Summer 2022

Non-Inferiority Testing: Kernel Estimation and Overlap Measure

Larie C. Ward

NON-INFERIORITY TESTING: KERNEL ESTIMATION AND OVERLAP MEASURE

by

LARIE WARD

(Under the Direction of Arpita Chatterjee)

ABSTRACT

In non-inferiority testing, the decision of whether a proposed treatment is non-inferior to a reference treatment depends on model assumptions and choices of acceptable tolerance limits. Here, we consider a method that employs kernels to estimate the probability density functions of both the experimental and reference populations from two independent samples. Based on these densities, we introduce a quantity called the overlap coefficient or overlap measure. A bootstrap technique is helpful in exploring the distribution and variance empirically. We derive the distribution of this measure and define a hypothesis test that can be applied to the non-inferiority setting under some simplifying assumptions about distributions of the populations.

INDEX WORDS: Non-inferiority testing, Kernel estimation, Overlap measure, Overlap coefficient, Bootstrap

2009 Mathematics Subject Classification: 62H30, 68H10

NON-INFERIORITY TESTING: KERNEL ESTIMATION AND OVERLAP MEASURE

by

LARIE WARD

B.S., University of Central Florida, 2008

M.S., University of Florida, 2010

Ph.D., University of Florida, 2016

A Thesis Submitted to the Graduate Faculty of Georgia Southern University in Partial

Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

NON-INFERIORITY TESTING: KERNEL ESTIMATION AND OVERLAP MEASURE

by

LARIE WARD

Major Professor:	Arpita Chatterjee
Committee:	Stephen Carden
	Hani Samawi

Electronic Version Approved:
July 2022

ACKNOWLEDGMENTS

I acknowledge my best friend and husband, Dan Gray, who is a being of rare wisdom and infinite patience.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF SYMBOLS

f_R	reference population
f_E	experimental population
μ_R	reference mean
μ_E	experimental mean
σ_R	reference standard deviation
σ_E	experimental standard deviation
n_R	reference sample size
n_E	reference sample size
δ_{NI}	non-inferiority margin
Δ	overlap
Δ_{NI}	overlap margin
$\hat{\Delta}$	estimated overlap

CHAPTER 1

INTRODUCTION

Two years into the Coronavirus pandemic, methods for testing, vaccination and treatment remain at the forefront of public health research on both the national and international level. At the core of much of this research lie questions of effectiveness. When newer, faster tests become available, it is natural to ask if the new product is as good at detecting the virus as the slower test. When a vaccine that is easier to manufacture or transport or administer appears, another effectiveness question arises. This time, we would like to know if the new vaccine provides as much protection as the first. Indeed, as is often the case in medicine, the social sciences and in industry, the question of comparative effectiveness is key.

Elementary hypothesis testing methods provide a foundation for answering such questions, but a more nuanced approach known as *non-inferiority testing* can take into account more structure. In particular, there are times when the benefits to a new product are so desirable, medical experts may be willing to tolerate a small drop in effectiveness. For instance, if a drug showed 90% effectiveness at fighting disease, but had a host of severe side-effects, a new drug that exhibits no side-effects might be largely preferable as long as it is at least 85% effective, say. The goal then is not necessarily to establish the new drug is superior (although it very well could be), but to establish that it is non-inferior.

There is a wealth of articles in the literature concerning non-inferiority testing, both in the frequentist and Bayesian arenas. For those in industry, FDA guidelines provide insight on non-inferiority clinical trials [1, 6]. In 2017, Althunian et al. published an overview pertaining to defining the non-inferiority margin [3]. The effect of programmable bacteria on tumor regression was studied using non-inferiority techniques by Chowdhury et al. in 2019 [5]. Samiran Ghosh et al. explored non-inferiority with Poisson distributed outcomes in 2020 [10]. Santu Ghosh et al., in 2017, published a paper exploring transformations of non-normal distributions [11], as well as a paper giving a hierarchical testing procedure

in 2022 [12]. In 2011, Hida and Tango wrote about non-inferiority testing with a pre-specified margin [15]. Kieser and Friede, in 2011, considered binary endpoints [16]. Koch and Röhmel, in 2004, described three-armed non-inferiority testing as the "gold standard" of testing [17]. Mütze et al. considered negative binomially distributed endpoints in 2009 [19]. Pigeot et al., in 2003, contributed to literature by studying three-armed non-inferiority testing with placebo [20].

A Bayesian perspective, while not the focus of this work, does lend itself to the non-inferiority method as well. Gamalo et al. used a Bayesian approach to non-inferiority trials on anti-infective products in 2014 [8]. Pulak Ghosh et al. also considered three-armed non-inferiority trials with a Bayesian technique in 2011 [9].

Our aim is to introduce the basic terminology of non-inferiority testing and then propose a new way of conducting these tests. The concept of overlap measure, which is essentially the integral of the minimum of two probability densities, will be defined. Some examples and applications of the overlap measure will provide an intuitive way of visualizing the measure. The concept of using overlap to address the non-inferiority problem, is, to our knowledge, a novel one. The proposed test is robust in that it can be used even if the underlying populations are unknown.

The overlap measure, also known as overlap coefficient, is already a well-studied quantity. It can be used to assess similarity and has other applications as well. Al-Saleh and Samawi considered overlap coefficient in two exponential populations in 2007 [2]. Samawi also used overlap to develop a nonparametric test of symmetry based on overlap in 2011 [21]. Anderson et al., in 2009, used overlap to study polarization [4]. In 2021, Franco-Pereira et al. used a binormal approach to perform inference on the overlap coefficient [7]. Giacoletti and Heyse discussed overlap as well as the proportion of similar responses in terms of vaccination research in 2015 [13]. Lei and Olson utilized overlap to gain insight into the similarity of two biologics in 2009 [18]. Tanaka-Mizuno et al. applied overlap to

assessing the similarity of ethnically different populations in terms of pharmacokinetic data in 2005 [22]. In 2016, Wang and Tian wrote about confidence interval estimation of the overlap coefficient [23]. There are several other articles in the literature concerning overlap measure as well.

This thesis, which proposes a test of non-inferiority based on overlap measure, features a simulation chapter that demonstrates how well the proposed test performs. We will consider the case first where both populations, the reference and the experimental, are normal. Within this case, different sample sizes will be considered, along with differing variances. Both the probability of type I error and the power are examined. For a second case, we turn to non-normal distributions, such as the chi-squared and exponential distributions. We investigate the usefulness of the test as compared to traditional methods. In some situations the overlap test provides more power then existing tests, which is an encouraging sign for this new test.

The subsequent sections are organized as follows: Section 2 is dedicated to non-inferiority testing, while section 3 is dedicated to overlap measure. Those two concepts are married together in section 4 where we formally introduce the technique of using overlap measure to test for non-inferiority. Section 5 is reserved for simulation and discussion.

CHAPTER 2

NON-INFERIORITY TESTING

2.1 Definitions and Notation

Suppose we have two populations. Further suppose that one of these populations receives a treatment of established efficacy, while the other population receives a treatment of unknown efficacy.

Definition 1. The **reference population** is the group that receives the reference treatment, which is of known efficacy.

Definition 2. The **experimental population** is the group that receives the experimental treatment, which is of unknown efficacy.

- μ_R is the mean efficacy for the reference population.
- μ_E is the mean efficacy for the experimental population.

If a researcher wished to establish that the experimental treatment was better than the reference treatment, such a test would have the form shown below in 2.1 known as a superiority trial.

$$H_0 : \mu_E \leq \mu_R \quad \text{versus} \quad H_a : \mu_E > \mu_R \tag{2.1}$$

In the medical and pharmaceutical communities, however, a test like the one above may take on a different form. If the experimental drug has some logistical advantage over the reference, such as ease of production or transportation, the researcher may relax what they wish to establish. A margin, which we will call the margin of non-inferiority, is agreed upon a priori by medical experts. Once this margin is chosen, we can ask whether the experimental treatment is within this margin of the reference.

Definition 3. The **non-inferiority margin**, denoted δ_{NI}, is a quantity established by medical experts wherein a treatment within this margin of the reference treatment is deemed to be of value.

The choice of δ_{NI} is of great consequence. If it is chosen to be too small, we may end up discarding a treatment that is actually of some value in helping patients who might benefit from the treatment. If it is chosen too large, we may adopt broad use of a treatment that is much less effective then the reference treatment. For a detailed discussion on how the margin is chosen in practice, [3] considers ways to make a selection. The test used to establish non-inferiority is shown in 2.2.

$$H_0 : \mu_E \leq \mu_R - \delta_{NI} \quad \text{versus} \quad H_a : \mu_E > \mu_R - \delta_{NI} \tag{2.2}$$

Definition 4. If the test shown in 2.2 is significant, we say the experimental treatment is **non-inferior** to reference treatment, or that **non-inferiority** has been established.

To visualize this, we can divide the real number line into two regions as shown in Figure 2.1. The mean of the experimental group μ_E may lie in the inferiority region or the non-inferiority region.

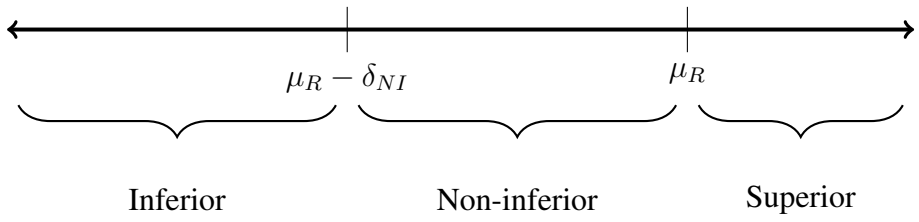

Figure 2.1: The experimental mean may lie in any region.

Using this language, our null hypothesis amounts to stating that μ_E falls in the inferiority region. Our alternative states μ_E falls in the non-inferiority region or the superior region.

CHAPTER 3

THE OVERLAP MEASURE

3.1 An Overlap Example

First introduced by M.S. Weizmann in 1970, the quantity known as Overlap Measure offers insight when it comes to measuring similarity between a pair of non-negative real-valued functions, f_1 and f_2. In some contexts, f_1 and f_2 may represent unknown probability distributions. Thus, it is helpful to define a quantity that represents an overlap estimate.

Definition 5. The **Overlap Measure**, Δ, of f_1 and f_2 is given by

$$\Delta = \int \min\{f_1(x), f_2(x)\}dx$$

Note if f_1 and f_2 are both probability densities, $0 \leq \Delta \leq 1$.

For instance, suppose f_1 and f_2 are normal distributions with a standard deviation of 1 unit and further suppose the mean of f_1, denoted by μ_1, is exactly 1 unit larger than the mean of f_2, denoted by μ_2. In this simple setting, we can readily compute Δ. Figure 3.1 illustrates the desired area.

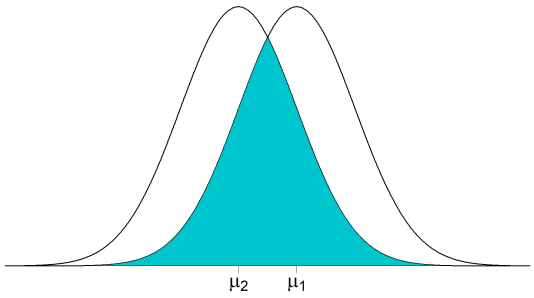

Figure 3.1: Overlap measure of two normal distributions of equal variance.

Since the area is symmetric about $x = \frac{\mu_1+\mu_2}{2}$, we find

$$\Delta = 2 \cdot P_{f_1}\left(X < \frac{\mu_1 + \mu_2}{2}\right) \approx 2 \cdot 0.3085375 = 0.617075$$

3.2 Applications and an Estimate

In terms of applications, Weizmann himself employed overlap to study the difference in income among two racial groups. Consequently, the overlap measure is often called "Weizmann's Overlap" in his honor.

More recently, in a 2011 paper, Samawi et al. [21] considered the special case where $f_2(x) = f_1(-x)$. In that setting, the overlap measure can be used to detect symmetry since a highly symmetric function will have a great deal of overlap with its mirror image. (The reflection is over the median.)

Definition 6. The **Estimated Overlap Measure**, $\hat{\Delta}$, of f_1 and f_2 is given by

$$\hat{\Delta} = \int \min\{\hat{f}_1(x), \hat{f}_2(x)\}dx$$

where $\hat{f}_1(x)$ and $\hat{f}_2(x)$ are estimates for f_1 and f_2.

If little or no information about the distribution of f_1 and f_2 is available, and instead we have a pair of random samples from f_1 and f_2, respectively, it is natural to take $\hat{f}_1(x)$ and $\hat{f}_2(x)$ to be kernel estimates with an appropriate choice of bandwidth. For a straightforward guide on kernel density estimation, the 2006 text *Nonparametric and Semiparametric Models* by Härdle et al., is a good starting point [14] . The only kernel function used in this work will be the Gaussian.

3.3 Asymptotic Results

According to Anderson (2009), the asymptotic distribution of $\hat{\Delta}$ can be obtained along the following lines. We start with partitioning the combined support of f_1 and f_2 into three

subsets of the real line:

$$\begin{aligned} C_{f_1,f_2} &= \{x \in \mathbb{R} : f_1(x) = f_2(x) > 0\} \\ C_{f_1} &= \{x \in \mathbb{R} : f_2(x) > f_1(x) > 0\} \\ C_{f_2} &= \{x \in \mathbb{R} : f_1(x) > f_2(x) > 0\} \end{aligned}$$

Next these three probabilities are defined:

$$\begin{aligned} p_0 &= P(X \in C_{f_1,f_2}) \\ p_1 &= P_{f_2}(X \in C_{f_1}) \\ p_2 &= P_{f_1}(X \in C_{f_2}) \end{aligned}$$

Asymptotically, as $n \to \infty$,

$$\sqrt{n}(\hat{\Delta} - \Delta) - a_n \quad \Longrightarrow \quad N(0, v)$$

where,

$$\begin{aligned} a_n &= b^{-d/2}||K||_2^2 \int_{C_{f,g}} f^{1/2}(x)dx \cdot E \min\{Z_1, Z_2\} \\ v &= p_0\sigma_0^2 + \sigma_1^2 \\ \sigma_1^2 &= p_1(1-p_1) + p_2(1-p_2) \\ \sigma_0^2 &= ||K||_2^2 \int_{T_0} \text{cov}\big(\min\{Z_1, Z_2\}, \min\{\rho(t)Z_1 + \sqrt{1-\rho(t)^2}Z_3, \rho(t)Z_2 + \sqrt{1-\rho(t)^2}Z_4\}\big)dt \\ T_0 &= \{t \in \mathbb{R}^d \big| ||t|| \leq 1\} \end{aligned}$$

Here, K is a kernel function and Z_i for $i \in \{1, 2, 3, 4\}$ are standard normal random variables. The variables b and d represent the bandwidth and dimension, respectively. Also,

$$||K||_2^2 = \int_{\mathbb{R}^d} K^2(u)du \text{ and } \rho(t) = \int_{\mathbb{R}^d} K(u)K(u+t)du/||K||^2$$

In the special case that both f_1 and f_2 are normal, many simplifications occur. If they have the same standard deviation, they have just one point of intersection. If they have

different standard deviations, they have just two points of intersection. This means the set $C_{1,2}$ has measure 0. Integrating over a set of 0 measure gives 0, so $a_n = 0$. Furthermore, $p_0 = P(X \in C_{f_1,f_2}) = 0$, so $v = p_0\sigma_0^2 + \sigma_1^2 = \sigma_1^2$. Thus, we have the test statistic,

$$z_0 = \frac{\hat{\Delta} - \Delta_0}{\hat{\sigma_1}} \sim N(0,1)$$

We can use this to find critical value and reject H_0 if $z_0 > z_\alpha$.

Finally, we have

$$\sqrt{n}(\hat{\Delta} - \Delta) \implies N(0, \sigma_1^2)$$

To summarize, the estimated overlap converges in distribution to a normal distribution.

CHAPTER 4

APPLYING THE OVERLAP MEASURE TO NON-INFERIORITY TESTING

4.1 MOTIVATION

The idea of using overlap measure to study non-inferiority is motivated by noticing an inverse relationship between the size of the overlap and the distance between the means. For now, we will assume both populations are unimodal, meaning they have one peak, and we will assume they are symmetric. Further assume the mean of the experimental population, μ_E is less than that of the reference population, μ_R. Under these conditions, the smaller the distance between the means, the larger the overlap. Figure 4.1 shows this relationship with a series of overlap areas.

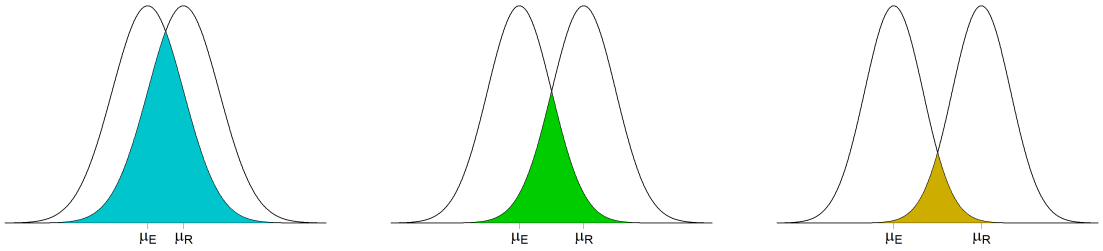

Figure 4.1: When the means are closer, the overlap is larger

There will be situations when the assumption of symmetry can be dropped which will allow for the study of skewed distributions. For example, this relationship between means and overlaps still holds in the case of chi-squared and exponential distributions.

4.2 Hypotheses for the Proposed Test

With a bit more notation, we will be able to revisit the non-inferiority hypotheses and rewrite them in terms of overlap. Let f_R and f_E be the probability density functions for the reference population and the experimental population, respectively.

Definition 7. The **Overlap Margin** is the overlap of f_R and f_E, when $\mu_E = \mu_R - \delta_{NI}$. Denote it by Δ_{NI}. We can think of this quantity as the *overlap at the boundary* of our hypothesis test. Let Δ be the **True Overlap** of f_R and f_E.

Using this terminology, the hypotheses for the proposed non-inferiority test are given in 4.1.

$$H_0 : \Delta \leq \Delta_{NI} \quad \text{(Small Overlap)} \quad \text{versus} \quad H_a : \Delta > \Delta_{NI} \quad \text{(Large Overlap)} \tag{4.1}$$

As long as Δ is a decreasing function of $\mu_R - \mu_E$, the proposed test is equivalent to the non-inferiority test laid out in Chapter 2. To see this, it is helpful to compare the hypotheses above to those in 2.2.

Recall that the null hypothesis from 2.2, which states $\mu_E \leq \mu_R - \delta_{NI}$, represents the assumption the experimental drug is inferior. The proposed null hypothesis above, which states $\Delta \leq \Delta_{NI}$, does the same, but with a different connotation. Instead of saying "the means are farther away than δ_{NI}," we are saying "the overlap is smaller than Δ_{NI}."

Also, recall that the alternative hypothesis from 2.2, which states $\mu_E > \mu_R - \delta_{NI}$, represents the assumption the experimental drug is non-inferior. The proposed alternative hypothesis above, which states $\Delta > \Delta_{NI}$, again does the same. Here, instead of saying "the means are closer together than δ_{NI}," we are saying "the overlap is larger than Δ_{NI}." Non-inferiority is established if we reject H_o in favor of H_a.

CHAPTER 5

RESULTS OF SIMULATION

5.1 INVESTIGATION

We now turn to learning more about the proposed test through simulation. Primarily, we are concerned with the probability of type I error, $P(\text{Type I})$, and with the power of the test. We'll address these questions for a wide variety of situations by sampling from the assumed reference population and from the experimental population.

Firstly, we'll restrict our attention to the case where both samples are drawn from a normal population. Within this setting, we offer several simulations according to sample size, specifically, where sample sizes are both set at 100, then where the sample sizes are both set at 50, and lastly, where the sample sizes are in a 2:1 ratio by setting the experimental sample size at 100 while setting the reference sample size at 50. Each of these three situations has an effect on the power function. Other possibilities such as a 3:1 ratio or a 3:2 ratio are not part of this work, but they may represent avenues for further research.

We also present simulation results according to several possible choices of spread in the two populations. Most simple of these is the situation where the standard deviation of both the reference and experimental populations are equal to 1. Next, is the situation where the standard deviation of the experimental is "somewhat larger" than the reference. Specifically, they are set at 1.5 and 1 respectively. Lastly, is the situation where the standard deviation of the experimental is "much larger" than the reference, set at 2 and 1 respectively. One might ask why we did not investigate any situations where the spread of the reference population is larger than the spread of the experimental. Certainly, this avenue could be explored with further simulation, but we chose not to at this time because a reference population is typically a well-studied group. On the other hand, an experimental population is a new group, and it is natural to expect more or equal spread.

With three situations under consideration for sample size choices and three situations under consideration for spread, this gives rise to a total of nine sub-cases. For example, there is one sub-case where the ratio of sample sizes is 2:1 and the variances are equal. For another example, there is one sub-case where the sample sizes are both 50, but the spread of the experimental is somewhat larger than that of the reference. Our aim will be to create a sketch of a power function for each of these nine cases and compare our proposed test to traditional testing to see how well it does.

Secondly, we will move to the case where both samples are drawn from non-normal populations. Of course, there are many non-normal distributions we could consider, but at this time, three in particular will be studied in detail. The chi-squared distribution with 1 degree of freedom, the chi-squared distribution with 3 degrees of freedom, and the exponential distribution with $\lambda = 1$.

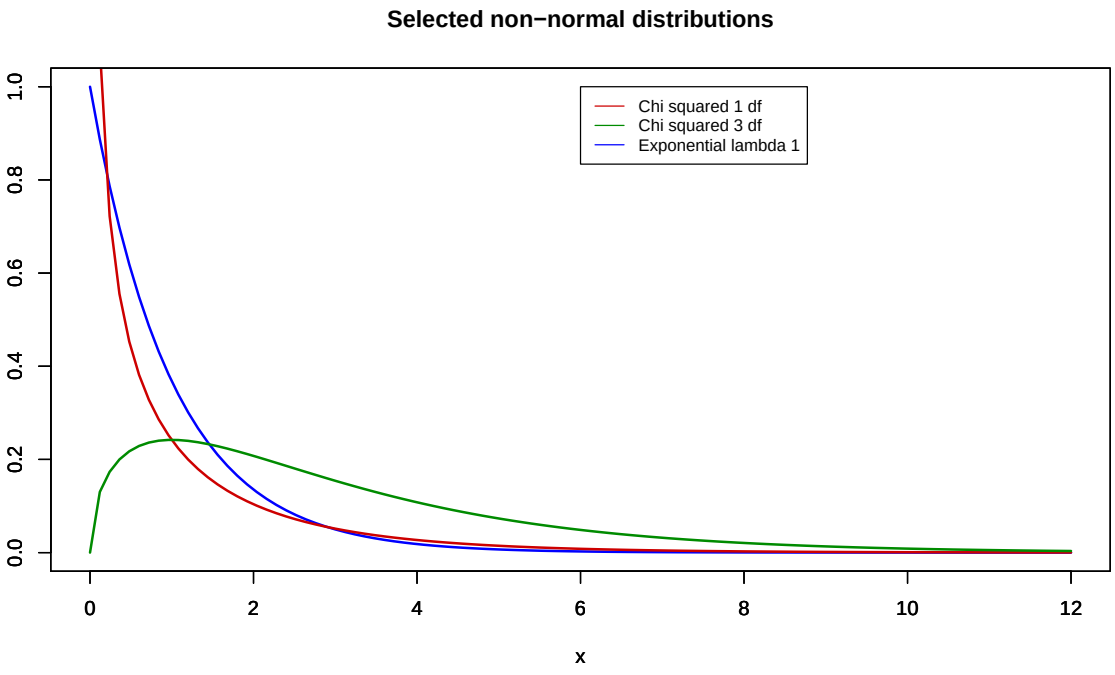

Figure 5.1: Selected non-normal distributions

The goal in considering three different non-normal distributions is to get a sense of how the proposed test does, compared to standard testing, when we sample in a skewed context. For reference, Figure 5.1 has, in red, the chi-squared with 1 degree of freedom, in green, the chi-squared with 3 degrees of freedom, and in blue, the exponential. It is with these non-normal distributions we will see the benefit of using overlap testing.

5.2 Sampling from Normal Populations–Methods

In this section, we present the results we found by sampling from normal populations. We found these results with the aid of a free software environment known as R to handle computing and produce graphs of the power functions.

We first fixed the sample sizes for the reference and experimental populations, respectively. Throughout, n_R and n_E denote these sample sizes. As noted above, we ran simulations when $n_R = n_E = 100$, when $n_R = n_E = 50$, and when $n_R = 50,\ n_E = 100$. Next, we fixed our μ_R, which is the reference mean, at 2.8. There is no special significance to this value since any real number can serve in this role, however, it does need to be a fixed value. The standard deviation of the reference population, σ_R, was fixed at 1 for ease. Also, the non-inferiority margin, δ_{NI}, was fixed at 1. The mean for the experimental population, μ_E is governed by expression below.

$$\mu_E = \mu_R - \delta_{NI} + \xi \tag{5.1}$$

The parameter, ξ in 5.1, can be adjusted at different levels to gain information about the probability of Type I error and the power of our proposed test. When $\xi = 0$, the experimental mean is at the boundary of our test. Thus, when we test a large number of times, the proportion of times H_0 is rejected provides an estimate for $P(\text{Type I})$. When $\xi > 0$, the experimental mean is within the non-inferiority region. The larger ξ is, the more firmly within the non-inferiority region, μ_E lies. Thus, we gain insight into the power of

the test by choosing a range of positive values for ξ. The values for ξ we investigated in this work are the values $\{0, 0.05, 0.2, 0.5, 0.8\}$. Consequently, these values will be marked on the horizontal axis in the estimated power functions we will present. The overall appearance of these graphs might be somewhat smoother if we had chosen more than these five points, however, they provided enough information to see the power tending toward 1 as ξ increases.

In this light, the mean of the experimental population, μ_E, depends on ξ so it will take on different values for different simulations. The last value we need to specify is the standard deviation of the experimental population, σ_R. As mentioned above, we ran simulations when $\sigma_E = 1$, when $\sigma_E = 1.5$ and when $\sigma_E = 2$. With these quantities, μ_R, σ_R, σ_E, n_R, n_E, δ_{NI} and ξ all specified, we can start sampling.

5.2.1 OBTAINING Δ_{NI}

Initially, we need to find Δ_{NI} by taking these steps:

1. For a fixed μ_R and δ_{NI}, obtain $\mu_E = \mu_R + \delta_{NI}$ at the **boundary** of NI hypothesis.
2. Generate samples of specified size for both arms.
3. Estimate the overlap based on the samples.
4. Repeat steps 2 and 3 a large number of times.
5. Δ_{NI} is obtained from the mean of the repeated estimates.

To estimate the variance of our estimated overlap, $var(\hat{\Delta})$, a bootstrap technique with a large number of iterations was used.

5.2.2 OBTAINING $\hat{var}(\hat{\Delta})$

These are the steps taken on each bootstrap iteration to find $\hat{var}(\hat{\Delta})$.

- For a given set of data, form both arms, X_R and X_E,
    1. Take n_E and n_R observations with replacement from each data set, respectively.
    2. Estimate the overlap, $\hat{\Delta}$.
    3. Repeat steps 1 and 2 over B times.
- Denote these estimates as $\hat{\Delta}_b,\ b = 1, 2, ..., B$.
- Obtain $\hat{var}(\hat{\Delta})$ as

$$\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\Delta}_b - \bar{\Delta}\right)^2$$

With this estimated variance in place, we can turn our attention toward calculating the power of the test as well as the probability of type I error. In the steps below, we should emphasize that if $\xi = 0$, the proportion of times the hypothesis is rejected tells us the $P(\text{Type I})$. On the other hand, if $\xi > 0$, the proportion of times the hypothesis is rejected tells the power of the test. In general, these are the steps for the normal case:

1. Specify μ_R, σ_R, σ_E, n_R, n_E, δ_{NI} and ξ.
2. Set $\mu_E = \mu_R - \delta_{NI} + \xi$
3. Generate $X_L \sim N(\mu_L, \sigma_L)$ with sample size n_L for $L \in \{R, E\}$.
4. Obtain $\hat{\Delta}$ and $\hat{var}(\hat{\Delta})$.
5. Get the test statistic, Z.
6. Repeat steps 3 through 5 over M times.
7. Obtain power, or $P(\text{Type I})$, depending on ξ, as the proportion of times the hypothesis is rejected.

Note that each of these seven steps below was carried out nine times since there are nine situations we considered. In step 4, kernel density estimation is used to estimate the density of each population, then the integral of the minimum is taken numerically. We will organize the results in a series of tables first, then several power function graphs.

5.3 Sampling from Normal Populations-Results

Below, we see 5.1 indeed has nine entries, and within each cell there are two numbers. Both numbers represent the probability of type I error. However, the numbers inside the parentheses represent the traditional two-sample hypothesis test while the numbers outside the parentheses represent the new proposed hypothesis test. In other words, we use the overlap test statistic in order to decide whether or not to reject H_0. Throughout, the significance level, $\alpha = 0.05$ is used.

Roughly speaking, all the entries are near our choice of $\alpha = 0.05$, as expected. In tables 5.2 through 5.5, the choice of ξ is increasing. That corresponds to placing $\mu_E = \mu_R - \delta_{NI} + \xi$ more firmly into the non-inferiority region. In turn, this corresponds to placing μ_E more firmly into the rejection region. Consequently, the power, which is the probability of rejecting H_0 ought to increase towards 1. The more rapidly a test does this, the more powerful it is.

$\xi = 0$	$n_R = n_E = 100$	$n_R = n_E = 50$	$n_R = 50,\ n_E = 100$
$\sigma_R = \sigma_E = 1$	(0.0580) 0.0460	(0.0570) 0.0545	(0.0520) 0.0445
$\sigma_R = 1,\ \sigma_E = 1.5$	(0.0560) 0.0475	(0.0515) 0.0495	(0.0410) 0.0355
$\sigma_R = 1,\ \sigma_E = 2$	(0.0505) 0.0345	(0.0475) 0.0355	(0.0515) 0.0470

Table 5.1: Probability of type I error when sampling from normal, $\xi = 0$.

What we notice that when the standard deviations are equal, the proposed test is almost as powerful as the traditional test, especially when both sample sizes at 100. When the

$\xi = 0.05$	$n_R = n_E = 100$	$n_R = n_E = 50$	$n_R = 50,\ n_E = 100$
$\sigma_R = \sigma_E = 1$	(0.1090) 0.0860	(0.1028) 0.0833	(0.0935) 0.0825
$\sigma_R = 1,\ \sigma_E = 1.5$	(0.0965) 0.0780	(0.0696) 0.0686	(0.0805) 0.0605
$\sigma_R = 1,\ \sigma_E = 2$	(0.0905) 0.0690	(0.0680) 0.0535	(0.0815) 0.0545

Table 5.2: Power when sampling from normal, $\xi = 0.05$.

$\xi = 0.2$	$n_R = n_E = 100$	$n_R = n_E = 50$	$n_R = 50,\ n_E = 100$
$\sigma_R = \sigma_E = 1$	(0.4015) 0.3605	(0.267) 0.231	(0.3155) 0.2580
$\sigma_R = 1,\ \sigma_E = 1.5$	(0.2900) 0.2240	(0.1785) 0.1730	(0.2290) 0.1745
$\sigma_R = 1,\ \sigma_E = 2$	(0.2182) 0.1278	(0.1565) 0.0980	(0.2068) 0.1218

Table 5.3: Power when sampling from normal, $\xi = 0.2$.

$\xi = 0.5$	$n_R = n_E = 100$	$n_R = n_E = 50$	$n_R = 50,\ n_E = 100$
$\sigma_R = \sigma_E = 1$	(0.9700) 0.9525	0.804 0.759	0.8915 0.8580
$\sigma_R = 1,\ \sigma_E = 1.5$	(0.8823) 0.7654	(0.6330) 0.4760	(0.7830) 0.5460
$\sigma_R = 1,\ \sigma_E = 2$	(0.7340) 0.3080	(0.4815) 0.1795	(0.6585) 0.2380

Table 5.4: Power when sampling from normal, $\xi = 0.5$.

$\xi = 0.8$	$n_R = n_E = 100$	$n_R = n_E = 50$	$n_R = 50,\ n_E = 100$
$\sigma_R = \sigma_E = 1$	(1.000) 1.000	(0.9241) 0.9031	(0.9995) 0.9980
$\sigma_R = 1,\ \sigma_E = 1.5$	(0.9950) 0.9290	(0.9355) 0.7485	0.9845 0.8250
$\sigma_R = 1,\ \sigma_E = 2$	(0.9695) 0.4590	(0.8240) 0.2790	(0.9500) 0.3415

Table 5.5: Power when sampling from normal, $\xi = 0.8$.

standard deviations are somewhat different, the proposed test is noticeably less powerful than the traditional test, even when both sample sizes at 100. When the standard deviations

are very different, the overlap test is not very powerful at all, regardless of sample size. It is helpful to look at some graphs of the power functions to help us visualize these findings.

In Figure 5.2, we include all the cases where $n_R = n_E = 100$. The red indicates the case where $\sigma_R = \sigma_E = 1$. The gold indicates the case where $\sigma_R = 1$ and $\sigma_E = 1.5$. The green indicates the case where $\sigma_R = 1$ and $\sigma_E = 2$. Throughout all colors, the dotted line represents traditional testing and the solid line represents the overlap testing.

We notice that the lines are rather close together and interpret that our proposed overlap test does quite well when the standard deviations are equal and both sample sizes are large. The gold has a noticeable disparity in power. Our proposed test is markedly less powerful than traditional testing when the standard deviations are somewhat different (meaning that $\sigma_R = 1$ and $\sigma_E = 1.5$). The green has a gap in power so severe, that the overlap test loses its usefulness completely when the standard deviations are very different ($\sigma_R = 1$ and $\sigma_E = 2$.)

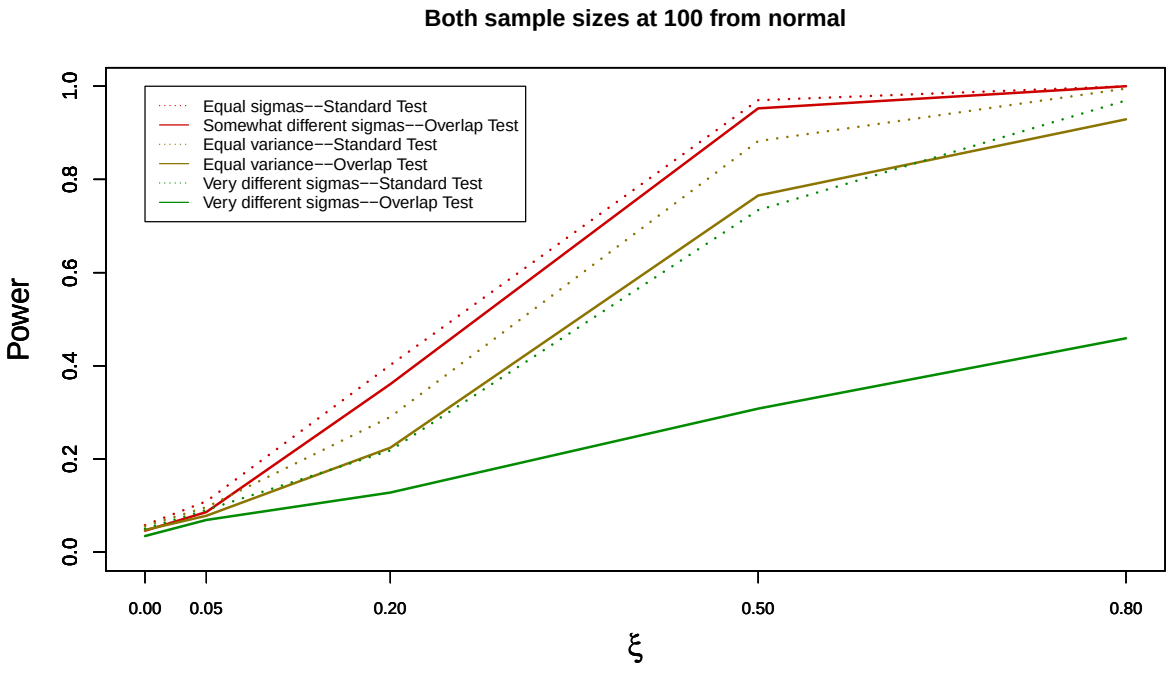

Figure 5.2: Both sample sizes at 100 from normal

This pattern of the overlap test performing very nicely for equal standard deviations, moderately nicely for somewhat different standard deviations and very poorly for very different standard deviations continued to appear even as the sample size was altered. Figure 5.3 includes all the cases where $n_R = n_E = 50$ and Figure 5.4 includes all the cases where $n_R = 50$ and $n_E = 100$. If any lesson can be learned from altering the sample size, it is that all cases, both traditional and overlap, are less powerful when the sample sizes are not both set at 100.

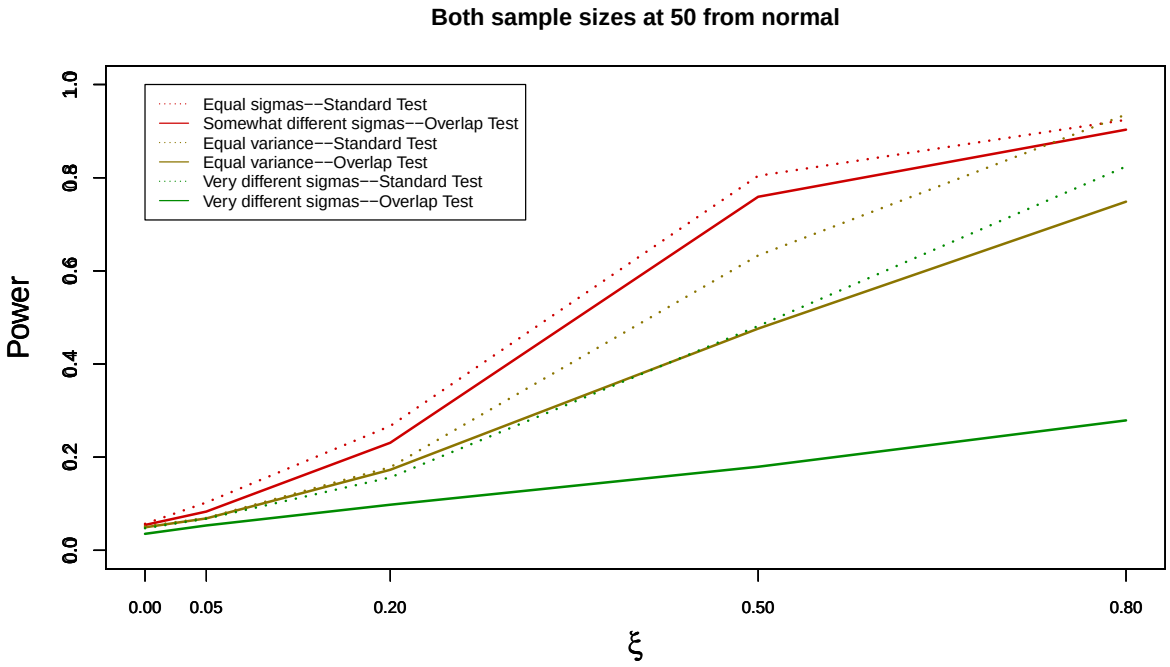

Figure 5.3: Both sample sizes at 50 from normal

At this stage, one might question the utility of the overlap test since all indications so far are telling us the proposed test is not as powerful as traditional tests. In some sense, we should not expect to see an improvement in power when we sample from a normal population because it is known that the standard test is uniformly most powerful. We need to try sampling from non-normal populations in order to see situations where the overlap test proves its worth.

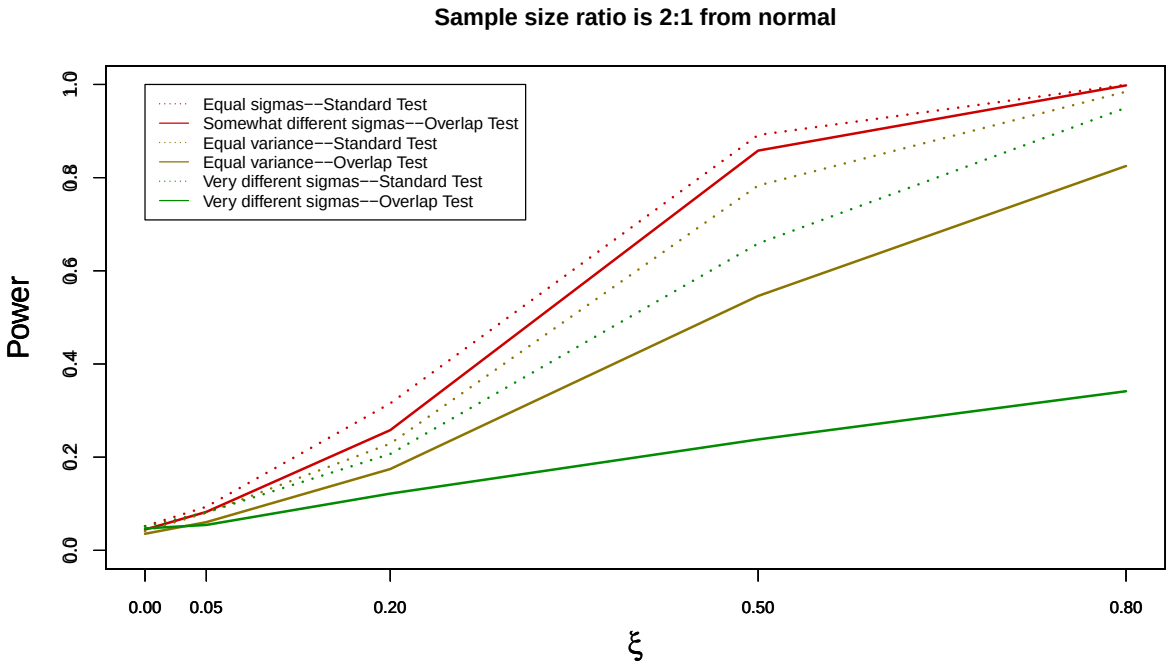

Figure 5.4: Sample size ratio is 2:1 from normal

5.4 Sampling from $\chi^2(1)$ Populations–Methods

Just as with the normal case, we ran simulations when $n_R = n_E = 100$, when $n_R = n_E = 50$ and when when $n_R = 50,\ n_E = 100$. For ease, we will sample the reference from a shifted chi-squared distribution. In particular, the reference is chosen from $\chi^2(k) - k$ and the experimental is chosen from $\chi^2(k) - k - \delta_{NI} + \xi$ where $k = 1$. The discussion from the previous section about obtaining Δ_{NI} and estimating $var(\hat{\Delta})$ is performed analogously. Here are the steps taken:

1. Specify k, n_R, n_E, δ_{NI} and ξ.

2. Generate $X_R \sim \chi^2(k) - k$ and $X_E \sim \chi^2(k) - k - \delta_{NI} + \xi$ with sample size n_R and n_E, respectively.

3. Obtain $\hat{\Delta}$ using the kernel functions.

4. And $\hat{var}(\hat{\Delta})$, using bootstrap samples.

5. Get the test statistic, Z.

6. Repeat steps 2 through 5 over M times.

7. Obtain power, or $P(\text{Type I})$, depending on ξ, as the proportion of times the hypothesis is rejected.

Unlike with the normal case, we did not consider various standard deviations. In fact, the experimental distribution is a horizontal translate of the reference distribution so it inherits the same standard deviation. With less cases to show, we can fit all choices of ξ into a single table. The graphs will be less busy as well. Each one will show just two lines instead of six.

5.5 Sampling from $\chi^2(1)$ Populations–Results

Table 5.6 has two numbers in each cell. Again, numbers inside the parentheses represent the traditional two-sample hypothesis test while the numbers outside the parentheses represent the new proposed hypothesis test. It appears the overlap test is more powerful than the standard test. This is easier to see with graphs.

$\chi^2(1)$	$n_R = n_E = 100$	$n_R = n_E = 50$	$n_R = 50,\ n_E = 100$
$\xi = 0$	(0.048) 0.049	(0.053) 0.0525	(0.0482) 0.053
$\xi = 0.05$	(0.094) 0.096	(0.075) 0.0525	(0.077) 0.087
$\xi = 0.2$	(0.266) 0.370	(0.182) 0.221	(0.234) 0.336
$\xi = 0.5$	(0.805) 0.977	(0.587) 0.822	(0.6606) 0.930
$\xi = 0.8$	(0.979) 1.000	(0.878) 0.962	(0.934) 0.996

Table 5.6: Probability of type I error and power from $\chi^2(1),\ \xi \in \{0, 0.05, 0.2, 0.5, 0.8\}$.

Figure 5.5 is very encouraging for our proposed test. It is noticeably more powerful in the setting where $n_R = n_E = 100$, particularly as we look beyond values of $\xi = 0.05$. This noticeable improvement is also evident in Figure 5.6 where $n_R = n_E = 50$ and in Figure 5.7 where $n_R = 50$ and $n_E = 100$. These results suggest that the overlap test is better at detecting non-inferiority than the standard test in at least some settings. Now we turn our attention to another chi-squared, but this time, with $k = 3$.

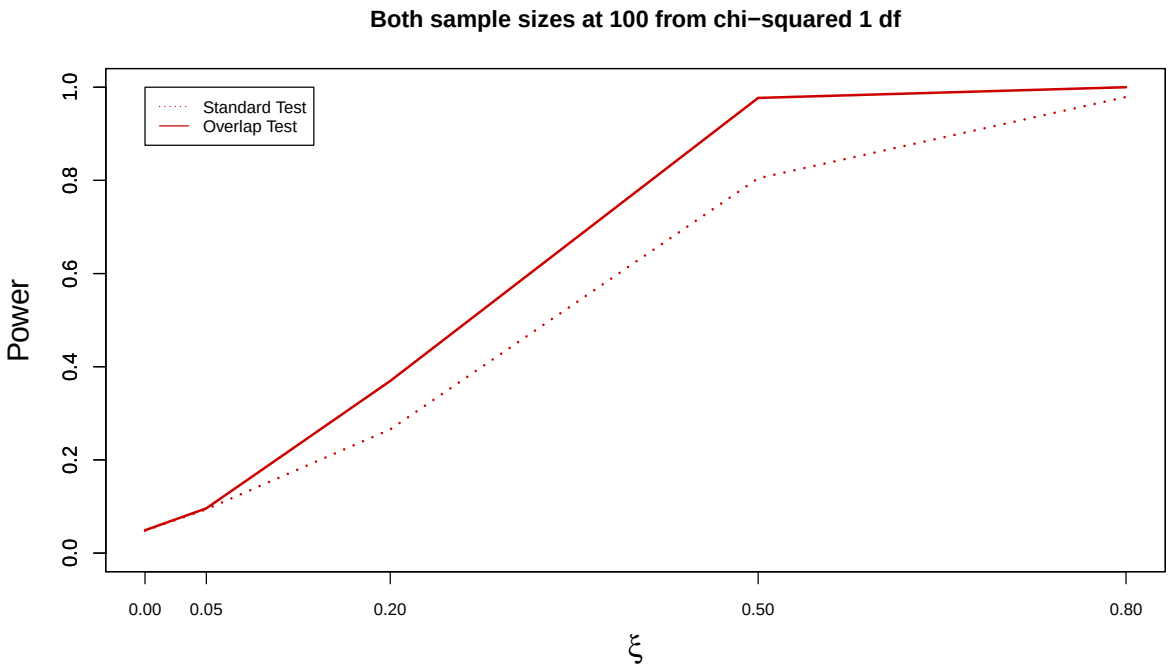

Figure 5.5: Both sample sizes at 100 from chi-squared 1 df

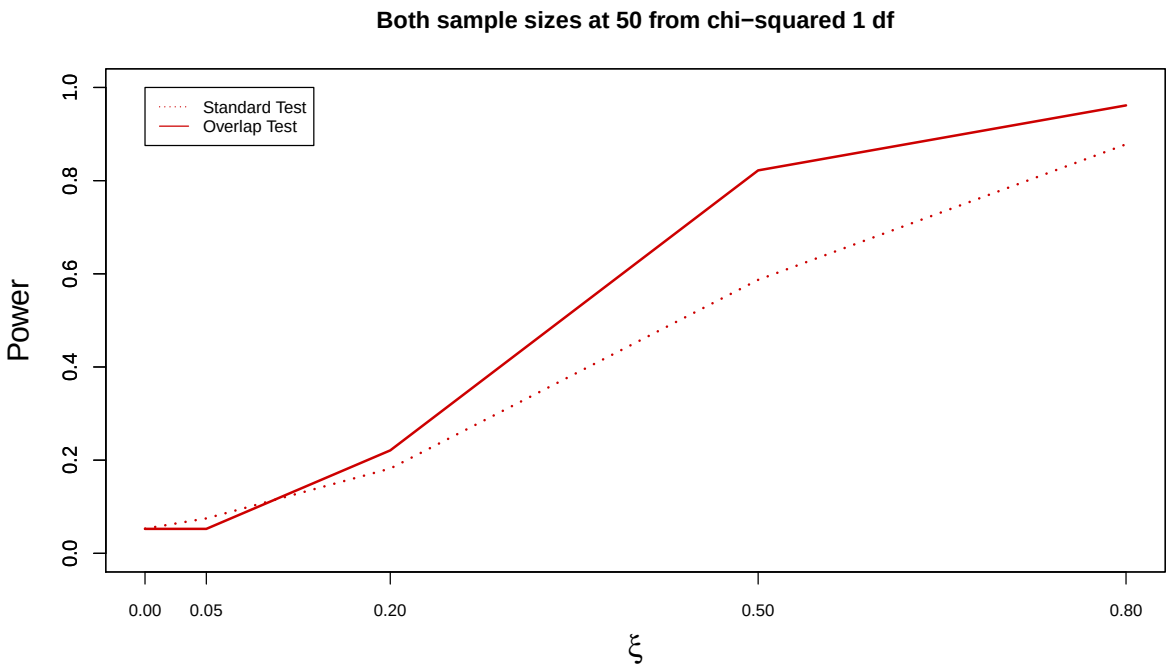

Figure 5.6: Both sample sizes at 50 from chi-squared 1 df

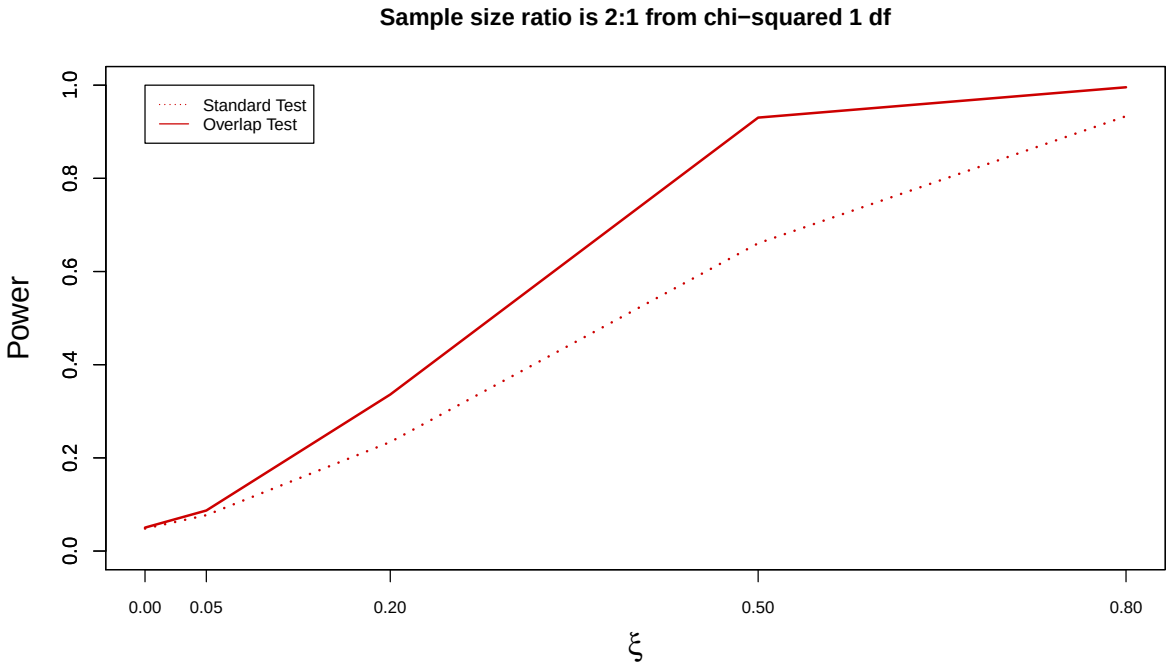

Figure 5.7: Sample size ratio is 2:1 from chi-squared 1 df

5.6 SAMPLING FROM $\chi^2(3)$ POPULATIONS–METHODS

While the previous section utilized the chi-squared distribution with just one degree of freedom, we now consider 3 degrees of freedom. In other words, $k = 3$, but no other major changes are made. In particular, the reader may consult the list of steps given in the previous section and replace everywhere $k = 1$ with $k = 3$. However, since there is more spread, in this case, it is necessary to view slightly larger values of ξ in order to observe the power tending towards 1.

5.7 SAMPLING FROM $\chi^2(3)$ POPULATIONS–RESULTS

With $k = 3$ degrees of freedom, the overlap test is noticeably more powerful than the traditional 2-means test for all cases. Each entry of Table 5.7 has the value inside the parentheses, which comes from a standard test, which is smaller than the value outside the parentheses which comes from the overlap test.

In Figures 5.8 through 5.10 display this with a moderately sized gap between the power for the overlap test in solid red and the standard test in dotted red. Interestingly, the gap is actually smaller for smaller values of ξ and larger for larger values of ξ. Thus, the advantage of using the overlap test becomes greater when the experimental population is well within the non-inferiority region.

$\chi^2(3)$	$n_R = n_E = 100$	$n_R = n_E = 50$	$n_R = 50,\ n_E = 100$
This $\xi = 0$	(0.048) 0.054	(0.046) 0.068*	(0.042) 0.061
$\xi = 0.05$	(0.063) 0.110	(0.068) 0.086*	(0.064) 0.078
$\xi = 0.2$	(0.147) 0.176	(0.118) 0.166*	(0.118) 0.136
$\xi = 0.5$	(0.448) 0.534	(0.346) 0.372	(0.333) 0.379
$\xi = 0.8$	(0.693) 0.736	(0.461) 0.620	(0.484) 0.599
$\xi = 1.1$	(0.724) 0.932	(0.616) 0.853	(0.631) 0.815

Table 5.7: Probability of type I error and power from $\chi^2(3),\ \xi \in \{0, 0.05, 0.2, 0.5, 0.8, 1.1\}$. (*indicates run with $M = 500$.)

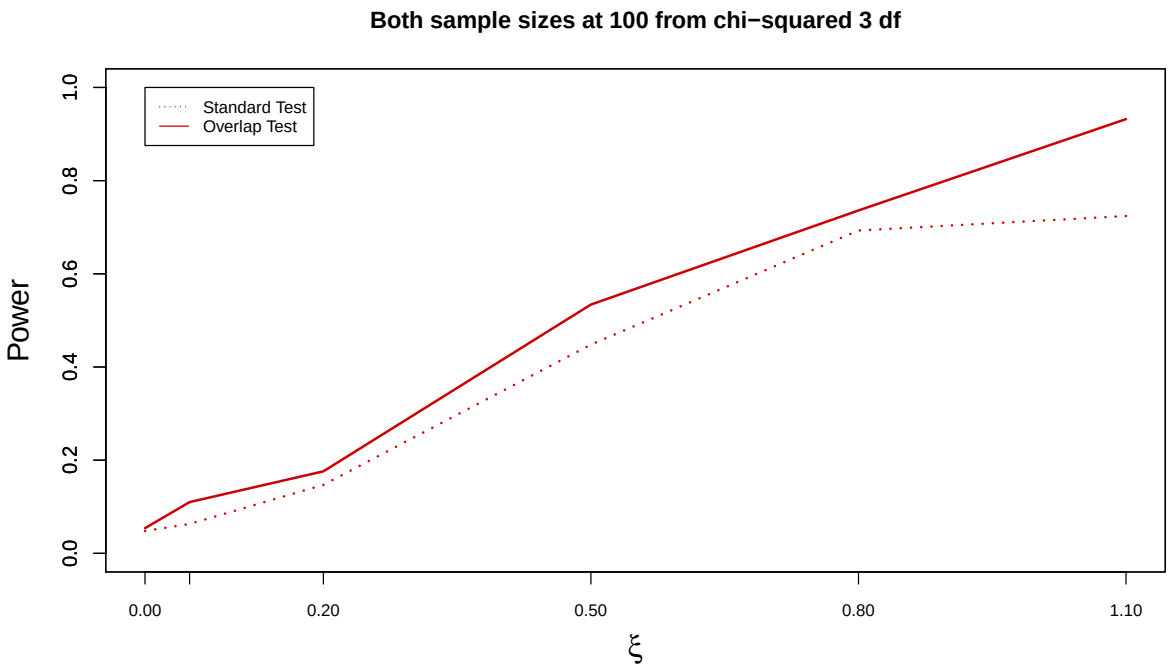

Figure 5.8: Both sample sizes at 100 from chi-squared 3 df

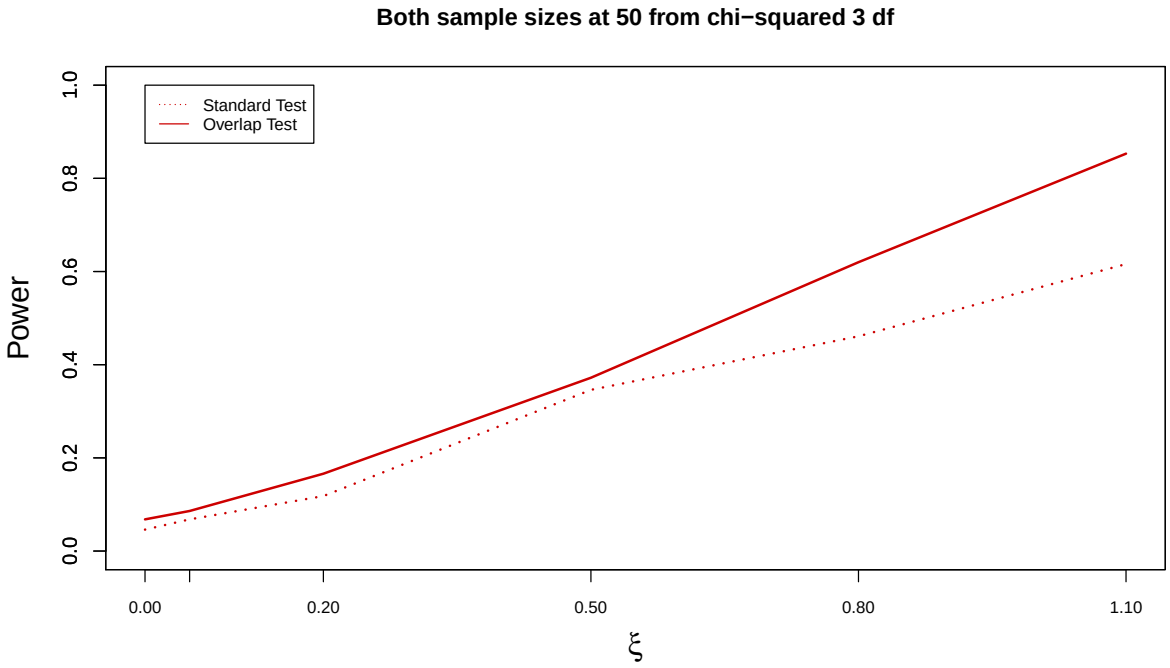

Figure 5.9: Both sample sizes at 50 from chi-squared 3 df

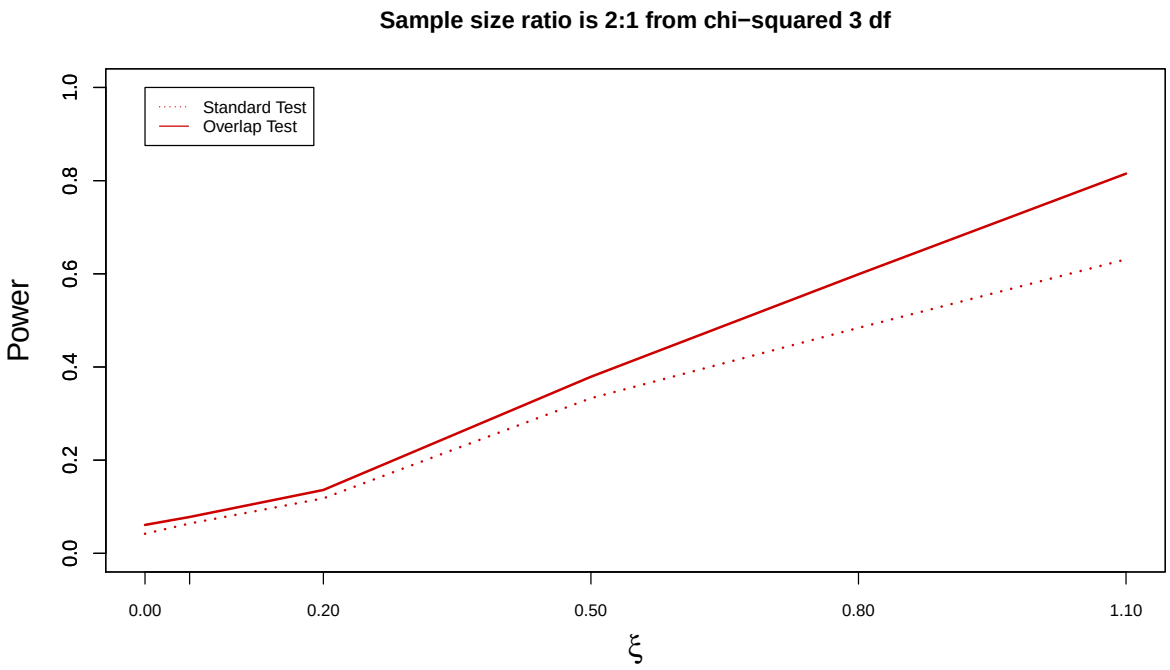

Figure 5.10: Sample size ratio is 2:1 from chi-squared 3 df

5.8 Sampling from Exponential Populations–Methods

Again, we we ran simulations when $n_R = n_E = 100$, when $n_R = n_E = 50$ and when when $n_R = 50,\ n_E = 100$. The reference is chosen from $Exp(\lambda) - \frac{1}{\lambda}$, a shifted exponential, and the experimental is chosen from $Exp(\lambda) - \frac{1}{\lambda} - \delta_{NI} + \xi$ where $\lambda = 1$. Here are the steps taken in the exponential case:

1. Specify λ, n_R, n_E, δ_{NI} and ξ.
2. Generate $X_R \sim Exp(\lambda) - \frac{1}{\lambda}$ and $X_E \sim Exp(\lambda) - \frac{1}{\lambda} - \delta_{NI} + \xi$ with sample size n_R and n_E, respectively.
3. Obtain $\hat{\Delta}$ using the kernel functions.
4. And $\hat{var}(\hat{\Delta})$, using bootstrap samples.
5. Get the test statistic, Z.
6. Repeat steps 2 through 5 over M times.
7. Obtain power, or $P(\text{Type I})$, depending on ξ, as the proportion of times the hypothesis is rejected.

5.9 SAMPLING FROM EXPONENTIAL POPULATIONS–RESULTS

Table 5.9 again has two numbers in each cell with the numbers inside the parentheses representing the traditional and the numbers outside the parentheses representing the proposed test. Just as in the chi-squared case, there is evidence the proposed test is more powerful.

On a cautionary note, one should be careful comparing too closely with the chi-squared case. There was an attempt to "smooth" the power function a bit more than previous by considering an additional point at $\xi = .3$. In the results section, the graphs will indeed appear smoother, in both the traditional and overlap power functions. One consequence of the extra information is that the gap between dotted and solid line may appear smaller than with chi-squared, but it is actually due to the added point.

$Exp(1)$	$n_R = n_E = 100$	$n_R = n_E = 50$	$n_R = 50,\ n_E = 100$
$\xi = 0$	(0.0470) 0.0585	(0.0560) 0.0565	(0.0515) 0.0470
$\xi = 0.05$	(0.1115) 0.1120	(0.0950) 0.0955	(0.0870) 0.0940
$\xi = 0.2$	(0.4235) 0.5195	(0.2710) 0.3220	(0.3030) 0.4280
$\xi = 0.3$	(0.6985) 0.8075	(0.4505) 0.5680	(0.5400) 0.7145
$\xi = 0.5$	(0.9685) 0.9945	(0.8105) 0.9045	(0.8865) 0.9740
$\xi = 0.8$	(0.9995) 1.0000	(0.9885) 0.9965	(0.9960) 0.9995

Figures 5.11 through 5.13 indicate the overlap test is more powerful than standard testing. The gap between the dotted and solid line is even more evident when the sample sizes are unequal.

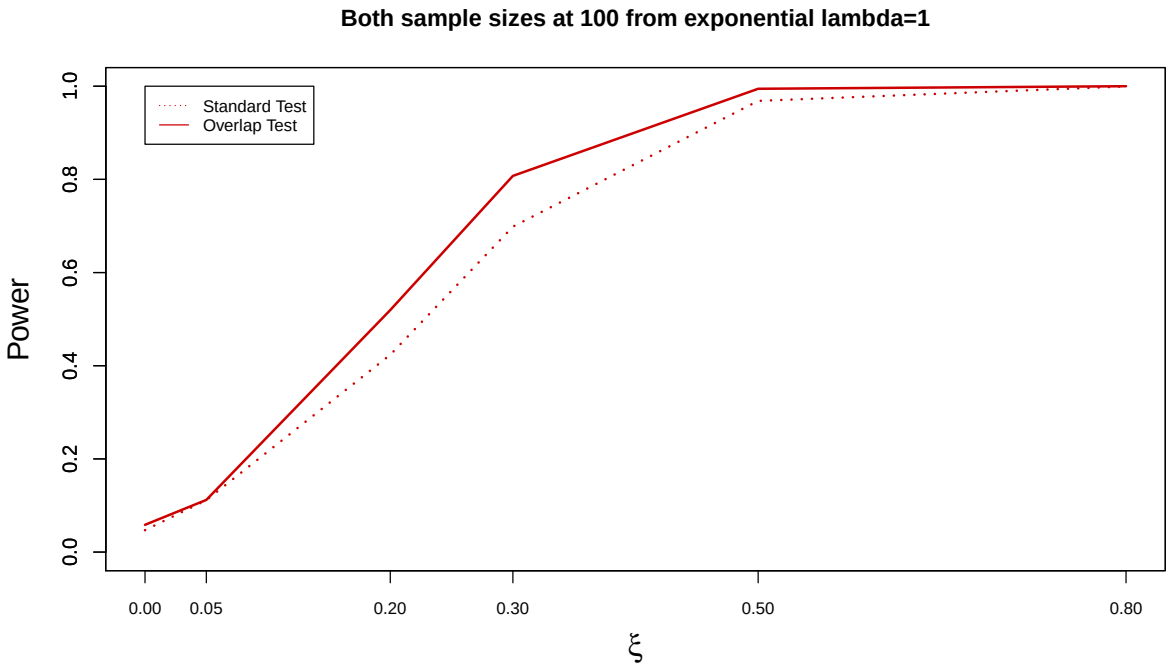

Figure 5.11: Both sample sizes at 100 from exponential, $\lambda = 1$

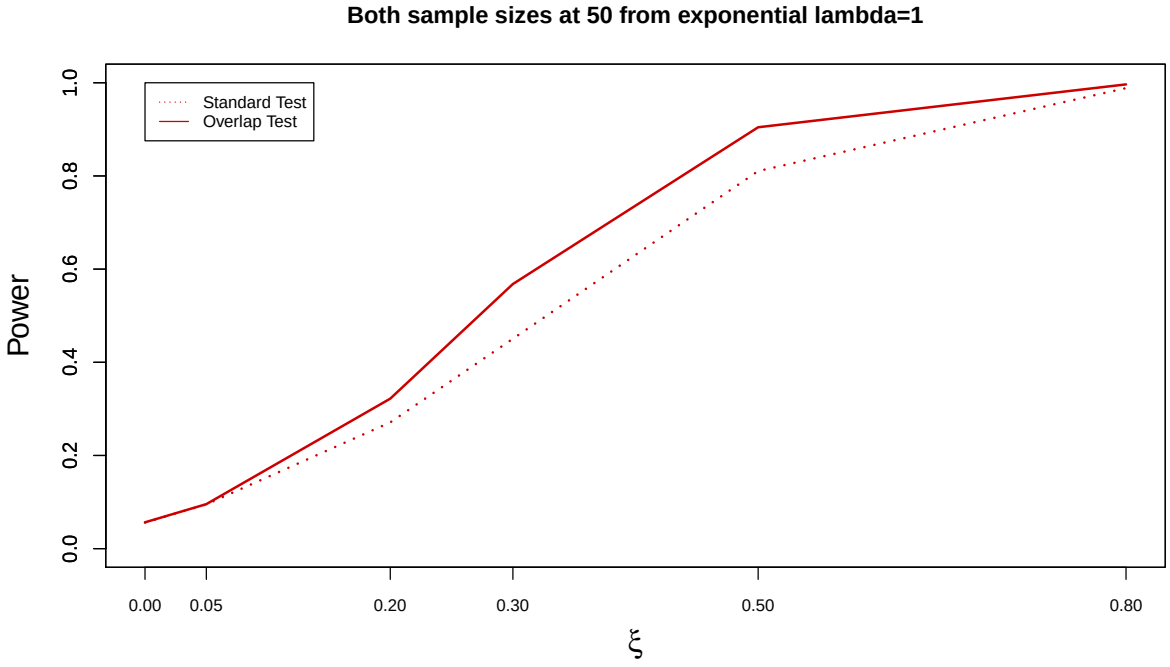

Figure 5.12: Both sample sizes at 50 from exponential, $\lambda = 1$

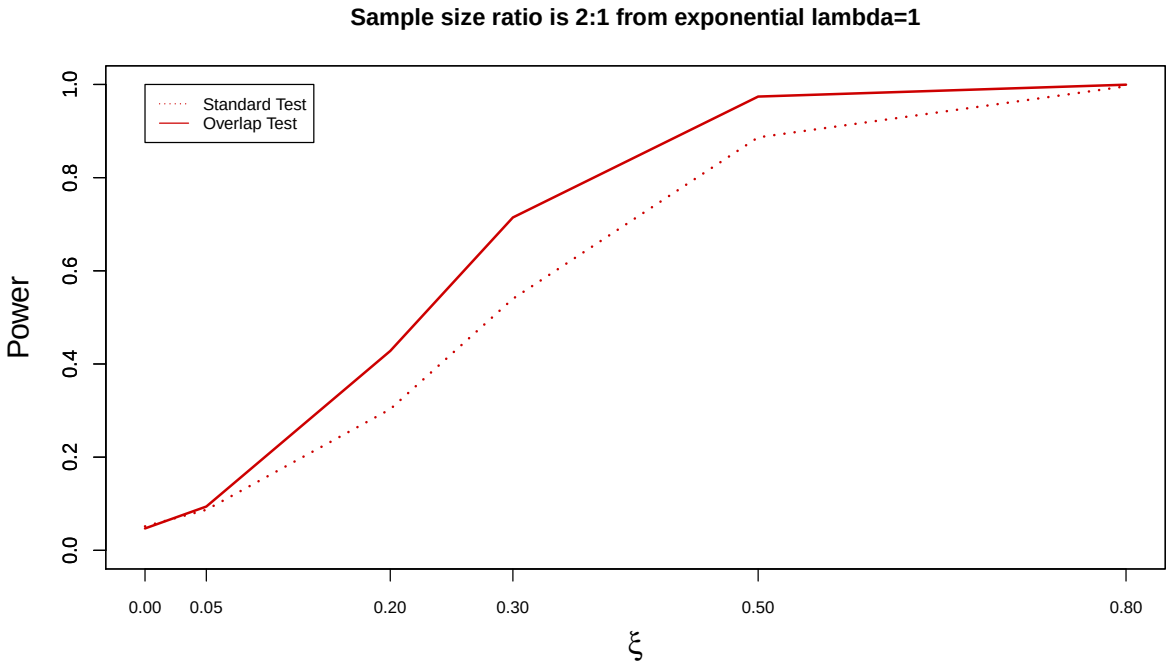

Figure 5.13: Sample size ratio is 2:1 from exponential, $\lambda = 1$

5.10 Concluding Remarks

Non-inferiority testing is a popular and widely used method of testing a claim about comparative effectiveness. Vaccines, medications, and therapies can be compared to existing ones in clinical trials, either two-armed, or three-armed with appropriate choices for the non-inferiority margin. Both frequentest and Baysian approaches are in use.

This thesis proposes a novel way of conducting such tests that requires little or no knowledge about the underlying distribution. Instead, the test relies on computing the estimated overlap from the kernel density arising from two samples. When the underlying distribution was normal with the same variance, the new test did quite well, almost as well as the standard test. However, for normal distributions with different standard deviations, the new test was less powerful the more unequal the standard deviations became. This trend held regardless of sample size considerations.

The overlap test shined brightest when we sampled from non-normal, skewed populations. Whether we sampled from a chi-squared distribution or from an exponential distribution, the overlap test showed itself to be more powerful than the standard test. This trend also held regardless of sample size considerations. It would be interesting, in the future to replicate these same simulations with other distributions or other smaller sample size.

REFERENCES

[1] *Guidance for industry non-inferiority clinical trials*, 2014, p. 2.

[2] Mohammad Al-Saleh and Hani Samawi, *Interference on overlapping coefficients in two exponential populations*, Journal of Modern Applied Statistical Methods **6** (2007), 503–516.

[3] Turki A Althunian, Anthonius de Boer, Rolf H. H. Groenwold, and Olaf H. Klungel, *Defining the noninferiority margin and analysing noninferiority: An overview*, British Journal of Clinical Pharmacology **83** (2017), 1636 – 1642.

[4] Gordon Anderson, Oliver Linton, and Yoon-Jae Wang, *Non Parametric Estimation of a Polarization Measure*, Working Papers tecipa-363, University of Toronto, Department of Economics, July 2009.

[5] Sreyan Chowdhury, Samuel Castro, Courtney Coker, Taylor E Hinchliffe, Nicholas Arpaia, and Tal Danino, *Programmable bacteria induce durable tumor regression and systemic antitumor immunity*, Nature medicine **25** (2019), no. 7, 1057–1063.

[6] FDA, *Fda guidance for industry: Non-inferiority clinical trials*, 01 2010.

[7] Alba María Franco-Pereira, Christos T Nakas, Benjamin Reiser, and María Carmen Pardo, *Inference on the overlap coefficient: The binormal approach and alternatives*, Statistical Methods in Medical Research **30** (2021), no. 12, 2672–2684, PMID: 34693817.

[8] Meg Gamalo, Ram Tiwari, and Lisa Lavange, *Bayesian approach to the design and analysis of non-inferiority trials for anti-infective products*, Pharmaceutical statistics **13** (2014).

[9] Pulak Ghosh, Farouk Nathoo, Mithat Gönen, and Ram C. Tiwari, *Assessing noninferiority in a three-arm trial using the bayesian approach*, Statistics in medicine **30** (2011), no. 15, 1794–1808.

[10] Samiran Ghosh, Erina Paul, Shrabanti Chowdhury, and Ram C. Tiwari, *New approaches for testing non-inferiority for three-arm trials with Poisson distributed outcomes*, Biostatistics **23** (2020), no. 1, 136–156.

[11] Santu Ghosh, Arpita Chatterjee, and Samiran Ghosh, *Non-inferiority test based on transformations for non-normal distributions*, Computational Statistics & Data Analysis **113** (2017), 73–87.

[12] Santu Ghosh, Wenge Guo, and Samiran Ghosh, *A hierarchical testing procedure for three arm non-inferiority trials*, Computational Statistics & Data Analysis (2022), 107521.

[13] Katherine ED Giacoletti and Joseph Heyse, *Using proportion of similar response to evaluate correlates of protection for vaccine efficacy*, Statistical Methods in Medical Research **24** (2015), no. 2, 273–286, PMID: 21865269.

[14] Wolfgang Karl Härdle, Marlene Müller, Stefan Sperlich, and Axel Werwatz, *Nonparametric and semiparametric models*, Springer Berlin, Heidelberg, 01 2006.

[15] Eisuke Hida and Toshiro Tango, *On the three-arm non-inferiority trial including a placebo with a prespecified margin*, Statistics in medicine **30** (2011), no. 3, 224–231.

[16] Meinhard Kieser and Tim Friede, *Planning and analysis of three-arm non-inferiority trials with binary endpoints*, Statistics in medicine **26** (2011), 253–73.

[17] Armin Koch and Joachim Röhmel, *Hypothesis testing in the "gold standard" design for proving the efficacy of an experimental treatment relative to placebo and a reference*, J Biopharm Stat. **14** (2004), no. 2, 315–325, PMID: 15206529.

[18] Lei Lei and Kurt Olson, *Evaluating statistical methods to establish clinical similarity of two biologics*, Journal of Biopharmaceutical Statistics **20** (2009), no. 1, 62–74.

[19] Tobias Mütze, Axel Munk, and Tim Friede, *Design and analysis of three-arm trials with negative binomially distributed endpoints.*, Statistics in medicine **35 4** (2016), 505–21.

[20] Iris Pigeot, Juliane Schäfer, Joachim Röhmel, and Dieter Hauschke, *Assessing non-inferiority of a new treatment in a three-arm clinical trial including placebo*, Statistics in medicine **22** (2003), 883–99.

[21] Hani M. Samawi, *A nonparametric test of symmetry based on the overlapping coefficient*, Journal of Applied Statistics **38** (2011), 885–898.

[22] Sachico Tanaka-Mizuno, Takuhiro Yamaguchi, Akira Fukushima, Yutaka Matsuyama, and Yasuo Ohashi, *Overlap coefficient for assessing the similarity of pharmacokinetic data between ethnically different populations*, Clinical Trials **2** (2005), 174–181.

[23] Dan Wang and Lili Tian, *Parametric methods for confidence interval estimation of overlap coefficients*, Computational Statistics & Data Analysis **106** (2016).