon test compared to the test based on the optimal scores can be measured by the ratios of the sample sizes they require.

3. Comparisons of sample sizes

In this section, we evaluate the sample sizes required for some commonly used location-scale distributions and show the impact that the mixture alternative has on the required sample size. For simplicity, we consider only the case $n = m$, so that $N = 2n$. For the calculations shown in the Tables 2–5 below, we assume $\delta = K\sigma$, so that defining $K' = K\theta$ the average treatment effect in the population is $\Delta = K'\sigma$. In our tables we hold K' fixed. For each value of θ, we set K to K'/θ. Then (θ, K) is used with (5) to compute $\gamma(\theta, \delta, F)$, $\xi_1(\theta, \delta, F)$ and $\xi_2(\theta, \delta, F)$, and (4) is used to compute the approximate sample size calculation for the case $m = n$.

3.1. Normal distributions

For the case when F is a normal distribution, $\Psi(u) = \Phi(u)$, and $\psi(u) = \phi(u)$. Table 2 shows the (rounded up) sample sizes for tests that have size 5% and power 80%. If F is normal, the optimal scores are $a_O(i) = E[\Phi^{-1}(U_{i:N})]$. The numbers reported in Table 2 that are in parentheses are the required sample sizes for this test and were computed as follows. First, asymptotically equivalent scores $\Phi^{-1}(i/N + 1)$ were used, which are obtained by replacing R_i in $a_O(i)$ by its expected value. Next, starting with the required n for the Wilcoxon test the power of the optimal score test was evaluated for decreasing sample sizes until the power became less than 0.8. The power for each sample size in the sequence was based on 100,000 simulated data sets where the $\{X_i\}_{i=1}^{n}$ came from F and the $\{Y_i\}_{i=1}^{n}$ came from the mixture G.

An alternative nonparametric test that might be considered for testing H_0 against mixtures is the KS test. Simulated power estimates of the KS test using the sample sizes shown in Table 2 averaged .681 and ranged between .608 and .725, which demonstrates the superiority of the Wilcoxon test.

3.2. Logistic distributions

For the case when F is a logistic distribution, $\Psi(u) = (1 + \exp(-cu))^{-1}$, and $\psi(u) = (c\exp(-cu)/[1 + \exp(-cu)]^2)$, where $c = \pi/\sqrt{3}$. Table 3 shows the (rounded up) sample sizes for tests that have size 5% and power 80%. If F is logistic the optimal scores are $a_O(i) = i$, and thus the optimal linear rank test is the Wilcoxon test itself [12]. Simulated power estimates of the KS test using the sample sizes shown in Table 3 averaged .707 and ranged from .605 to .763.

3.3. Laplace distributions

For the case when F is a Laplace distribution, $\Psi(u) = \begin{cases} e^{cu}/2, & \text{if } u < 0 \\ 1 - e^{-cu}/2, & \text{if } u \geq 0 \end{cases}$, and $\psi(u) = c\,e^{-|cu|}/2$, where $c = \sqrt{2}$. Table 4 shows the (rounded up) sample sizes for tests that have size 5% and power 80%. Let B denote a binomial random variable with parameters N and 0.5. It is shown in Appendix B that if F is Laplace, optimal scores are $a_O(i) = 2\Pr(B < i - 1) - 1$. Sample sizes required for the optimal linear rank test are shown as the numbers in parentheses in Table 4. Simulated power estimates of the KS test using the sample sizes shown in Table 4 averaged .758 and ranged from .672 to .819.

3.4. Location-scale t_3 distributions

For the location-scale t-distribution with 3 degrees of freedom, $\psi(u) = (2/\pi(1 + u^2)^2)$. Table 5 shows the (rounded up) sample sizes for tests that have size 5% and power 80%. It is shown in Appendix B that if F is location-scale t_3, optimal scores are $a_O(i) = E[(\Psi^{-1}(U_{i:N}))/3 + \Psi^{-1}(U_{i:N})^2]$. Asymptotically equivalent scores of $((t_{3,i/(N+1)})/3 + t^2_{3,i/(N+1)})$ were used in our calculations. Sample sizes required for the optimal linear rank test are shown as the numbers in parentheses in Table 5. Simulated power estimates of the KS test using the sample sizes shown in Table 5 averaged .747 and ranged from .635 to .838.

Table 2. Sample sizes $m = n$ for a 5% test based on W that will achieve 80% power for the mixture alternative (Ψ is standardized normal distribution).

	K'			
θ	.4	.6	.8	1
.5	89 (86)	44 (43)	29 (28)	22 (22)
.6	86 (84)	41 (40)	26 (26)	19 (19)
.7	84 (82)	40 (39)	24 (24)	17 (17)
.8	83 (80)	38 (38)	23 (23)	16 (16)
.9	83 (80)	38 (37)	22 (22)	15 (15)
1.0	82 (79)	37 (36)	21 (21)	14 (14)

Notes: Numbers in parentheses are sample sizes for a 5% test based on optimal linear rank test statistic.

Table 3. Sample sizes $m = n$ for a 5% test based on W that will achieve 80% power for the mixture alternative (Ψ is standardized logistic distribution).

	K'			
θ	.4	.6	.8	1
.5	80	41	28	22
.6	77	38	24	18
.7	75	36	22	16
.8	74	34	21	15
.9	73	34	20	14
1.0	72	33	19	13

Table 4. Sample sizes $m = n$ for a 5% test based on W that will achieve 80% power for the mixture alternative (Ψ is standardized Laplace distribution).

	K'			
θ	.4	.6	.8	1
.5	67 (67)	37 (37)	27 (27)	22 (22)
.6	62 (61)	33 (33)	22 (22)	17 (17)
.7	59 (55)	30 (29)	20 (19)	15 (14)
.8	58 (52)	28 (27)	18 (17)	13 (13)
.9	56 (50)	27 (25)	17 (16)	12 (12)
1.0	55 (49)	26 (25)	16 (15)	11 (11)

Table 5. Sample sizes $m = n$ for a 5% test based on W that will achieve 80% power for the mixture alternative (Ψ is standardized t_3 distribution).

	K'			
θ	.4	.6	.8	1
.5	53 (52)	30 (30)	23 (23)	20 (20)
.6	49 (47)	26 (26)	19 (18)	15 (15)
.7	46 (44)	24 (23)	16 (16)	13 (12)
.8	45 (43)	22 (21)	14 (14)	11 (11)
.9	44 (42)	21 (20)	13 (13)	10 (9)
1.0	43 (41)	20 (20)	13 (12)	9 (9)

3.5. Discussion

Tables 2–5 show that calculating the sample size based on the assumption that $\theta = 1$ can result in an underpowered experiment when the true alternative is a mixture. The sample sizes needed for the four distributions order themselves smallest to largest according to t_3, Laplace, logistic and normal. To intuitively understand this ordering, consider Figure 4 where $K' = .4$ and for each $\theta \in [.5, 1]$ the Kullback-Leibler (KL) distances [16] between F and G are shown (as lines) for the four distributions: t_3 (blue), Laplace (green), logistic (red) and normal (black). It can be seen that the largest-to-smallest KL distances order the distributions the same way that the smallest-to-largest sample sizes order them. Other values of K' lead to the same ordering of distributions.

Tables 2–5 were constructed by holding the average treatment effect in the population fixed. That is, $K' = K\theta$ was held fixed as θ was varied. In this way, θ and K do not

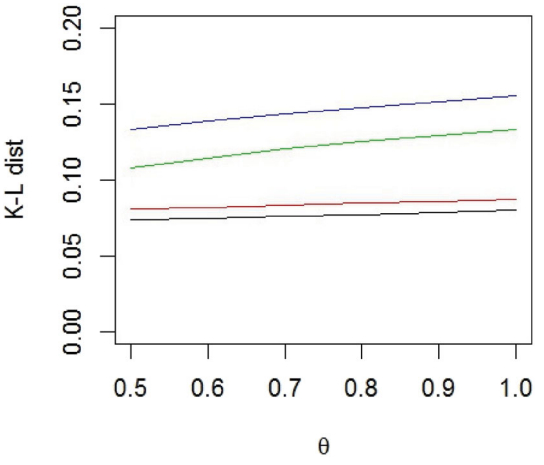

Figure 4. K–L distances for $K' = .4$ that provide intuition for the ordering of sample sizes across Tables 1–4. Coloured lines: t_3 (blue), Laplace (green), logistic (red) and normal (black).

Table 6. Sample sizes $m = n$ for a 5% test based on W that will achieve 80% power for the mixture alternative (Ψ is standardized normal distribution).

	K			
θ	.4	.6	.8	1
.5	331	152	89	60
.6	230	105	61	41
.7	169	77	45	30
.8	129	59	34	23
.9	102	46	27	18
1.0	82	37	21	14

vary independently. An alternative perspective is that a researcher could think about and select K and θ independently from one another. Specifically, a researcher might first choose the shift size K based on what they believe is plausible for a treatment effect, or perhaps choose it as the smallest shift size that would have any practical consequence. Separately, the researcher might use other information that informs about the choice of θ for what fraction of a treated population has potential to respond to the treatment. To illustrate this alternative perspective, Table 6 is analogous to Table 2, except that K and θ are allowed to independently vary. The results similarly show the importance of accounting for $\theta < 1$ when designing the study.

4. Estimating the treatment effect with the mixture model

Let (μ_F, σ_F^2) denote the mean and variance of the control group response, and (μ_G, σ_G^2) be the same for the treatment group response. The treatment effect $\Delta = \mu_G - \mu_F$ is appropriate under the assumption that $\theta = 1$. Here we assume that the difference is an appropriate scale for comparing the means; in some settings, some other quantities, such as the ratio, are more appropriate. In a mixture setting, the treatment effect is better thought of as the pair (θ, δ). In this section, we propose a simple moment estimator for (θ, δ) that does not

depend on the distribution assumptions for F and G. To that end, Proposition 4.1 (proved in Appendix A) provides insight into the inadequacy of Δ in mixture settings.

Proposition 4.1: *The treatment effect parameters* (θ, δ) *in the mixture setting satisfy*

$$\theta = \left\{1 + \frac{\sigma_G^2 - \sigma_F^2}{\Delta^2}\right\}^{-1} \tag{7}$$

$$\delta = \Delta\left\{1 + \frac{\sigma_G^2 - \sigma_F^2}{\Delta^2}\right\}. \tag{8}$$

Note that the mixture model for the treatment group implies $\sigma_G^2 \geq \sigma_F^2$*, with equality if and only if* $\theta = 1$*. Substituting* (7) *into* (8) *gives* $\Delta = \delta\theta$*. As* θ *decreases from* 1, Δ *becomes increasingly inadequate as a measure of the treatment effect within the sub-population where the treatment effect is a shift of the distribution that governs responses in the control group.*

Proposition 4.1 motivates the following modified method of moments estimator,

$$\hat{\theta} = \left\{1 + \frac{(S_Y^2 - S_X^2)_+}{(\bar{Y} - \bar{X})_+^2 + \varepsilon}\right\}^{-1} \tag{9}$$

$$\hat{\delta} = (\bar{Y} - \bar{X})_+\left\{1 + \frac{(S_Y^2 - S_X^2)_+}{(\bar{Y} - \bar{X})_+^2 + \varepsilon}\right\}, \tag{10}$$

where $t_+ = t$ if $t > 0$ and 0 otherwise, $(\bar{X}, S_X^2)$ are the mean and variance of the control group observations, $(\bar{Y}, S_Y^2)$ are the same for the treatment group observations, and ε is a small positive number that bounds the denominators away from zero. In our experiments with $\varepsilon \in [.01, .1]$ we found that a value of ε around 0.05 usually works well.

Tables 7 and 8 summarize a simulation study that investigates the bias and MSE of the proposed moment estimators when F is a normal distribution and a Laplace distribution, respectively. For each considered combination of (θ, δ), bias and MSE were evaluated for the minimum sample size n that is needed to achieve 80% power, and then also for $2n$. The results show positive relative bias in $\hat{\theta}$ that ranges from 10% to 20% and the bias does not seem to decrease when the sample size increases, which might be due to the paucity of information about mixture models when the components are close and the sample size is not too large. The relative bias in $\hat{\delta}$ tends to be much smaller. The MSE results show an anticipated n^{-1} rate of decrease for $\hat{\delta}$. However, the MSE of $\hat{\theta}$ does not show an anticipated n^{-1} rate of decrease for the considered simulation settings and it decreases slower than the MSE of $\hat{\delta}$. We note that the estimation accuracy of δ is much worse for the heavy-tailed Laplace distribution. Results for the logistic distribution showed similar patterns, however, results with the $t(3)$ distribution showed both estimators have comparatively worse MSE performance as a result of the heavy tails. Alternative estimators that are robust to heavy-tailed distributions and outliers would be a useful future research direction.

5. Example

Conover and Salsburg discuss how treatments can affect only a subset of treated patients [2]. They refer to that subset of patients as 'responders'. In their paper, they discuss a

Table 7. Bias and MSE when $\sigma = 1$, F is a normal distribution and $\varepsilon = 0.03$.

(θ, δ)	$m = n$	Bias $(\hat{\theta})$	MSE $(\hat{\theta})$	Bias $(\hat{\delta})$	MSE $(\hat{\delta})$
(.5, .5)	215	0.147	0.117	−0.021	0.081
	430	0.166	0.102	−0.053	0.048
(.5, .75)	100	0.120	0.106	0.002	0.161
	200	0.099	0.075	−0.021	0.087
(.5, 1)	60	0.082	0.089	0.000	0.231
	120	0.075	0.063	−0.005	0.124
(.6, .5)	149	0.076	0.098	0.048	0.103
	298	0.088	0.080	0.002	0.053
(.6, .75)	69	0.059	0.087	0.071	0.173
	138	0.066	0.071	0.023	0.098
(.6, 1)	41	0.050	0.086	0.096	0.259
	82	0.058	0.060	0.010	0.127

Table 8. Bias and MSE when $\sigma = 1$, F is a Laplace distribution and $\varepsilon = 0.03$.

(θ, δ)	$m = n$	Bias $(\hat{\theta})$	MSE $(\hat{\theta})$	Bias $(\hat{\delta})$	MSE $(\hat{\delta})$
(.5, .5)	148	0.131	0.141	0.089	0.223
	296	0.133	0.124	0.032	0.113
(.5, .75)	74	0.117	0.127	0.098	0.339
	148	0.113	0.109	0.049	0.212
(.5, 1)	48	0.104	0.115	0.107	0.470
	96	0.099	0.094	0.036	0.266
(.6, .5)	103	0.049	0.123	0.172	0.303
	206	0.075	0.105	0.074	0.138
(.6, .75)	51	0.045	0.114	0.189	0.439
	102	0.071	0.093	0.094	0.210
(.6, 1)	33	0.046	0.102	0.218	0.702
	66	0.062	0.084	0.090	0.288

control group versus treatment group design that was aimed for studying acute painful diabetic neuropathy. In that study, individual patients scored their pain at baseline and after 4 weeks of treatment on a continuous scale. The larger the score, the more severe pain the patient feels. The response variable recorded from each patient was calculated as log(baseline score/final score). If the treatment is effective, we would see higher values of the response variable.

Table 1 in Conover and Salsburg shows responses from 28 patients in the control group and 30 patients from the treatment group. We assume that the study planned to have group sizes $m = n = 30$ but two patients in the control group dropped out. While the rationale for recommending 30 patients in each group is not known to us, the following seems plausible. Suppose it is determined that a size $\alpha = .05$ Wilcoxon test will be carried out. Further, suppose it is desired to have 80% power to detect the alternative where the distribution of the treatment group responses is a pure shift of the control group distribution and the size of the shift is $K = 2/3$ standard deviations. Computing from Equation (5) with Ψ a standard normal distribution and $\theta = 1$, and then using Equation (4), the resulting group size is 30 patients.

Conover and Salsburg suggest the neuropathy application is an example where the response distribution in the treatment group could be anticipated to be a mixture. Keeping $K' = K\theta$ fixed at 2/3, the second column in Table 9 shows the reduction of power for values of θ less than unity. The third column in Table 9 shows the group size needed to maintain 80% power for a shift of $K = 2/3$ standard deviations that is applicable only to the fraction

Table 9. Reduction of power and necessary group sizes in order to achieve design objectives for neuropathy study for alternative fractions of treated patients that are 'responders' when $K' = K\theta$ is fixed at 2/3 (Ψ is standardized normal distribution).

θ	Power	Required group size
.5	.71	37
.6	.75	34
.7	.77	32
.8	.78	31
.9	.79	31
1.0	.80	30

θ of the treated patients. The impact on the necessary sample size that the mixture alternative has can be appreciable. If as few as half of treated patients are 'responders', the group size needs to be increased from 30 to 37.

6. Conclusion and summary

Powering studies for comparing two treatments is usually done by assuming that the treatment effect is constant. Our position is that powering the design under a mixture model for the responses from the treatment group is more realistic. Using the mixture model will reduce the chance of running an underpowered experiment. For the Wilcoxon test statistic, we have shown how to power the design under a proposed mixture alternative. The required sample size to detect a mixture alternative can be substantially larger than what is required to detect a pure shift alternative. Using the mixture alternative when calculating sample sizes entails extending the definition of the treatment effect from δ to (θ, δ). The mixture parameter θ is an additional piece of information that goes into the sample size calculation. Alternative candidates for F should be considered as part of the sensitivity study that is typically done with sample size calculations.

If the treatment group observations follow a mixture model, a natural way to report the treatment effect is by an estimate of (θ, δ). We proposed a modified method of moment estimator for (θ, δ). Understanding the extent to which a treatment alters the distribution of responses within a sub-population can prompt follow-up research that is aimed for identifying specific baseline characteristics of that sub-population.

Acknowledgements

The authors appreciate helpful comments from two referees and an associate editor.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Yao's research is supported by NSF [grant number DMS-1461677] and Department of Energy [award number 10006272].

ORCID

Daniel R. Jeske http://orcid.org/0000-0002-0214-7992

References

[1] Good PI. Detection of a treatment effect when not all experimental subjects will respond to treatment. Biometrics. 1979;35:483–489.
[2] Conover WJ, Salsburg DS. Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to respond to treatment. Biometrics. 1988;44:189–196.
[3] Lindsay BG. Mixture models. Theory, geometry and applications. Hayward: Institute for Mathematical Statistics; 1995.
[4] McLachlan GJ, Peel D. Finite mixture models. New York: John Wiley and Sons; 2000.
[5] FDA Report. (2013). Paving the way for personalized medicine: FDA's role in a new era of medical product development.
[6] Spear BB, Heath-Chiozzi M, Huff J. Clinical applications of pharmacogenetics. Trends Mol Med. 2001;7:201–204.
[7] Manegold C, Adjei A, Bussilino F, et al. Novel active agents in patients with advanced small cell lung cancer without driver mutations who have progressed after first-line chemotherapy. ESMO Open Cancer Horizons. 2017;2:e000118. doi:10.1136/esmoopen-2016-000118.
[8] Rosenblatt JD, Benjamini Y. On mixture alternatives and Wilcoxon's signed-rank test. Am Stat. 2017;72:344–347.
[9] Wang S-J, O'Neill RT, Hung HMJ. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. Pharm Stat. 2007;6:227–244.
[10] Wilcoxon F. Individual comparisons by ranking methods. Biometrics. 1945;1:80–83.
[11] Brumback LC, Pepe MS, Alonzo TA. Using the ROC curve for gauging treatment effects in clinical trials. Stat Med. 2006;25:575–590.
[12] Randles RH, Wolfe DA. Introduction to the theory of nonparametric statistics. New York: John Wiley & Sons; 1979.
[13] Kitchen CM. Nonparametric versus parametric tests of location in biomedical research. Am J Ophthalmol. 2009;147:571–572.
[14] Gibbons JD, Chakraborti S. Nonparametric statistical inference. Berlin, Heidelberg: Springer; 2011.
[15] Hollander M, Wolfe DA. Nonparametric statistical methods. New York: John Wiley & Sons; 1999.
[16] Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat. 1951;22:79–86.

Appendices

Appendix A: Proofs of Propositions 2.3, 2.4, 4.1

Proposition 2.3: *It suffices to show that γ, ξ_1 and ξ_2 only depend on (θ, K) In what follows, let U equal 0 or 1, depending on whether $Y \sim \mu + \sigma Z$ or $Y \sim \delta + \mu + \sigma Z$, and let Z_1, Z_2, Z_3 be independent and identically distributed realizations from Ψ. Then,*

$$\begin{aligned}\gamma &= P(X_1 < Y_1)\\ &= (1-\theta)\,P(X_1 < Y_1|U_1 = 0) + \theta\,P(X_1 < Y_1|U_1 = 1)\\ &= (1-\theta)\,P(Z_1 < Z_2) + \theta\,P(\mu + \sigma Z_1 < \delta + \mu + \sigma Z_2)\\ &= (1-\theta)/2 + \theta\,P(Z_1 - Z_2 < \delta/\sigma)\end{aligned}$$

Therefore γ depends on parameters only through (θ, K). Next consider,

$$\begin{aligned} &P(X_1 < Y_1, X_1 < Y_2) \\ &= \sum_{u_1=0}^{1} \sum_{u_2=0}^{1} \theta^{u_1+u_2}(1-\theta)^{2-(u_1+u_2)} P(X_1 < Y_1, X_1 < Y_2 | U_1 = u_1, U_2 = u_2) \\ &= \sum_{u_1=0}^{1} \sum_{u_2=0}^{1} \theta^{u_1+u_2}(1-\theta)^{2-(u_1+u_2)} P(Z_1 - Z_2 < u_1\delta/\sigma, Z_1 - Z_3 < u_2\delta/\sigma). \end{aligned}$$

The joint distribution of $(Z_1 - Z_2, Z_1 - Z_3)'$ does not depend on any parameters.

Therefore, $P(X_1 < Y_1, X_1 < Y_2)$, and hence ξ_1, depends on parameters only through (θ, K). Similarly,

$$\begin{aligned} P(X_1 < Y_1, X_2 < Y_1) &= \sum_{u_1=0}^{1} \theta^{u_1}(1-\theta)^{1-u_1} P(X_1 < Y_1, X_2 < Y_1 | U_1 = u_1) \\ &= \sum_{u_1=0}^{1} \theta^{u_1}(1-\theta)^{1-u_1} P(Z_1 - Z_2 < u_1\delta/\sigma, Z_3 - Z_2 < u_1\delta/\sigma), \end{aligned}$$

showing that ξ_2 depends on parameters only through (θ, K). Consequently, the asymptotic power of the Wilcoxon rank-sum test only depends on (θ, K).

Proposition 2.4: *Let Q_i $(i = 1, \ldots, m)$ and R_j $(j = 1, \ldots, n)$ be the ranks of the $\{X_i\}_{i=1}^m$ and $\{Y_j\}_{j=1}^n$, respectively, in the combined sample of $N = m + n$ observations. Let $\boldsymbol{R}^* = (Q_1, \ldots, Q_n, R_1, \ldots, R_n)$ denote the vector of ranks, and let $\boldsymbol{r}^*$ denote a particular realization of $\boldsymbol{R}^*$. Applying lemma* 4.3.12 *in reference* [12] *with $G_1(u) = F(u)$, $G_2(u) = G(u)$ and $H(u) = G_1(u)$ gives*

$$\begin{aligned} Pr_{F,G}(\boldsymbol{R}^* = \boldsymbol{r}^*) &= \frac{1}{N!} E\left[\frac{\prod_{i=1}^n [(1-\theta) f(V_{R_i:N}) + \theta f(V_{R_i:N} - \delta)]}{\prod_{i=1}^n f(V_{R_i:N})} \right] \\ &= \frac{1}{N!} E\left[\prod_{i=1}^n \left(1 - \theta + \theta \frac{f(V_{R_i:N} - \delta)}{f(V_{R_i:N})} \right) \right] \end{aligned}$$

where $V_{1:N} < V_{2:N} < \cdots < V_{N:N}$ are order statistics of a random sample of size N from F.

Theorem 9.1.20 *in reference* [2] *proves that* (6) *are the optimal scores for testing $H_0 : F = G$ versus pure shift alternatives of the form $G(u) = F(u - \delta)$. They show that the critical region for the locally most powerful linear rank test contain the rank configuration that have the largest values of $(\partial\, Pr_{F,G}(\boldsymbol{R}^* = \boldsymbol{r}^*)/\partial\delta)|_{\delta=0}$. Their starting point was the expression for $Pr_{F,G}(\boldsymbol{R}^* = \boldsymbol{r}^*)$, but with $\theta = 1$.*

Following their approach, we first fix θ and find that the locally most powerful linear rank test will reject for larger values of $(\partial\, Pr_{F,G}(\boldsymbol{R}^* = \boldsymbol{r}^*)/\partial\delta)|_{\delta=0} = \theta \sum_{i=1}^n E[(-f'(V_{R_i:N})/f(V_{R_i:N}))]$. As θ is just a multiplier in this expression, the test equivalently rejects for large values of $\sum_{i=1}^n E[(-f'(V_{R_i:N})/f(V_{R_i:N}))]$, and therefore does not depend on θ. Additionally, the joint distribution of $V_{1:N} < V_{2:N} < \cdots < V_{N:N}$ is the same as the joint distribution of $F^{-1}(U_{1:N}) < F^{-1}(U_{2:N}) < \cdots < F^{-1}(U_{N:N})$, and thus the test equivalently rejects for large values of $\sum_{i=1}^n E[-f'(F^{-1}(U_{R_i:N}))/f(F^{-1}(V_{R_i:N}))]$, which is of the form $S = \sum_{i=1}^n a(R_i)$, with $a(i) = E[-f'(F^{-1}(U_{i:N}))/f(F^{-1}(U_{i:N}))]$.

Proposition 4.1: *It is straight forward to show that the treatment group has the following mean and variance,*

$$\begin{aligned} \mu_y &= (1-\theta)\mu_x + \theta(\mu_x + \delta) \\ \sigma_y^2 &= \sigma_x^2 + \theta(1-\theta)\delta^2. \end{aligned}$$

Solving these two equations for θ and δ gives Equations (7) *and* (8).

Appendix B: Optimal scores for selected distributions

Equation (6) gives the general formula for the optimal scores. As mentioned, for location-scale distributions the optimal scores do not depend on the location or scale parameters. Thus, an equivalent formula for the optimal scores is $a_O(i) = E[-\psi'(\Psi^{-1}(U_{i:N}))/\psi(\Psi^{-1}(U_{i:N}))]$. The optimal scores for the normal and logistic distributions are widely known. In this appendix, we calculate the less familiar optimal scores for the Laplace and the location-scale t_3 distributions.

For the Laplace distribution,

$$\Psi^{-1}(U_{i:N}) = \begin{cases} \log\ 2\,U_{i:N}, & \text{if } 0 < U_{i:N} \le 1/2 \\ -\log\ 2(1-U_{i:N}), & \text{if } 1/2 \le U_{i:N} < 1 \end{cases},$$

$$\frac{-\psi'(u)}{\psi(u)} = \begin{cases} 1, & \text{if } u > 0 \\ -1, & \text{if } u < 0. \end{cases}$$

Hence,

$$\frac{-\psi'(\Psi^{-1}(U_{i:N}))}{\psi(\Psi^{-1}(U_{i:N}))} = \begin{cases} 1, & \text{if } \Psi^{-1}(U_{i:N}) > 0 \text{ or equivalently if } 1/2 < U_{i:N} < 1 \\ -1, & \text{if } \Psi^{-1}(U_{i:N}) < 0 \text{ or equivalently if } 0 < U_{i:N} < 1/2. \end{cases}$$

It follows that

$$\begin{aligned} E\left[\frac{-\psi'(\Psi^{-1}(U_{i:N}))}{\psi(\Psi^{-1}(U_{i:N}))}\right] &= \Pr(U_{i:N} > 1/2) - \Pr(U_{i:N} < 1/2) \\ &= 2\Pr(U_{i:N} > 1/2) - 1 \\ &= 2\Pr(B < i-1) - 1, \end{aligned}$$

which are the scores reported in Section 3.2.

For the location-scale t_3 distribution $\psi(u) = (6\sqrt{3}/\pi(3+u^2)^2)$, and therefore $(-\psi'(u)/\psi(u)) = (4u/3+u^2)$. Hence, the optimal scores are $a_O(i) = 4E[(t_{3,U_{i:N}})/3 + t^2_{t_{3,U_{i:N}}}]$. The multiplier of 4 can be discarded since doing so merely scales down the critical value of the test by the same factor of 4. While the expectation could be evaluated numerically, we instead use the asymptotically equivalent scores $((t_{3,i/(N+1)})/3 + t^2_{3,i/(N+1)})$, obtained by replacing $U_{i:N}$ by its expectation.

Sequential Analysis
Design Methods and Applications

ISSN: 0747-4946 (Print) 1532-4176 (Online) Journal homepage: https://www.tandfonline.com/loi/lsqa20

Designing one-sided group sequential clinical trials to detect a mixture alternative

Hua Peng, Daniel R. Jeske, Ashis SenGupta & Weixin Yao

To cite this article: Hua Peng, Daniel R. Jeske, Ashis SenGupta & Weixin Yao (2018) Designing one-sided group sequential clinical trials to detect a mixture alternative, Sequential Analysis, 37:2, 268-291, DOI: 10.1080/07474946.2018.1466545

To link to this article: https://doi.org/10.1080/07474946.2018.1466545

SEQUENTIAL ANALYSIS
2018, VOL. 37, NO. 2, 268–291
https://doi.org/10.1080/07474946.2018.1466545

Taylor & Francis
Taylor & Francis Group

Designing one-sided group sequential clinical trials to detect a mixture alternative

Hua Peng[a], Daniel R. Jeske[a], Ashis SenGupta[b] and Weixin Yao[a]

[a]Department of Statistics, University of California Riverside, Riverside, California, USA; [b]Applied Statistics Unit, Indian Statistical Institute, Kolkata, India

ABSTRACT

We consider the construction of one-sided group sequential designs where the stopping rule includes boundaries for early stopping to accept for futility and to reject for efficacy. The traditional assumption that all patients have the same likelihood of benefiting from the treatment is sometimes unrealistic and can underestimate the required sample size. This motivates us to power the design for an alternative where the treatment group observations come from a mixture of normal distributions. For the proposed setting, we use standardized test statistics based on sample means, and the test turns out to be an L-optimal similar test. Stopping boundaries and arm size for the design are determined by Type I and Type II error spending equations. We demonstrate the need for larger arm sizes when trying to detect a mixture alternative compared to trying to detect a pure shift alternative. The unknown variance case is discussed. With the mixture model, we discuss a more general definition of treatment effect. The maximum likelihood estimator for the treatment effect is discussed.

ARTICLE HISTORY
Received 6 December 2016
Revised 14 April 2018
Accepted 15 April 2018

KEYWORDS
Error spending approach; group sequential design; mixture model; power; sample size; treatment effect

SUBJECT CLASSIFICATIONS
62L05; 62H10; 62F10

1. Introduction

When early termination is desired for either a small or large treatment difference, a group sequential design with both early stopping to reject the null hypothesis for efficacy and early stopping to accept the null hypothesis for futility can be constructed. Numerous techniques for formulating this design have been developed. Popular methods are fixed boundary shape methods and error spending methods. The fixed boundary shape (i.e., boundaries are parameterized by one or two constants) methods derive boundaries with specified boundary shapes, and these methods include Haybittle-Peto (O'Brien and Fleming, 1979; Peto et al., 1976; Pocock, 1977; Wang and Tsiatis, 1987). The error spending method uses an error spending function to specify the errors at each stage, and the stopping boundaries are determined by Type I or Type II error equations. Lan and DeMets (1983) proposed a group sequential design for early stopping with rejection boundaries, and the design is based on a Type I error spending

CONTACT Daniel R. Jeske daniel.jeske@ucr.edu Department of Statistics, University of California, Riverside, 900 University Avenue, 1340 Olmsted Hall, Riverside, CA 92521 USA.
Recommended by S. Chattopadhyay
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lsqa

function. Chang et al. (1998) extended the method of Lan and DeMets to early stopping with rejection and acceptance boundaries and proposed a group sequential design using both Type I and Type II error spending functions. Several software applications are available for designing a group sequential design. Zhu et al. (2011) summarized four popular software applications: EAST v5.2 (Cytel Inc., 2010), ADDPLAN v5.0, the gsDesign package v2.3 in R (Anderson, 2009), and the SEQDESIGN and SEQTEST procedures in SAS v9.2 (SAS Institute Inc., 2009).

For the conventional group sequential design discussed above, it is assumed that both the control and treatment group responses come from normal distributions that potentially have different means. We call this setting a pure shift setting in later sections. As discussed in *Paving the Way for Personalized Medicine* by the Food and Drug Administration (2013), the actual safety and effectiveness of a treatment or product may vary from one individual to the next. Reasons for different results include genetic and environmental factors, as well as the interaction of these factors. For some drug therapies, we can anticipate that only a percentage of treated patients will see an effect. A study conducted by Spear et al. (2001) showed that the response rates of patients to medications from different therapeutic classes varied from 80% (analgesics) to 25% (oncology). This motivated us to set up a group sequential design where the responses from the treatment group patients potentially follow a mixture of normal distributions. Similar to the conventional group sequential design, we test the null hypothesis that there is no difference between the distribution of responses in the control and treatment group. However, power is set through the specification of two parameters, namely, a mixing proportion parameter and mean shift parameter. We refer to this as a mixture alternative in later sections. Our examples will illustrate the need for larger arm sizes when trying to detect a mixture alternative compared to trying to detect a pure shift alternative.

No literature is available for designing group sequential designs to detect a mixture alternative. A related idea is the evaluation of a treatment effect in randomized clinical trials using a genomic biomarker by Wang et al. (2007). Our work and Wang et al.'s paper both consider that a subset of treatment group patients might have a higher likelihood of benefiting from a treatment. However, we differ from Wang et al.'s work in the following aspects: (1) Wang et al.'s test utilizes known subgroup information, whereas the subgroup information is unknown in our setting; (2) Wang et al.'s work considers two test statistics (one using biomarker positive patients and one using all the patients), whereas our work uses only one test statistic; and (3) Wang et al.'s work analyzes the treatment effect for the biomarker-positive subgroup and the overall population simultaneously, whereas our work focuses on the overall treatment effect.

The rest of the article is organized as follows. In Section 2, we introduce the fixed sample test when the treatment group responses potentially follow a mixture of normal distributions. In Section 3, we describe the proposed one-sided group sequential designs to detect a mixture alternative. In Section 4, we discuss implementation details and practical concerns related to the proposed group sequential designs, which includes accounting for the unknown variance parameter and estimating the treatment effect. We conclude in Section 6 with a brief summary and conclusion.

2. Fixed sample test with the mixture alternative

Jeske and Yao (2017) discussed the sample size determination and treatment effect estimation if a subpopulation has a higher likelihood of benefiting from a treatment. Their setting was a fixed sample size and a nonparametric test. This article focuses on a group sequential context and normal theory inference. Let X_{Ci} and X_{Ti} denote the response of subjects from the control and treatment groups, respectively. Suppose that X_{Ci} are independently identically distributed as $F \sim N(\mu_0,\ \sigma^2)$, and X_{Ti} comes from a mixture of normal distributions $G \sim (1-\theta)N(\mu_0,\ \sigma^2) + \theta N(\mu_0+\delta,\ \sigma^2)$, for $i = 1, \ldots, m$, with known σ^2. The whole parameter space is $\theta \in [0,1],\ \delta \geq 0,\ \mu_0 \in R$. We consider the null hypothesis of $H_0 : F = G$ against the one-sided alternative that G is a mixture of F and a shift of F.

Define the standardized test statistic Z as

$$Z = \frac{1}{\sqrt{2m\sigma^2}}\left(\sum_{i=1}^{m} X_{Ti} - \sum_{i=1}^{m} X_{Ci}\right) = \sqrt{\frac{m}{2\sigma^2}}(\overline{X}_T - \overline{X}_C). \tag{2.1}$$

For testing H_0, Z has the motivation of being L-optimal against a one-sided alternative. A detailed discussion about the L-optimal similar property is presented in Appendix A.

Under the null hypothesis, Z has standard normal distribution. However, the distribution of Z under the mixture alternative is not straightforward. When m is large enough, the distribution is approximately

$$Z \dot{\sim}\ N\left(\theta\delta\sqrt{\frac{m}{2\sigma^2}},\ 1 + \frac{\theta(1-\theta)\delta^2}{2\sigma^2}\right). \tag{2.2}$$

For a Type I error probability α, we reject H_0 if $Z > Z_\alpha$, where Z_α is the upper α quantile for standard normal distribution. In order to have power $1-\beta$ at a specified value of $(\theta,\ \delta)$, the required arm size m is

$$m \approx \frac{\left(\sqrt{2}Z_\alpha + \sqrt{2 + \theta(1-\theta)(\delta/\sigma)^2}Z_\beta\right)^2}{\theta^2(\delta/\sigma)^2}. \tag{2.3}$$

In practice, we round m upward to obtain an integer number.

Table 1 shows the required arm size for some choices of θ and δ/σ based on (2.3). The numbers that are in parentheses are the nearly exact required arm sizes based on Monte Carlo simulation as follows. Starting with the required arm size for $\theta = 1$ (pure shift setting), the power of the test was evaluated based on 100,000 simulated data sets for a sequence of increasing (by one) arm sizes until the power became greater than 0.8. The required arm sizes based on the approximate formula (2.3) and the Monte Carlo simulation are almost same, which demonstrates that the approximation in (2.2) works well in a variety of practical settings.

3. Group sequential design with the mixture alternative

3.1. Mixture setting

In a group sequential design with a maximum of K stages, the data are analyzed after every group of $2m$ patients has been accrued. For the analysis at stage k, the test statistic

Table 1. Arm size for a 5% level fixed sample one-sided test based on Z that will achieve 80% power for the mixture alternative.[a]

	Treatment effect shift size, δ/σ			
θp	0.25	0.50	0.75	1
0.5	794 (792)	200 (200)	90 (91)	52 (52)
0.6	551 (551)	139 (139)	63 (63)	36 (36)
0.7	405 (406)	102 (102)	46 (46)	27 (27)
0.8	310 (310)	78 (79)	35 (35)	20 (20)
0.9	245 (244)	62 (62)	28 (28)	16 (16)
1	198 (99)	50 (50)	22 (22)	13 (13)

[a]Numbers in parentheses are arm sizes for a 5% test based on Monte Carlo simulation (simulation size 100,000).

is computed as

$$Z_k = \frac{1}{\sqrt{2mk\sigma^2}}\left(\sum_{i=1}^{mk} X_{Ti} - \sum_{i=1}^{mk} X_{Ci}\right). \tag{3.1}$$

When m is large, the distribution for the sequence of test statistics $\{Z_1, ..., Z_K\}$ can be approximated by a multivariate normal distribution,

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \cdots \\ Z_K \end{pmatrix} \dot{\sim}\ MVN\left(\begin{pmatrix} \theta\delta\sqrt{m/(2\sigma^2)} \\ \theta\delta\sqrt{2m/(2\sigma^2)} \\ \cdots \\ \theta\delta\sqrt{Km/(2\sigma^2)} \end{pmatrix},\ \left(1+\frac{\theta(1-\theta)\delta^2}{2\sigma^2}\right)\begin{pmatrix} 1 & \sqrt{1/2}\ldots & \sqrt{1/K} \\ \sqrt{1/2} & 1\ldots & \sqrt{2/K} \\ & \cdots & \\ \sqrt{1/K} & \sqrt{2/K}\ldots & 1 \end{pmatrix}\right). \tag{3.2}$$

A general one-sided group sequential test is defined by stopping boundaries (a_k, r_k) with $a_k < r_k$ for $k = 1, ..., K-1$ and $a_K = r_K = u$. Figure 1 shows an illustrative picture of the stopping boundaries, which in general take the form:

After stage $k = 1, ..., K-1$

If $Z_k \geq r_k$ stop, reject H_0
If $Z_k \leq a_k$ stop, accept H_0
Otherwise, continue to stage $k+1$.

After stage K

If $Z_K \geq u$ stop, reject H_0
If $Z_K < u$ stop, accept H_0.

The final boundaries coincide, $a_K = r_K = u$, to ensure that the test terminates at stage K.

For the error spending function approach, the stopping boundaries are determined through error spending functions. Error spending functions $f(t)$ and $g(t)$ are defined for Type I and Type II errors, respectively, which are nondecreasing and satisfy $f(0) = 0,\ g(0) = 0$ and $f(t) = \alpha,\ g(t) = \beta$ for $t \geq 1$. Chang et al. (1998) used conservative error spending functions introduced by Hwang et al. (1990). We consider simple power family functions suggested by Jennison and Turnbull (1999), of the form

$$f(t) = \min\{\alpha,\ \alpha t^{\rho}\} \text{ and } g(t) = \min\{\beta,\ \beta t^{\rho}\},\ t \in [0,\ 1]$$

for a chosen value of $\rho > 0$. The choice of ρ controls how much Type I error and Type II error probability is spent at each stage. For equal error spending at each stage, we

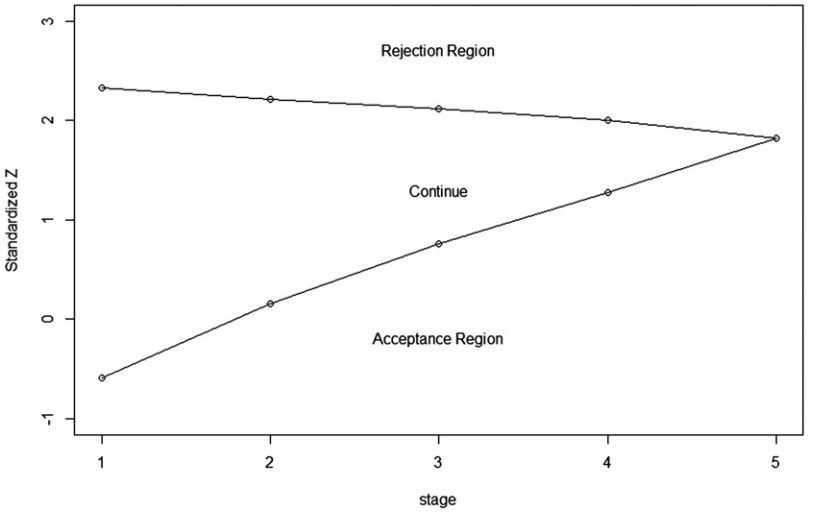

Figure 1. General picture for one-sided group sequential design stopping boundaries with $K = 5$.

take $\rho = 1$. A relatively large ρ value is recommended if practitioners want to be stingy about Type I error and Type II error at early stages. For the case of equal increments of sample size, t at stage k is $t_k = k/K$.

The stopping boundaries and arm size are determined using numerical methods by solving the system of $2K$ equations shown in equation (3.3). Starting with a value of m, the first $2K-2$ equations are solved to obtain $(a_k, r_k), k = 1, 2, ..., K-1$. Then the last two equation are checked to see whether they yield the same solution for u. If the solutions for u are not the same, we increase m and try again. For the proposed mixture setting, the power is designed at a mixture alternative with specification of the mixing proportion θ and mean shift parameter δ, compared to the pure shift setting where the power is designed for a specified value of δ. This difference will be reflected in Type II error probability equations.

$$
\begin{aligned}
P_{F=G}\{Z_1 \geq r_1\} &= \pi_{1,1} \\
P_{\theta,\delta}\{Z_1 \leq a_1\} &= \pi_{2,1} \\
P_{F=G}\{a_1<Z_1<r_1, ..., a_{k-1}<Z_{k-1}<r_{k-1},\ Z_k \geq r_k\} &= \pi_{1,k},\ \ k = 2, ..., K-1 \\
P_{\theta,\delta}\{a_1<Z_1<r_1, ..., a_{k-1}<Z_{k-1}<r_{k-1},\ Z_k \leq a_k\} &= \pi_{2,k},\ \ k = 2, ..., K-1 \\
P_{F=G}\{a_1<Z_1<r_1, ..., a_{K-1}<Z_{K-1}<r_{K-1},\ Z_K \geq u\} &= \pi_{1,K} \\
P_{\theta,\delta}\{a_1<Z_1<r_1, ..., a_{K-1}<Z_{K-1}<r_{K-1},\ Z_K<u\} &= \pi_{2,K}.
\end{aligned} \tag{3.3}
$$

The $\pi_{1,k}$ and $\pi_{2,k}$ values in equation (3.3) are the Type I and Type II error probabilities, respectively, spent at stage k,

$$
\begin{aligned}
\pi_{1,1} &= f(t_1) \\
\pi_{1,k} &= f(t_k) - f(t_{k-1}) \ \text{ for } \ k = 2, 3, ..., K \\
\pi_{2,1} &= g(t_1) \\
\pi_{2,k} &= g(t_k) - g(t_{k-1}) \ \text{ for } \ k = 2, 3, ..., K.
\end{aligned} \tag{3.4}
$$

Table 2. Run time for $\alpha = 0.05,\ \beta = 0.2,\ \sigma^2 = 1,\ \delta = 0.5, \theta = 0.7,\ \rho = 2,$ and specified $(\theta,\ \delta) = (0.7, 0.5)$.

K	Runtime (sec)
3	25.24
5	59.96
10	141.69

Table 3. Constants $R(K, \alpha, \beta, \rho)$ for one-sided test with Type I error rate 1%, 5%, and 10%.

		$1-\beta = 0.8$			$1-\beta = 0.9$			$1-\beta = 0.95$		
αp	K	$\rho = 1$	$\rho = 2$	$\rho = 3$	$\rho = 1$	$\rho = 2$	$\rho = 3$	$\rho = 1$	$\rho = 2$	$\rho = 3$
0.01	2	1.135	1.043	1.015	1.132	1.043	1.016	1.129	1.043	1.016
	3	1.191	1.070	1.030	1.188	1.071	1.032	1.183	1.072	1.032
	4	1.223	1.087	1.041	1.219	1.089	1.043	1.213	1.089	1.044
	5	1.244	1.098	1.049	1.238	1.100	1.051	1.232	1.100	1.052
0.05	2	1.145	1.043	1.014	1.143	1.045	1.015	1.139	1.045	1.016
	3	1.207	1.070	1.028	1.203	1.072	1.030	1.198	1.073	1.031
	4	1.242	1.088	1.038	1.237	1.089	1.040	1.230	1.090	1.042
	5	1.264	1.099	1.046	1.258	1.101	1.048	1.251	1.102	1.051
0.1	2	1.148	1.043	1.013	1.146	1.044	1.014	1.143	1.045	1.015
	3	1.212	1.069	1.026	1.209	1.072	1.029	1.203	1.072	1.030
	4	1.248	1.086	1.036	1.243	1.089	1.039	1.237	1.089	1.041
	5	1.271	1.097	1.043	1.266	1.100	1.046	1.258	1.101	1.049

For the pure shift setting, stopping boundaries and arm size can be determined using existing software applications. For the proposed mixture setting, an R program with user documentation is provided in Appendix B. Generalization of the subdensity computational approach to our mixture setting is described in Appendix C. Table 2 illustrates the efficiency of the efficiency of the computational approach.

3.2. Properties of the proposed group sequential designs

From equation (3.3), it follows that the maximum arm size mK depends on the specified $\alpha, \beta, \theta,\ \rho,\ K,$ and δ/σ. Table D.1 in Appendix D shows the ratio of the maximum arm size to the arm size for the fixed sample test (2.3) for the case $\alpha = 0.05,\ \beta = 0.2,$ and $\rho = 2$. We can see that there is very little dependence on the specified θ or δ/σ, although dependence on K is evident. Similar results are obtained for different choices of $\alpha,\ \beta,$ and ρ. As it is for the pure shift setting, it is useful to tabulate the ratio $R(K, \alpha, \beta, \rho)$ of the maximum arm size for the group sequential test to the arm size of the fixed sample test. Values of $R(K, \alpha, \beta, \rho)$ are listed in Table 3 for $K \in \{2, 3, 4, 5\}$, $\alpha \in \{0.01,\ 0.05,\ 0.1\},\ 1-\beta \in \{0.8,\ 0.85,\ 0.9\},$ and $\rho \in \{1,\ 2,\ 3\}$.

Table 3 implies that the ratio of maximum arm size for the group sequential test to the arm size for the fixed sample test is approximately the same for the pure shift and mixture settings. It follows that because the fixed sample test arm size in the mixture setting is larger than that in the pure shift setting, the same will be true for the arm size of the group sequential test.

Appendix D discusses the fact that when the specified values of θ and δ/σ are in practical ranges, they have a small impact on the stopping boundaries. Thus, the

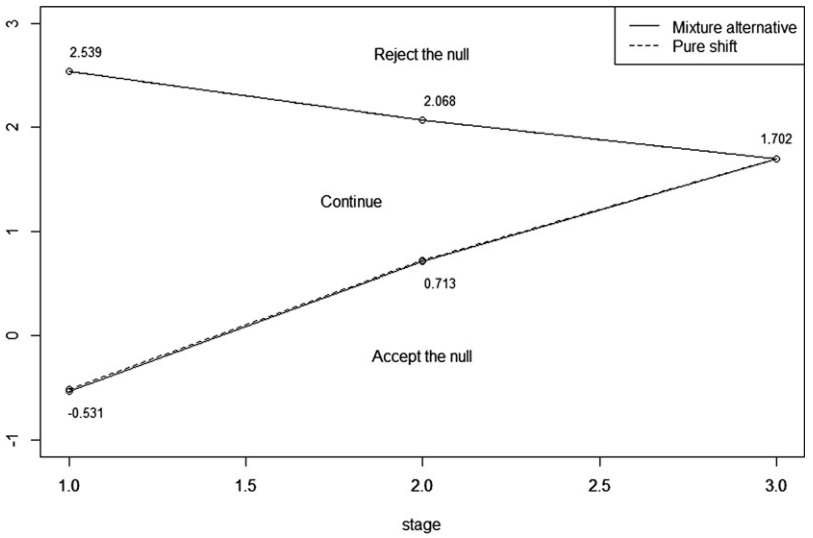

Figure 2. Stopping boundaries for $K = 3,\ \alpha = 0.05,\ \beta = 0.2, \sigma^2 = 1, \delta = 0.5, \theta = 0.7,\ \rho = 2.$

stopping boundaries for the pure shift setting could serve as good approximations for the stopping boundaries in the mixture setting.

3.3. Example illustration

Consider a group sequential design with a maximum number of analyses $K = 3$, Type I error rate 5%, Type II error rate 20%, variance $\sigma^2 = 1$, and $\rho = 2$ for both Type I and Type II spending functions. Assume that we wish to detect the mixture that has $\theta = 0.7$ and $\delta = 0.5$.

Using the R code in Appendix B, it is determined that arm size per stage is $m = 37$, which is significantly larger than the arm size $m = 18$ that is needed to detect a pure shift ($\theta = 1$). The solid lines and numbers listed in Figure 2 illustrate the stopping boundaries. In addition, the stopping boundaries for the pure shift setting are plotted in Figure 2. The insignificant difference between the stopping boundaries for the mixture setting and pure shift setting illustrate adequacy of the previously discussed (Section 3.2) near-equivalence of the two sets of stopping boundaries.

Table 4 compares the target and achieved Type I and Type II error probabilities at each stage. The achieved Type I and Type II error probabilities are obtained through simulation. The simulated error probabilities are obtained via 100,000 sample paths with stopping boundaries and m stated above. As shown in Table 4, the target and achieved Type I and Type II error probabilities are very close.

4. Implementation details and practical concerns

4.1. Unknown σ^2

In practice, σ^2 is unknown and an estimate needs to be used instead. Jennison and Turnbull (1999) discussed the group sequential t-test for a pure shift setting. Following

Table 4. Target and achieved Type I and Type II error at each stage for $K = 3,\ \alpha = 0.05,\ \beta = 0.2, \sigma^2 = 1, \delta = 0.5, \theta = 0.7, \rho = 2$.

	Type I error $\pi_{1,k}$		Type II error $\pi_{2,k}$	
	Target	Simulated	Target	Simulated
$k = 1$	0.0056	0.0053	0.0222	0.0227
$k = 2$	0.0167	0.0168	0.0667	0.0687
$k = 3$	0.0278	0.0282	0.1111	0.1053
Overall	0.0500	0.0503	0.2000	0.1968

that, we introduce the group sequential one-sided t-test for the mixture setting, which is formulated by replacing the unknown σ^2 with an estimator s_k^2 to get

$$T_k = \frac{1}{\sqrt{2mks_k^2}}\left(\sum_{i=1}^{mk} X_{Ti} - \sum_{i=1}^{mk} X_{Ci}\right). \tag{4.1}$$

Our studies suggest using

$$s_k^2 = \frac{1}{mk-1}\sum_{i=1}^{mk}\left(X_{Ci} - \overline{X}_C^k\right)^2. \tag{4.2}$$

Because the observed information depends on σ^2, it will not be possible to guarantee both Type I and Type II error probabilities using prespecified arm sizes. Following Jennison and Turnbull (1999), the priority is to maintain the Type I error rate close to its target value. The starting point is the stopping rule with critical values a_k and r_k. When σ^2 is known, the rejection region at stage k is $Z_k \geq r_k$, and the marginal rejection probability is $1 - \Phi(r_k)$. In order to keep the same marginal rejection probability at stage k, the test static T_k should use a rejection boundary of $t_{mk-1,1-\Phi(r_k)}$. Making this change at each stage gives the stopping rule

After stage $k = 1, \ldots, K-1$

If $T_k \geq t_{mk-1,1-\Phi(r_k)}$ stop, reject H_0
If $T_k \leq t_{mk-1,1-\Phi(a_k)}$ stop, accept H_0
Otherwise, continue to stage $k+1$.

After stage K

If $T_K \geq t_{mk-1,1-\Phi(u)}$ stop, reject H_0
If $T_K \leq t_{mk-1,1-\Phi(u)}$ stop, accept H_0.

Similar to the pure shift setting, our empirical results show that this group sequential t-test can maintain the Type I error pretty accurately, with only a slight erosion in power.

4.2. Treatment effect estimation

In the mixture setting, the treatment effect is thought of as the pair (θ, δ). We propose to use the maximum likelihood estimator (MLE) for (θ, δ). In the pure shift setting, it is known that the post termination MLE of treatment effect is biased. Simulation results shown in Figure 3 indicate that the MLE for (θ, δ) in the mixture setting is also biased. In order to reduce the bias of MLE, we considered using bootstrap bias correction (Y. Wang and Leung, 1997). At termination of the trial, we have maximum likelihood

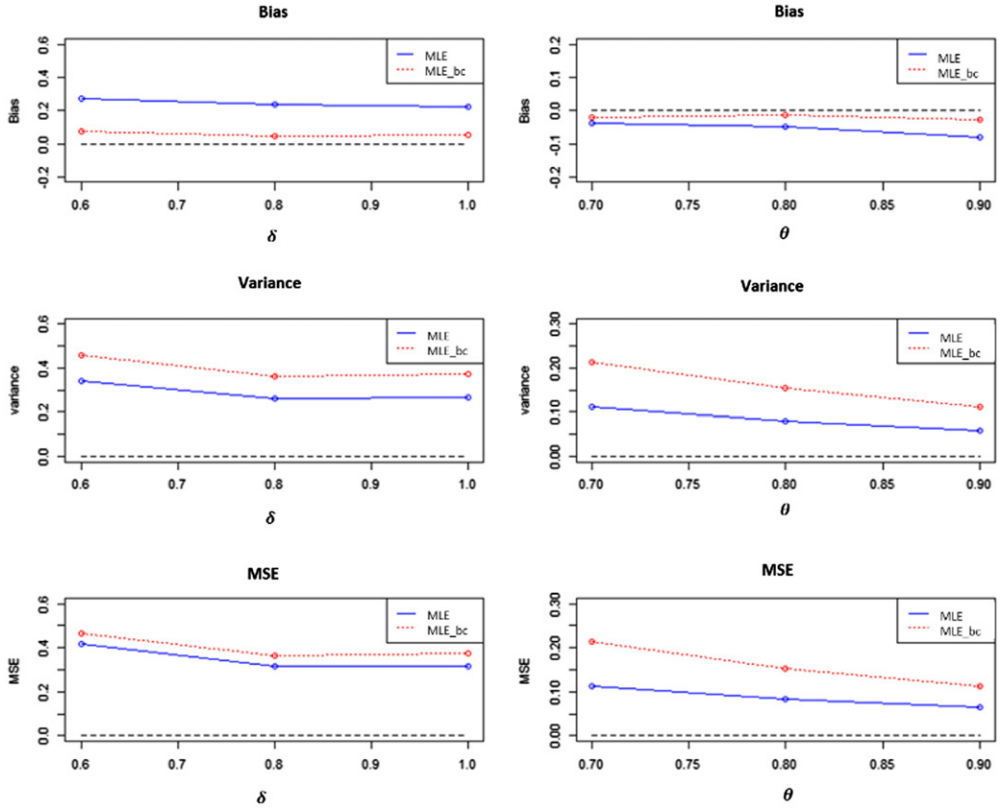

Figure 3. Bias, variance, and MSE of MLE and bootstrap bias-corrected MLE of (θ, δ). $K = 3, \alpha = 0.05,\ \beta = 0.2,\ \sigma^2 = 1,\ \rho = 2,\ \mu_0 = 0$, specified values of $(\theta, \delta) = (0.8,\ 0.8)$, $B = 1,000$, and 1,000 sample paths.

estimators $\hat{\mu_0}, \hat{\sigma^2},\ \hat{\theta},$ and $\hat{\delta}$ obtained from the pooled control and treatment group responses. The bootstrap bias reduction algorithm is outlined below:

1. Generate B bootstrap group sequential sample paths via simulation using $\hat{\mu_0}, \hat{\sigma^2}, \hat{\theta},$ and $\hat{\delta}$. Each sample path will yield a post termination estimate of θ and δ, which are denoted as $\hat{\theta_j^*}$ and $\hat{\delta^*},\ j = 1, \ldots, B$.
2. The bootstrap bias estimates of θ and δ are

$$\frac{\sum_{j=1}^{B} \hat{\theta_j^*}}{B} - \hat{\theta},\ \ \frac{\sum_{j=1}^{B} \hat{\delta_j^*}}{B} - \hat{\delta}; \tag{4.3}$$

3. The bootstrap bias-corrected MLE estimates of θ and δ are

$$\hat{\theta}_{bc} = 2\hat{\theta} - \frac{\sum_{j=1}^{B} \hat{\theta_j^*}}{B},\ \ \hat{\delta}_{bc} = 2\hat{\delta} - \frac{\sum_{j=1}^{B} \hat{\delta_j^*}}{B}. \tag{4.4}$$

Figure 3 compares the bias, variance, and mean squared error (MSE) of the MLE and the bias-corrected MLE of (θ, δ). In this example, we consider $K = 3$,

$\alpha = 0.05,\ \beta = 0.2,\ \sigma^2 = 1,\ \rho = 2,\ \mu_0 = 0,$ and specified values of $(\theta, \delta) = (0.8,\ 0.8)$. It can be seen that the bias-corrected MLE has significantly less bias. However, the increase in variance results in higher MSE. Similar results are observed for various specified values of (θ, δ). Therefore, we suggest using the MLE instead of the bias-corrected MLE.

5. Conclusion and summary

The traditional group sequential design is powered under an alternative where the treatment effect is a pure shift in the mean. In this article, we argue that this alternative is often unrealistic. Instead, we consider an alternative where only a percentage of treated patients have a higher likelihood of benefiting from a treatment and propose the use of a mixture alternative as a way to capture this realism. The required arm size to detect a specified mean shift of δ for a specified percentage of patients is larger than what is required to detect the mean shift of δ for all treated patients. The stopping boundaries depend on the specified θ and δ/σ, but the dependence is weak. However, the arm size depends significantly on both the specified θ and δ/σ. We also have discussed adjustment for the unknown σ^2case and post termination MLE of (θ, δ).

Appendix A: L-optimal similar test

The null hypothesis, $H_0 : F = G$, can be written as the union of two hypotheses:

$$H_{01} :\ \theta = 0 \cup H_{02} :\ \delta = 0.$$

Denote η as $\eta = \theta\delta$. Then $H_{01} \cup H_{02}$ is equivalent to $H_0 : \eta = 0$. Following SenGupta (2007), η is called a pivotal parametric product (P^3) for H_0. To conclude that a test ψ based on the test statistic $Z = \sqrt{\frac{m}{2\sigma^2}}\left(\overline{X}_T - \overline{X}_C\right)$ is L-optimal similar (see definition 3 in SenGupta [2007]; also see SenGupta [1991]) for testing $H_0 : \eta = 0$ against $H_1 : \eta > 0$, we need to check whether ψ is LMP (locally most powerful) similar for testing $H_0 : \delta = 0$ against $H_1 : \delta > 0$ for each known $\theta \in (0, 1)$ and unknown nuisance parameter μ_0.

Spjotvoll (1968) presented the method of derivation of the LMP similar test with some generality. Let L and l denote the likelihood and log-likelihood based on the sample observations. From Gokhale and SenGupta (1986), it is seen that the LMP similar test has critical region ω,

$$\omega :\ \frac{\partial l}{\partial \delta}|_{\delta=0} > c(t), \tag{A.1}$$

where $c(t)$ generically denotes a constant depending on a fixed value $T = t$ of the sufficient statistic for μ_0 under H_0, and it is so determined using the conditional distribution of X given $T = t$ to satisfy the size requirement.

Following the structure of the general nonexponential family and also the arguments as presented by Spjotvoll (1968, p. 773), it is seen that the probability measure for $\delta = 0$ under H_0 and some $\mu_0 = \mu_0{}^*$ can be used as a dominating measure for any $\mu_0{}^*$. The log-likelihood function l at $\mu_0 = \mu_0{}^*$ is

$$l = \sum_{i=1}^{m} \ln\phi(x_{Ci}; \mu_0{}^*,\ \sigma^2) + \sum_{i=1}^{m} \ln\{(1-\theta)\phi(x_{Ti}; \mu_0{}^*,\ \sigma^2) + \theta\phi(x_{Ti}; \mu_0{}^* + \delta, \sigma^2)\}, \tag{A.2}$$

where ϕ is the normal probability density function.

$$\begin{aligned} \frac{\partial l}{\partial \delta}|_{\delta=0} &= \frac{\partial}{\partial \delta}\sum_{i=1}^{m} \ln\{(1-\theta)\phi(x_{Ti}; \mu_0{}^*, \sigma^2) + \theta\phi(x_{Ti}; \mu_0{}^* + \delta, \sigma^2)\}|_{\delta=0} \\ &= \sum_{i=1}^{m} \frac{\frac{\partial}{\partial \delta}\theta\phi(x_{Ti}; \mu_0{}^* + \delta, \sigma^2)}{(1-\theta)\phi(x_{Ti}; \mu_0{}^*, \sigma^2) + \theta\phi(x_{Ti}; \mu_0{}^* + \delta, \sigma^2)}|_{\delta=0} \\ &= \sum_{i=1}^{m} \frac{\theta}{\sigma^2}(x_{Ti} - \mu_0{}^*). \end{aligned} \tag{A.3}$$

Using (A.1), the critical region ω reduces to

$$\omega:\ \sum_i (X_{Ti} - \mu_0{}^*) > c(t)\frac{\sigma^2}{\theta}. \tag{A.4}$$

Under H_0, likelihood function L is given by

$$L = \prod_{i=1}^{m} \phi(x_{Ci}; \mu_0, \sigma^2)\phi(x_{Ti}; \mu_0, \sigma^2) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^{2m}} e^{-\frac{\sum_{i=1}^{m}(x_{Ci}-\mu_0)^2 + (x_{Ti}-\mu_0)^2}{2\sigma^2}}. \tag{A.5}$$

Therefore, the sufficient statistic for μ_0 under H_0 is $T = \frac{1}{2}\left(\overline{X_T} + \overline{X_C}\right)$.
Now rewriting the left-hand side of (A.4) by introducing T, we have

$$\sum_i (X_{Ti} - \mu_0{}^*) = \sum_i (X_{Ti} - T) + m(T - \mu_0{}^*) = 0.5m(\overline{X_T} - \overline{X_C}) + m(T - \mu_0{}^*). \tag{A.6}$$

Then (A.4) can be rewritten as

$$\omega:\ \ D = \overline{X_T} - \overline{X_C} > 2c\left\{(t)\frac{\sigma^2}{m\theta} - (t - \mu_0{}^*)\right\}. \tag{A.7}$$

Further, under H_0, D has a distribution free of μ_0 and thus from Basu's theorem it is independent of T. Hence, ω reduces equivalently to

$$\omega : D = \overline{X_T} - \overline{X_C} > C, \tag{A.8}$$

where C is merely a constant determined to unconditionally satisfy the size condition; that is,

$$P_{H_0}\left(\left(\overline{X_T} - \overline{X_C}\right) > C\right) = \alpha, \text{or}$$

$$P_{H_0}\left(Z=\sqrt{\frac{m}{2\sigma^2}}(\overline{X_T}-\overline{X_C})>Z_\alpha\right)=\alpha. \tag{A.9}$$

Because the distribution of D depends only on δ and D does not depend on ${\mu_0}^*$, the test based on D in (A.8) is LMP similar.

Define the test ψ as

$$\psi=\begin{cases}1, & \text{if } Z=\sqrt{\frac{m}{2\sigma^2}}(\overline{X_T}-\overline{X_C})>Z_\alpha\\ 0, & \text{otherwise}\end{cases}. \tag{A.10}$$

Therefore, for each known value of $\theta\in(0,1)$, the test ψ is an LMP similar test (with respect to the nuisance parameter μ_0) for testing $H_0:\delta=0$ against $H_1:\delta>0$. Hence, by definition, the test ψ is L-optimal similar for testing $H_0:\eta=0$ against $H_1:\eta>0$.

Appendix B: The R code for the stopping boundaries

B.1. Manual

This is the guide to function **GSDMix()** written in R to implement the group sequential design to detect a mixture alternative.

GSDMix *Determining stopping boundaries and arm size per stage for a group sequential design to detect a mixture alternative*

Description

GSDMix() is used to determine stopping boundaries and trial size required for a group sequential design in mixture setting.

Usage

GSDMix(K,alpha,beta,rho,theta,effectsize)

Arguments

K	maximum number of analyses planned, including interim and final.
alpha	Type I error rate for the design.
beta	Type II error rate for the design.
rho	exponent in error spending functions.
theta	the specified mixing proportion parameter in the mixture distribution.
effectsize	the specified standardized mean shift parameter, which is δ/σ.

Details

GSDMix() is used for one-sided group sequential design in the mixture setting. The stopping boundaries are early stopping to reject for efficacy or early stopping to accept for futility. The input *rho* is the power exponent in Type I and Type II error spending functions. The group sequential design in mixture setting will be reduced to group sequential design in pure shift setting if choosing *theta* $= 1$.

Value

lower boundaries of acceptance for the group sequential design with mixture alternative.

upper boundaries of rejection for the group sequential design with mixture alternative.

arm_size arm size per stage.

Examples

>test<-GSDMix(3,0.05,0.2,2,0.8,0.5)

> test

$lower

[1] -0.5332111 0.7047889 1.7031848

$upper

[1] 2.539185 2.068185 1.703185

$arm_size

[1] 28

B.2. R code

```
GSDMix=function(K,alpha,beta,rho,theta,effectsize){

    # function powerdiff is defined to evaluate difference of power with m and
target power
      powerdiff=function(m){
      pi1=vector("numeric",K)
      pi2=vector("numeric",K)
      b=vector("numeric",2*K+1)
      a=vector("numeric",K)
      r=vector("numeric",K)
      t=vector("numeric",K)
      In=vector("numeric",K)

      for (i in 1:K){
        t[i]=i/K
      }

      pi1[1]=alpha*(t[1])^rho
      pi2[1]=beta*(t[1])^rho

      for (i in 2:K){
        pi1[i]=alpha*(t[i])^rho-alpha*(t[i-1])^rho
        pi2[i]=beta*(t[i])^rho-beta*(t[i-1])^rho
      }

      ratiomean = theta
      ratiovariance =1+theta*(1-theta)*effectsize^2/2
```

```
      for (i in 1:K){
        In[i]=m*i/2
      }

      r[1]=qnorm(1-pi1[1])
      a[1]=ratiomean*effectsize*sqrt(In[1])+sqrt(ratiovariance)*qnorm+(p-
i2[1])

      n = 320
      h1=matrix(numeric(0),K,n+1)
      h2=matrix(numeric(0),K,n+1)
      z=matrix(numeric(0),K,n+1)
      w=vector("numeric",n+1)
      f=vector("numeric",K)
      g=vector("numeric",K)

      g[1]=pi2[1]

      for (i in 2:K){
        ### weight for Trapezoidal rule
        w[1]=(r[i-1]-a[i-1])/(2*n)
        w[n+1]=(r[i-1]-a[i-1])/(2*n)
        for (j in 2:n) {
          w[j]=(r[i-1]-a[i-1])/n
        }
        h=(r[i-1]-a[i-1])/n

        for (j in 1:(n+1)){
          z[i,j]=a[i-1]+(j-1)*h
        }
        if(i==2){
          for (j in 1:(n+1)) {
            h1[i,j]=w[j]*dnorm(z[i,j])
            h2[i,j]=w[j]*dnorm((z[i,j]-ratiomean*effectsize*sqrt(In[i-1]))/
sqrt(ratiovariance))/sqrt(ratiovariance)
          }
        } else{
           for (j in 1:(n+1)) {
             h1[i,j] = 0
             h2[i,j] = 0
             for(l in 1:(n+1)){
               h1[i,j]=h1[i,j]+h1[i-1,l]*w[j]*sqrt(In[i-1]/(In[i-1]-In[i-
2]))*dnorm((z[i,j]*sqrt(In[i-1])-z[i-1,l]*sqrt(In[i-2]))/sqrt(In[i-1]-In[i-2]))
               h2[i,j]=h2[i,j]+h2[i-1,l]*w[j]*sqrt(In[i-1]/((In[i-1]-In[i-
2])*ratiovariance))*
```

```
                +dnorm((z[i,j]*sqrt(In[i-1])-z[i-1,l]*sqrt(In[i-2]))/
sqrt((In[i-1]-In[i-2])*rationvariance)-ratiomean*effectsize*sqrt((In[i-1]-In[i-
2])/rationvariance))
            }
          }
        }
        d=-0.001
        r[i]=r[i-1]
        f[i]=0
        for (j in 1:(n+1)) {
          f[i]=f[i]+h1[i,j]*pnorm((z[i,j]*sqrt(In[i-1])-r[i]*sqrt(In[i]))/
sqrt(In[i]-In[i-1]))
        }

        e1=(f[i]-pi1[i])^2

        r[i]=r[i]+d
        f[i]=0
        for (j in 1:(n+1)) {
          f[i]=f[i]+h1[i,j]*pnorm((z[i,j]*sqrt(In[i-1])-r[i]*sqrt(In[i]))/
sqrt(In[i]-In[i-1]))
        }
        e2=(f[i]-pi1[i])^2

        if (e1<e2) {
          d=-d
        }

        e1=0.9
        e2=1

        while (e1<e2){
          r[i]=r[i]+d
          e2=e1
          f[i]=0
          for (j in 1:(n+1)) {
            f[i]=f[i]+h1[i,j]*pnorm((z[i,j]*sqrt(In[i-1])-r[i]*sqrt(In[i]))/
sqrt(In[i]-In[i-1]))
          }
          e1=(f[i]-pi1[i])^2
        }
        r[i]=r[i]-d
        if(i==K){
          a[i]=r[i]
          g[i]=0
            for (j in 1:(n+1)) {
```

```
            g[i]=g[i]+h2[i,j]*pnorm((-z[i,j]*sqrt(In[i-
1])+a[i]*sqrt(In[i])-ratiomean*effectsize*(In[i]-In[i-1]))/sqrt((In[i]-In[i-
1])*rationvariance))
        }
        pK=g[i]
      } else{
        d=0.001
        a[i]=a[i-1]
          g[i]=0
          for (j in 1:(n+1)) {
            g[i]=g[i]+h2[i,j]*pnorm((-z[i,j]*sqrt(In[i-1])+a[i]*sqrt(In[i])-
ratiomean*effectsize*(In[i]-In[i-1]))/sqrt(In[i]-In[i-1]))
          }

          e1=(g[i]-pi2[i])^2

          a[i]=a[i]+d
          g[i]=0
          for (j in 1:(n+1)) {
            g[i]=g[i]+h2[i,j]*pnorm((-z[i,j]*sqrt(In[i-1])+a[i]*sqrt(In[i])-
ratiomean*effectsize*(In[i]-In[i-1]))/sqrt((In[i]-In[i-1])*rationvariance))
          }
          e2=(g[i]-pi2[i])^2

          if (e1<e2) {
            d=-d
          }

          e1=0.9
          e2=1

          while (e1<e2){
            a[i]=a[i]+d
            e2=e1
            g[i]=0
            for (j in 1:(n+1)) {
              g[i]=g[i]+h2[i,j]*pnorm((-z[i,j]*sqrt(In[i-
1])+a[i]*sqrt(In[i])-ratiomean*effectsize*(In[i]-In[i-1]))/sqrt((In[i]-In[i-
1])*rationvariance))
            }
            e1=(g[i]-pi2[i])^2
          }
          a[i]=a[i]-d
        }
      }
      1-sum(g)-(1-beta)
    }
```

```
    n = 2*(qnorm(1-alpha)+qnorm(1-beta)*sqrt(1+theta*(1-
theta)*effectsize^2/2))^2/(theta*effectsize)^2
    l=max(n/K,2)
    u = n/K*2
    m=uniroot(powerdiff,lower = l,upper = u)$root
    m=ceiling(m)

    ratiomean = theta
    ratiovariance =1+theta*(1-theta)*effectsize^2/2
    pi1=vector("numeric",K)
    pi2=vector("numeric",K)
    b=vector("numeric",2*K+1)
    a=vector("numeric",K)
    r=vector("numeric",K)
    t=vector("numeric",K)
    In=vector("numeric",K)

    for (i in 1:K){
      t[i]=i/K
    }

    pi1[1]=alpha*(t[1])^rho
    pi2[1]=beta*(t[1])^rho

    for (i in 2:K){
      pi1[i]=alpha*(t[i])^rho-alpha*(t[i-1])^rho
      pi2[i]=beta*(t[i])^rho-beta*(t[i-1])^rho
    }
    for (i in 1:K){
      In[i]=m*i/2
    }

    r[1]=qnorm(1-pi1[1])
    a[1]=ratiomean*effectsize*sqrt(In[1])+sqrt(ratiovariance)*qnorm+(pi2[1])

    n = 320
    h1=matrix(numeric(0),K,n+1)
    h2=matrix(numeric(0),K,n+1)
    z=matrix(numeric(0),K,n+1)
    w=vector("numeric",n+1)
    f=vector("numeric",K)
    g=vector("numeric",K)

    for (i in 2:K){
      ### weight for Trapezoidal rule
      w[1]=(r[i-1]-a[i-1])/(2*n)
```

```
      w[n+1]=(r[i-1]-a[i-1])/(2*n)
      for (j in 2:n) {
        w[j]=(r[i-1]-a[i-1])/n
      }
      h=(r[i-1]-a[i-1])/n

      for (j in 1:(n+1)){
        z[i,j]=a[i-1]+(j-1)*h
      }
      if(i==2){
        for (j in 1:(n+1)) {
          h1[i,j]=w[j]*dnorm(z[i,j])
          h2[i,j]=w[j]*dnorm((z[i,j]-ratiomean*effectsize*sqrt(In[i-1]))/
sqrt(ratiovariance))/sqrt(ratiovariance)
        }
      } else{
        for (j in 1:(n+1)) {
          h1[i,j]=0
          h2[i,j]=0
          for(l in 1:(n+1)){
            h1[i,j]=h1[i,j]+h1[i-1,l]*w[j]*sqrt(In[i-1]/(In[i-1]-In[i-
2]))*dnorm((z[i,j]*sqrt(In[i-1])-z[i-1,l]*sqrt(In[i-2]))/sqrt(In[i-1]-In[i-2]))
            h2[i,j]=h2[i,j]+h2[i-1,l]*w[j]*sqrt(In[i-1]/((In[i-1]-In[i-
2])*ratiovariance))*
              +dnorm((z[i,j]*sqrt(In[i-1])-z[i-1,l]*sqrt(In[i-2]))/sqrt((In[i-1]-
In[i-2])*ratiovariance)-ratiomean*effectsize*sqrt((In[i-1]-In[i-2])/ratiovariance))
          }
        }
      }
      d=-0.001
      r[i]=r[i-1]
      f[i]=0
      for (j in 1:(n+1)) {
        f[i]=f[i]+h1[i,j]*pnorm((z[i,j]*sqrt(In[i-1])-r[i]*sqrt(In[i]))/
sqrt(In[i]-In[i-1]))
      }

      e1=(f[i]-pi1[i])^2

      r[i]=r[i]+d
      f[i]=0
      for (j in 1:(n+1)) {
        f[i]=f[i]+h1[i,j]*pnorm((z[i,j]*sqrt(In[i-1])-r[i]*sqrt(In[i]))/
sqrt(In[i]-In[i-1]))
      }
      e2=(f[i]-pi1[i])^2
```

```
      if (e1<e2) {
        d=-d
      }

      e1 = 0.9
      e2 = 1
      while (e1<e2){
        r[i]=r[i]+d
        e2 = e1
        f[i] = 0
        for (j in 1:(n+1)) {
          f[i]=f[i]+h1[i,j]*pnorm((z[i,j]*sqrt(In[i-1])-r[i]*sqrt(In[i]))/
sqrt(In[i]-In[i-1]))
        }
        e1=(f[i]-pi1[i])^2
      }
      r[i]=r[i]-d

      if(i==K){
        a[i]=r[i]
        g[i] = 0
        for (j in 1:(n+1)) {
          g[i]=g[i]+h2[i,j]*pnorm((-z[i,j]*sqrt(In[i-1])+a[i]*sqrt(In[i])-
ratiomean*effectsize*(In[i]-In[i-1]))/sqrt((In[i]-In[i-1])*rationariance))
        }
        pK = g[i]
      } else{
        d = 0.001
        a[i]=a[i-1]
        g[i] = 0
        for (j in 1:(n+1)) {
          g[i]=g[i]+h2[i,j]*pnorm((-z[i,j]*sqrt(In[i-1])+a[i]*sqrt(In[i])-
ratiomean*effectsize*(In[i]-In[i-1]))/sqrt(In[i]-In[i-1]))
        }

        e1=(g[i]-pi2[i])^2
        a[i]=a[i]+d
        g[i] = 0
        for (j in 1:(n+1)) {
          g[i]=g[i]+h2[i,j]*pnorm((-z[i,j]*sqrt(In[i-1])+a[i]*sqrt(In[i])-
ratiomean*effectsize*(In[i]-In[i-1]))/sqrt((In[i]-In[i-1])*rationariance))
        }
        e2=(g[i]-pi2[i])^2

        if (e1<e2) {
          d=-d
```

```
        }

        e1=0.9
        e2=1

        while (e1<e2){
        a[i]=a[i]+d
        e2=e1
        g[i]=0
        for (j in 1:(n+1)) {
          g[i]=g[i]+h2[i,j]*pnorm((-z[i,j]*sqrt(In[i-1])+a[i]*sqrt(In[i])-
ratiomean*effectsize*(In[i]-In[i-1]))/sqrt((In[i]-In[i-1])*ratiovariance))
        }
        e1=(g[i]-pi2[i])^2
        }
        a[i]=a[i]-d
      }
    }

    for (i in 1:K){
      b[2*i-1]=a[i]
      b[2*i]=r[i]
      b[2*K+1]=m
    }

    return(list(lower=a,upper=r,arm_size=m))
  }
```

Appendix C: Subdensity

The distribution of the sequence of test statics $\{Z_1, ..., Z_K\}$ are approximately multivariate normal with mean and covariance

$$E(Z_k) = \theta\delta\sqrt{\frac{mk}{2\sigma^2}},\ \ k = 1, ..., K$$

$$cov(Z_{k1}, Z_{k2}) = \left(1 + \frac{\theta(1-\theta)\delta^2}{2\sigma^2}\right)\sqrt{k_1/k_2},\ \ 1 \le k_1 \le k_2 \le K. \tag{C.1}$$

This allows use of the recursive formula of Armitage et al. (1969). Specifically, for $k = 2, ..., K$,

$$Z_k\sqrt{\frac{mk}{2\sigma^2}} - Z_{k-1}\sqrt{\frac{m(k-1)}{2\sigma^2}} \ \sim N\left(\theta\delta\frac{m}{2\sigma^2}, \left(1 + \frac{\theta(1-\theta)\delta^2}{2\sigma^2}\right)\frac{m}{2\sigma^2}\right), \tag{C.2}$$

and this distribution is independent of $Z_1, ..., Z_{k-1}$. Defining

$$G_k(z; \theta, \delta) = P_{\theta,\delta}\{a_1 < Z_1 < r_1, ..., a_{k-1} < Z_{k-1} < r_{k-1},\ \ Z_k \ge z\} \tag{C.3}$$

and $g_k(z; \theta, \delta)$ to be the derivative of $G_k(z; \theta, \delta)$ with respect to z, this is called

subdensity of Z_k at stage k. Let $\phi(z)$ denote the standard normal density for $k = 2, ..., K$. Then g_k is given recursively by

$$g_1(z;\theta,\delta) = \phi\left(z - \theta\delta\sqrt{\frac{m}{2\sigma^2}}\right),$$

$$g_k(z;\theta,\delta) = \int_{a_{k-1}}^{r_{k-1}} g_{k-1}(u;\theta,\delta)\frac{\sqrt{\frac{mk}{2\sigma^2}}}{\sqrt{\left(1+\frac{\theta(1-\theta)\delta^2}{2\sigma^2}\right)\frac{m}{2\sigma^2}}}\phi\left(\frac{z\sqrt{\frac{mk}{2\sigma^2}} - u\sqrt{\frac{m(k-1)}{2\sigma^2}} - \theta\delta\frac{m}{2\sigma^2}}{\sqrt{\left(1+\frac{\theta(1-\theta)\delta^2}{2\sigma^2}\right)\frac{m}{2\sigma^2}}}\right)du. \tag{C.4}$$

The stopping boundaries (a_k, r_k), $k = 1, 2, ..., K-1$, and u are calculated in sequence. When $(a_i, r_i), i = 1, 2, ..., k-1$ are known, the above formula can be used to evaluate g_k numerically by a succession of $k-1$ univariate integrals and hence for the specified values of (θ, δ), one can solve

$$\pi_{1,k} = \int_{r_k}^{\infty} g_k(z_k;\theta,\delta|\theta\delta = 0)dz_k$$

$$\pi_{2,k} = \int_{-\infty}^{a_k} g_k(z_k;\theta,\delta)dz_k \tag{C.5}$$

to obtain r_k and a_k for the group sequential design with mixture alternative.

Appendix D: Stopping boundaries and ratio of maximum arm size to arm size for the fixed sample test

By the central limit theorem, the variance–covariance matrix is approximately

$$\left(1+\frac{\theta(1-\theta)\delta^2}{2\sigma^2}\right)\begin{pmatrix} 1 & \sqrt{1/2}... & \sqrt{1/K} \\ \sqrt{1/2} & 1... & \sqrt{2/K} \\ & ... & \\ \sqrt{1/K} & \sqrt{2/K}...1 & \end{pmatrix}. \tag{D.1}$$

Cohen (1988) suggested that 0.2 represents a small effect size, 0.5 represents a medium effect size, and 0.8 represents a large effect size. If we take $\delta/\sigma = 1$ as the largest practical value we might see, then the multiplier in front of the variance covariance matrix satisfies

$$1 \le 1 + \frac{\theta(1-\theta)\delta^2}{2\sigma^2} \le 1 + \frac{0.25}{2} = 1.125. \tag{D.2}$$

As such, the choices of the specified values of θ and δ/σ do not have a big impact for the variance–covariance matrix of the test statistics $\{Z_1, ..., Z_K\}$, and

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \dots \\ Z_K \end{pmatrix} \dot{\sim} \; MVN\left(\begin{pmatrix} \theta\delta\sqrt{m/(2\sigma^2)} \\ \theta\delta\sqrt{2m/(2\sigma^2)} \\ \dots \\ \theta\delta\sqrt{Km/(2\sigma^2)} \end{pmatrix}, 1.125 \begin{pmatrix} 1 & \sqrt{1/2}\dots & \sqrt{1/K} \\ \sqrt{1/2} & 1\dots & \sqrt{2/K} \\ & \dots & \\ \sqrt{1/K} & \sqrt{2/K}\dots & 1 \end{pmatrix} \right). \tag{D.3}$$

Under $H_0: \; F = G$,

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \dots \\ Z_K \end{pmatrix} \sim \; MVN\left(\begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sqrt{1/2}\dots & \sqrt{1/K} \\ \sqrt{1/2} & 1\dots & \sqrt{2/K} \\ & \dots & \\ \sqrt{1/K} & \sqrt{2/K}\dots & 1 \end{pmatrix} \right). \tag{D.4}$$

Under the specified θ and δ, Type II error probabilities become

$$P_{\theta,\delta}\{a_1 < Z_1 < r_1, \dots, \; a_{k-1} < Z_{k-1} < r_{k-1}, \; Z_k \leq \; a_k\}$$

$$\approx P_{\theta,\delta}\left\{ \frac{a_1 - \theta\delta\sqrt{\frac{1}{K}}\sqrt{\frac{mK}{2\sigma^2}}}{\sqrt{1.125}} < Z_1^* < \frac{r_1 - \theta\delta\sqrt{\frac{1}{K}}\sqrt{\frac{mK}{2\sigma^2}}}{\sqrt{1.125}}, \dots, \; Z_k^* \leq \; \frac{a_k - \theta\delta\sqrt{\frac{k}{K}}\sqrt{\frac{mK}{2\sigma^2}}}{\sqrt{1.125}} \right\}, \tag{D.5}$$

where $Z_i^* = \frac{1}{\sqrt{1.125}}\left\{ Z_i - \theta\delta\sqrt{\frac{i}{K}}\sqrt{\frac{mK}{2\sigma^2}} \right\}$ for $i = 1, \dots, k$, and Z_i^* is standard normal at the specified values of θ and δ. Solving for m is equivalent to solving first for $\lambda = \theta\delta\sqrt{\frac{mK}{2\sigma^2}}$ in (D.6) and then computing $m = 2\lambda^2\sigma^2/(\theta^2\delta^2 K)$.

$$P_{H_0}\{Z_1 \geq r_1\} = \pi_{1,1}$$

$$P_{\theta,\delta}\left\{ Z_1^* \leq \frac{a_1 - \lambda\sqrt{1/K}}{\sqrt{1.125}} \right\} = \pi_{2,1}$$

$$P_{H_0}\{a_1 < Z_1 < r_1, \dots, Z_k \geq r_k\} = \pi_{1,k}, \;\; k = 2, \dots, K-1$$

$$P_{\theta,\delta}\left\{ \frac{a_1 - \lambda\sqrt{1/K}}{\sqrt{1.125}} < Z_1^* < \frac{r_1 - \lambda\sqrt{1/K}}{\sqrt{1.125}}, \dots, Z_k^* \leq \frac{a_k - \lambda\sqrt{k/K}}{\sqrt{1.125}} \right\} = \pi_{2,k}, \;\; k = 2, \dots, K-1$$

$$P_{\theta=0}\{a_1 < Z_1 < r_1, \dots, Z_K \geq u\} = \pi_{1,K}$$

$$P_{\theta,\delta}\left\{ \frac{a_1 - \lambda\sqrt{1/K}}{\sqrt{1.125}} < Z_1^* < \frac{r_1 - \lambda\sqrt{1/K}}{\sqrt{1.125}}, \dots, Z_K^* < \frac{u - \lambda}{\sqrt{1.125}} \right\} = \pi_{2,K}. \tag{D.6}$$

Because the distribution of $\{Z_1, \dots, Z_K\}$ under the null hypothesis and $\{Z_1^*, \dots, Z_K^*\}$ under the mixture alternative are both approximately $MVN\left(\begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sqrt{1/2}\dots & \sqrt{1/K} \\ \sqrt{1/2} & 1\dots & \sqrt{2/K} \\ & \dots & \\ \sqrt{1/K} & \sqrt{2/K}\dots & 1 \end{pmatrix} \right)$, the solution to the stopping boundaries and λ only depends on K, α, β, and ρ. Once λ is solved for, we can determine the maximum arm size, $mK = 2\lambda^2\sigma^2/(\theta^2\delta^2)$, which is proportional to $\sigma^2/(\theta^2\delta^2)$. For the fixed

Table D.1. Ratio of the maximum arm size for the group sequential test to the arm size for the fixed sample test for 5% level that will achieve 80% power for the mixture alternative with $\rho = 2$.

θ	K	Treatment effect shift size, δ/σ							
		0.25	0.5	0.75	1	1.5	2	2.5	3
1	2	1.043	1.043	1.043	1.043	1.043	1.043	1.043	1.043
	3	1.070	1.070	1.070	1.070	1.070	1.070	1.070	1.070
	4	1.088	1.088	1.088	1.088	1.088	1.088	1.088	1.088
	5	1.099	1.099	1.099	1.099	1.099	1.099	1.099	1.099
0.9	2	1.043	1.043	1.043	1.043	1.043	1.042	1.040	1.037
	3	1.070	1.070	1.071	1.070	1.069	1.067	1.064	1.060
	4	1.088	1.087	1.087	1.087	1.086	1.084	1.079	1.073
	5	1.099	1.098	1.099	1.098	1.096	1.094	1.089	1.083
0.8	2	1.043	1.044	1.043	1.043	1.042	1.039	1.035	1.029
	3	1.070	1.070	1.070	1.069	1.067	1.063	1.056	1.046
	4	1.087	1.087	1.087	1.086	1.084	1.078	1.069	1.056
	5	1.099	1.099	1.098	1.097	1.094	1.087	1.077	1.062
0.7	2	1.043	1.043	1.043	1.043	1.040	1.036	1.030	1.022
	3	1.071	1.070	1.069	1.069	1.065	1.059	1.049	1.036
	4	1.087	1.087	1.087	1.085	1.081	1.073	1.060	1.043
	5	1.099	1.099	1.098	1.096	1.091	1.082	1.067	1.047
0.6	2	1.043	1.043	1.043	1.042	1.040	1.035	1.028	1.018
	3	1.071	1.070	1.070	1.068	1.065	1.057	1.045	1.029
	4	1.087	1.087	1.086	1.085	1.080	1.070	1.055	1.035
	5	1.099	1.098	1.097	1.096	1.090	1.078	1.060	1.037
0.5	2	1.043	1.043	1.043	1.043	1.040	1.035	1.027	1.017
	3	1.071	1.070	1.070	1.069	1.064	1.056	1.043	1.027
	4	1.087	1.087	1.086	1.085	1.079	1.069	1.053	1.032
	5	1.098	1.099	1.098	1.096	1.089	1.077	1.058	1.034

sample test in the mixture setting, the arm size approximation formula is shown in (2.3). Again using (D.2), the arm size will be approximately proportional to $\sigma^2/(\theta^2\delta^2)$. It follows that the ratio of the maximum arm size for the group sequential test to the arm size for the fixed sample test primarily only depends on K, α, β, and ρ.

To illustrate the result, Table D.1 shows the exact ratio of the maximum arm size for the group sequential test to the arm size for the fixed sample test for the case $\alpha = 0.05$, $\beta = 0.2$, and $\rho = 2$. What can be seen in Table D.1 is that unless δ/σ is large, the values in the table only depend on K.

References

Anderson, K. M. (2009). *gsDesign: An R Package for Designing Group Sequential Clinical Trials Version 2.0 Manual*, https://r-forge.r-project.org/scm/viewvc.php/*checkout*/pkg/gsDesign/inst/doc/gsDesignManual.pdf?revision=183&root=gsdesign.

Armitage, P., McPherson, C. K., and Rowe, B. C. (1969). Repeated Significance Tests on Accumulating Data, *Journal of Royal Statistical Society. Series A* 132: 235–244.

Chang, M. N., Hwang, I. K., and Shin, W. J. (1998). Group Sequential Designs Using Both Type I and Type II Error Probability Spending Functions, *Communications in Statistics – Theory and Methods* 27: 1323–1339.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, Hillsdale: Erlbaum Associates.

Cytel Inc. (2010). *EAST Advanced Clinical Trial Design, Simulation and Monitoring*, Cambridge: Cytel Inc.

Food and Drug Administration. (2013). *Paving the Way for Personalized Medicine*, https://www.fda.gov/downloads/scienceresearch/specialtopics/personalizedmedicine/ucm372421.pdf.

Gokhale, D. V. and SenGupta, A. (1986). Optimal Tests for the Correlation Coefficient in a Symmetric Multivariate Normal Population, *Journal of Statistical Planning & Inference* 14: 256–263.

Hwang, I. K., Shih, W. J., and De Cani, J. S. (1990). Group Sequential Design Using a Family of Type I Error Probability Spending Functions, *Statistics in Medicine* 9: 1439–1445.

Jennison, C. and Turnbull, B. W. (1999). *Group Sequential Methods with Applications to Clinical Trials*, London: CRC Press.

Jeske, D. R. and Yao, W. (2017). Sample Size Determination and Treatment Effect Estimation If a Treatment Is Only Effective in a Sub-Population.

Lan, K. K. and DeMets, D. L. (1983). Discrete Sequential Boundaries for Clinical Trials, *Biometrika* 70: 659–663.

O'Brien, P. C. and Fleming, T. R. (1979). A Multiple Testing Procedure for Clinical Trials, *Biometrics* 35: 549–556.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., and Smith, P. G. (1976). Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient. I. Introduction and Design, *British Journal of Cancer* 34: 585–612.

Pocock, S. J. (1977). Group Sequential Methods in the Design and Analysis of Clinical Trials, *Biometrika* 64: 191–199.

SAS Institute Inc. (2009). *SAS/STAT® 9.2 User's Guide*, Cary: SAS Institute.

SenGupta, A. (1991). A Review of Optimality of Multivariate Tests, *Statistics & Probability Letters* 12: 527–535.

SenGupta, A. (2007). P3 Approach to Intersection-Union Testing of Hypotheses, *Journal of Statistical Planning and Inference*. 137: 3753–3799.

Spear, B. B., Heath-Chiozzi, M., and Huff, J. (2001). Clinical Application of Pharmacogenetics, *Trends in Molecular Medicine* 7: 201–204.

Spjotvoll, E. (1968). Most Powerful Test for Some Non-Exponential Families, *Annals of Mathematical Statistics* 39: 772–784.

Wang, S.-J., O'Neill, T. R., and Hung, H. J. (2007). Approaches to Evaluation of Treatment Effect in Randomized Clinical Trials with Genomic Subset, *Pharmaceutical Statistics* 6: 227–244.

Wang, S. K. and Tsiatis, A. A. (1987). Approximately Optimal One-Parameter Boundaries for Group Sequential Trials, *Biometrics* 43: 193–200.

Wang, Y. and Leung, D. (1997). Bias Reduction via Resampling for Estimation Following Sequential Tests, *Sequential Analysis* 16: 249–267.

Zhu, L., Ni, L., and Yao, B. (2011). Group Sequential Methods and Software Application, *American Statistician* 65: 127–135