Power Analysis for Alternative Tests for the Equality of Means

---

A thesis

presented to

the faculty of the Department of Mathematics and Statistics

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Science in Mathematical Sciences

---

by

Haiyin Li

May 2011

Edith Seier, Ph.D., Chair

Robert M. Price, Ph.D.

Yali Liu, Ph.D.

Keywords: bootstrapping, randomization, t-test, skewness, kurtosis.

ABSTRACT

Power Analysis for Alternative Tests for the Equality of Means

by

Haiyin Li

The two sample t-test is the test usually taught in introductory statistics courses to test for the equality of means of two populations. However, the t-test is not the only test available to compare the means of two populations. The randomization test is being incorporated into some introductory courses. There is also the bootstrap test. It is also not uncommon to decide the equality of the means based on confidence intervals for the means of these two populations. Are all those methods equally powerful? Can the idea of non-overlapping t confidence intervals be extended to bootstrap confidence intervals? The powers of seven alternative ways of comparing two population means are analyzed using small samples with data coming from distributions with different degrees of skewness and kurtosis. The analysis is done using simulation; programs in GAUSS were especially written for this purpose.

DEDICATION

I would like to dedicate this work to my parents, An Li and Yi Hai, and my professors, Dr. Edith Seier, Dr. Robert Price and Dr. Yali Liu, for their continuous encouragement and support throughout my undergraduate and graduate study at East Tennessee State University.

ACKNOWLEDGMENTS

First, I would like to give my immense thanks to my advisor Dr. Edith Seier for all her support, guidance, advice, encouragement and patience throughout my four years undergraduate and graduate study at East Tennessee State University. Furthermore, a special thanks to my thesis committee members Dr. Robert Price and Dr. Yali Liu. Thank you so much for your continuous help and trust. I also would like to express my appreciation to all people in Department of Mathematics and Statistics with whom my academic journey are full of joy and happiness. Last but not the least, my gratitude goes to my parents, friends and those who have supported and helped me in my life.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Traditionally in introductory statistics courses, the means of two populations are compared using the two-sample t-test. Recently, the randomization test is making its way into such courses. This opens the discussion: is the randomization test really more powerful than the t-test when the assumptions for the t-test are not fulfilled or is it being included because it has fewer prerequisites and can be taught earlier in the semester? Instructors of introductory statistics courses are probably going to ask this question and they deserve information about the comparison of these two tests. Also, there are two common versions of the randomization test, one that uses the simple difference of the two sample means and one that calculates the t-statistic for the randomized samples. Is there really a difference between these two versions in terms of power? According to Efron and Tibshirani [5], the randomization test is not really a test for means but a test to compare distributions. Would it happen that, if the means are equal but the shapes of the distributions are very different, then we would be likely to reject the hypothesis of equal means? Efron and Tibshirani [5] defined a bootstrap test that is beyond the scope of an introductory course. However, it will be interesting to compare the power of the bootstrap test with that of the randomization test.

It is common practice among some teachers of introductory statistics to arrive at conclusions about the null hypothesis of equality of means based on confidence intervals for the mean of each one of the two populations. If the confidence intervals overlap, the null hypothesis is not rejected; while if the confidence intervals do not overlap, the null hypothesis is rejected. Schenker and Gentleman [23] analyzed the

consequences of using confidence intervals in lieu of a formal test of hypothesis. One question that comes to mind is: are the consequences in terms of power similar to those obtained by Schenker and Gentleman [23] if instead of using t-confidence intervals, bootstrap confidence intervals are used?

The focus of this work is on the behavior of power for small samples (n=5, 10 and 15) since for large samples most methods tend to behave well. Since the t-procedures assume normality, it is interesting to explore the impact of different degrees of skewness and kurtosis in the power of the different tests. The analysis of empirical power was done with data simulated with different distributions such as the uniform, Normal, SU(0.9), exponential, Laplace, lognormal, Tukey(10), and scale contaminated normal. Those distributions are described in Chapter 2. The variability of the data of course plays a role in the power of a test, the effect sizes were written in terms of the standard deviation of the distribution. Programs in Gauss were written in order to perform the simulations.

2 HYPOTHESIS TESTING METHODS

2.1 Introduction to Hypothesis Testing

Hypothesis testing, also referred as test of hypothesis or significance test, is one of the major parts of statistical inference. The procedure is used to examine whether the data constitute evidence against the null hypothesis. In introductory statistics, we emphasize statistical inference on parametric methods or sometimes called classical methods, such as z-test, t-test and Analysis of Variance (ANOVA). In case of two independent samples, the two sample t-test is the most common classical method that is based on strict distribution theories; nevertheless, it has assumptions that are hard to fulfill in many cases. In the modern world of statistics, non-parametric inferential methods are becoming more and more popular and a number of computer-intensive methods have been well developed. The most famous ones include randomization tests, bootstrap and Monte Carlo methods (Manly [18]). Another alternative to hypothesis testing is to examine overlapping confidence intervals. Although this method has limitations, it is relatively efficient and easier under certain circumstances.

Hypothesis testing involves Type I errors (Reject a true H_0) and Type II errors (Not reject a false H_0). It is known that if the probability of making one type of error is reduced, simultaneously, the probability of making the other error will increase. Our goal is to choose an appropriate significance level in order to control Type I error. However, it is important to be aware that the probability of making the Type I error in practice is not always equal to the theoretical significance level, i.e not always the real α is equal to the nominal value α of the test. If the probability of Type I error

is equal to the assigned significance level, then we say the hypothesis test is exact, otherwise, the test is either conservative or liberal. Certain assumptions should be held in order to make a test exact, i.e a one sample t-test is exact only if data come from a normal distribution.

In order to evaluate the efficiency of the hypothesis testing method, it is necessary to calculate the power of the test. The power of a statistical test is defined as the probability of rejecting a false null hypothesis and it can also be calculated as $1 - P(Type\ II\ error)$ [13]. The power of the test depends on the difference between the two population means, as well as the significance level being used. In introductory statistics courses, it is already emphasized that for a fixed significance level, the power of the test increases as the sample sizes increase. High power indicates the statistical test is highly efficient. In this study, the focus is on the comparison of different hypothesis tests under the condition of small samples.

2.2 Two Sample T-Test

The t-test is formally called 'Student's t-test' in honor of the famous British statistician William Sealy Gosset whose pseudonym was 'Student' [26]. Gosset [26] introduced the t statistic in 1908 and the probability distribution he derived was called t-distribution or Student's t probability model. In case of the two sided t-test, the null and alternative hypotheses are written as:

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

Before performing the two sample t-test, usually these three assumptions should be

checked:

1. Two independent samples are randomly selected from two distinct populations.

2. Both of the populations are normally distributed, which is called the assumption of normality.

3. The two populations have similar variances.

In order to compute the test statistic, let $\bar{x}_1$ and $\bar{x}_2$ be the two sample means respectively and n_1 and n_2 be the corresponding sample sizes. Learning from solving the one sample case, without knowing the two population standard deviations, we replace the population standard deviations σ_1 and σ_2 by the sample standard deviations s_1 and s_2. The two sample t-statistic is written as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}} \tag{1}$$

Under the assumption of equal variances, the pooled variance ${s_p}^2$ is used to estimate the unknown population variance rather than ${s_1}^2$ and ${s_2}^2$ because the pooled estimator in equation (3) is based on a larger sample ($n_1 + n_2$ observations) than ${s_1}^2$ or ${s_2}^2$ separately. In this case, the t statistic is written as

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2}$$

where

$$s_p = \sqrt{\frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}} \tag{3}$$

and degrees of freedom $= n_1 + n_2 - 2$.

On the other hand, if the consistency or equal variance assumption is dropped, Welch's t-test [13] is an adaption of the two sample t-test by using expression (4) and approximating the degree of freedom from the data,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}} \tag{4}$$

with degrees of freedom approximated by the integer part of

$$df = \frac{(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2})^2}{\frac{{s_1}^4}{{n_1}^2(n_1 - 1)} + \frac{{s_2}^4}{{n_2}^2(n_2 - 1)}} \tag{5}$$

Expression (5) is called Welch-Satterthwaite equation [13].

2.3 Overlapping T Confidence Intervals

Statistical significance is often associated with confidence intervals. For example, consider the two sample t-test with the null hypothesis and alternative hypothesis stated as in Section 2.2. For each of the population means, a confidence interval can be calculated as follows:

$$\bar{x}_1 \pm t_1^* \frac{s_1}{\sqrt{n_1}}$$
$$\bar{x}_2 \pm t_2^* \frac{s_2}{\sqrt{n_2}}$$

Here, the t critical value t_1^* and t_2^* can be found in the t table or can be computed using software.

In introductory statistics courses, it is sometimes said when the two confidence intervals for the two population means are not overlapping, we can conclude that the

two means are not the same. In other words, the null hypothesis should be rejected. However, failure to reject the null hypothesis using overlapping confidence intervals does not necessarily imply failure to reject H_0 using the corresponding hypothesis test. Schenker and Gentleman [23] concluded that due to the conservatism and low power relative to the standard methods, the overlap method should not be used as a formal significance testing unless this is the only option for the data analyst.

2.4 Randomization Tests

The randomization test is considered as one of the great revolutions in statistics in the twentieth century and it is becoming one of the major approaches in statistical education. The basic idea of randomization was first introduced by Sir Ronald A. Fisher in 1923 and was greatly explained by him again in his book on experimental design [8]. Now it is widely applied in data analysis, especially in biological sciences and health sciences. The basic idea behind the randomization test is to generate an empirical distribution of the statistic of interest by regrouping data from the original samples. Depending on different situations, there are two main probability models for explaining the probability basis for statistical inference in randomization tests. Lehmann [16] called the first model the *randomization model* in which the available subjects are randomly assigned to treatments. The second, the *population model,* is used when subjects are randomly sampled from different populations. Edgington [4] pointed out that the methods used for the randomization model are randomization tests, while the same methods used for the population model are called permutation tests. The two names, randomization and permutation, are frequently considered

interchangeable in the hypotheses testing context. However, there is an important distinction between two models in the design of a study. The randomization model is used in the design of experiments and the population model in the design of observational studies.

Compared to classical methods, the main advantages of randomization tests are that they can be applied even without random samples and that they are almost distribution free. Thus, randomization tests are relatively less restrictive than the classical methods such as the t-test. However, the limitations of the randomization tests are obvious, too. First, they are only applicable to the comparison of two or more populations. Second, by its nature, a randomization test can only tell us whether a certain pattern of data is likely to have arisen purely by chance. Therefore, a randomization test can only test whether populations are equal instead of arriving at conclusions about the values of the parameters of populations.

In a two sample case, the null hypothesis and alternative hypothesis are

$$H_0 : F = G$$

$$H_a : F \neq G$$

If the null hypothesis is true, we can consider these two populations are really a single population. There are four steps to conduct the randomization tests for the difference between two population means as described below.

1. Find the true difference D_1 between two sample means.

2. Put the two samples together and then randomly reallocate n_1 elements as the first new group and the remaining n_2 as the second group. The difference, D,

between these two groups can be now obtained.

3. Repeat the random selection in step 2 a large number of times, for example ten thousand times in the approximated version of the test. In the exact test, all the possible re-groupings of $n_1 + n_2$ observations are considered. For each re-grouping obtain the difference between the two reallocated groups. Keeping the difference between the means of the two groups for each re-grouping, an empirical distribution for the differences is obtained by randomization.

4. Use the empirical distribution to obtain the achieved significance level and arrive at a conclusion about the null hypothesis.

If the true difference, D_1, is a value that looks extremely large or small based on the empirical distribution obtained by randomization, we reject the null hypothesis. Alternatively, we calculate the achieved significance level, defined by Efron [5]. Similar to the p-value in classical test, the achieved significance level of the test, abbreviated ASL, is the probability of observing at least that large a value when the null hypothesis is true.

$$ASL = Prob_{H_0}(|D| \geq |D_1|)$$

If the ASL is very small, the null hypothesis would not seem reasonable and the alternative hypothesis would be preferred. Otherwise, the allocation in reality seems to be random and we do not reject the null hypothesis.

The interesting thing is that there is more than one way to obtain the test statistic of the randomization test, some of which are equivalent. Instead of using the difference

between the two sample means, the pooled two sample t statistic could be used and it will give the exactly the same result as using difference between two means. Some statisticians such as Manly [18] and Ernst [6] pointed out that the sum of the responses in one group is often used as the statistic rather than the t statistic and the difference in means because it is computationally more efficient as the two sample sizes increase.

How many random re-groupings of the original data should be considered? The exact randomization test requires us to do all the possible regrouping of the data:

$$\binom{n_1 + n_2}{n_1}$$

Randomization tests are exact [5], which means the probability of making a Type I error is always equal to the defined significance level. For example, if there are two samples with ten subjects each, then there will be $\binom{20}{10} = 184756$ possible re-arrangements of the twenty individuals in two groups of ten individuals each. In introductory statistics, we probably would prefer not to perform all possible reallocations. The number of re-groupings can be reduced to a certain level while keeping the significance level estimated close to the nominal value of the exact significance level. This type of randomization test is called the approximate randomization test. In this study, the approximate randomization test is used in the simulations.

2.5 Testing with Bootstrap Confidence Intervals

Schenker [23] compared the t-test with the overlapping t-confidence intervals. Bootstrapping [5] is an alternative way of building confidence intervals. A new comparison to be done is that of the randomization test with overlapping bootstrap confi-

dence intervals. It would be interesting to see whether or not the relationship between the randomization test and overlapping bootstrap confidence intervals is similar or not to the relationship, between the t-test and overlapping t-confidence intervals, found by Schenker [23].

In statistics, resampling means sampling from the sample. Bootstrap is a method defined by Efron in 1980 [5]. The bootstrap method relies only on an empirical distribution obtained from the data in order to do inference about the parameters. Resampling is a very practical method to obtain such an empirical distribution. To do resampling is to select random samples from the original sample. The new samples are called "bootstrap samples" [5] and they are of the same size as the original sample. The statistic of interest is calculated for each bootstrap sample in order to obtain an empirical bootstrap distribution for the statistic. The bootstrap method is usually a good choice when the assumptions necessary for more classical methods are not fulfilled or when extremely complicated calculations were necessary. The biggest difference from the sampling point of view between bootstrap and the randomization test is whether to sample with replacement or without replacement.

There are several ways of building confidence intervals based on the bootstrap empirical distribution. The percentile bootstrap confidence interval was described by Efron in his earlier work [5]. This method is also referred as "the first percentile method" or "simple percentile confidence limits". The two bounds are the values that encompass the central 100(1-α)% of the bootstrap empirical distribution. Other improved percentile methods also exist, such as the second percentile method by Hall [18]. Furthermore, some better confidence intervals have also been defined [5], such

as the bootstrap t-confidence interval, the accelerated bias-corrected percentile limits (BCa intervals) and the approximate bootstrap confidence interval (ABC). In this study, the percentile bootstrap confidence intervals and the bootstrap t-confidence intervals will be used in the simulations because those are the bootstrap type tests most likely to be used in an introductory statistics course.

2.6 Testing with Bootstrap Test

The bootstrap test is not as popular as the bootstrap confidence intervals. Fisher and Hall [9] and Hall and Wilson [12] point out that an important difference between bootstrap confidence intervals and the bootstrap test is the accuracy of the estimators of the critical values for the test statistic. The basic procedure of the bootstrap test is as follows:

1. Standardize the observations z_i and y_i of the two samples

$$z\prime_i = z_i - \bar{z} + \bar{x}$$

$$y\prime_i = y_i - \bar{y} + \bar{x}$$

   where $\bar{z}$ and $\bar{y}$ are group means and $\bar{x}$ is the mean of the combined sample.

2. Regroup the two groups of the standardized values 1000 times. Each time, obtain the sample means $\bar{z^*}$ and $\bar{y^*}$ and standard deviations ${s_1}^2$ and ${s_2}^2$.

3. For each re-grouping, compute the t statistic using the formula below

$$t^* = \frac{\bar{z^*} - \bar{y^*}}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

4. Approximate ASL by

$$ASL = \#\{t^* \geq t_{obs}\}/B$$

In this study, the power of the bootstrap test is compared with the power of other methods using simulation as shown in the results section.

3 ENVIRONMENT OF POWER COMPARISON

3.1 Statistical Power

Statistical power is the probability of rejecting the null hypothesis when the null hypothesis is false. As power increases, the probability of type II error (not rejecting Ho when Ho is false) will decrease. Statistical power is usually associated with sample size and the effect size. The larger the sample sizes and the effect sizes are, the more powerful the statistical test will be. When the probability of one of the error decreases, the other will increase. The change of the probabilities of these two types of errors is never in the same direction. The nominal value of the significance level, α, is set before the test is performed, and 0.05 is a common value. The statistical power is usually a criterion to judge and compare different statistical tests. However, better alpha control should also be considered. All the procedures mentioned in the previous section will be compared pairwise, as shown in Figure 1, by plotting the values of α = P(Type I error) and power.

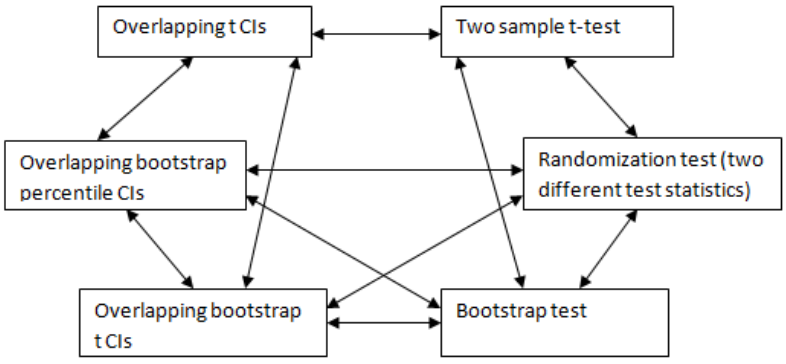


Figure 1: Comparison of different methods

3.2 Different Skewness Environments

In probability theory and statistics, skewness is defined as a measure of asymmetry of the probability density function of a real-valued random variable. The value of skewness is negative when the distribution is left skewed and a positive value indicates the distribution is right skewed; and when the distribution is symmetric, the value of skewness is zero. There are many different ways to calculate skewness. The skewness coefficient was originally defined by Pearson using the third central moment $\gamma_1 = \mu_3/\sigma^3$ where $\mu_3 = E(X-\mu)^3$ [2]. There are also more simple ways of measuring skewness such as (mean-mode)/standard deviation, for the Pearson's mode or first coefficient of skewness, and (mean-median)/standard deviation, for the Pearson's median or second coefficient of skewness.

For inference about the population mean, skewness is one of the factors that should be taken into consideration since it can affect the power of the statistical test and lead to a misleading result. For example, the t procedures always emphasize the assumption of normality. The one sample t-test and the one sample t-confidence interval are robust enough under mild skewness when the sample size is greater than 15. However, the impact of skewness will be significant when the sample size is smaller or equal to 15. In cases of severe skewness, we need much a larger sample size to apply the one sample t procedures. The question is how skewness can affect the power of the two sample test when sample sizes are smaller than or equal to 15. For instance, such skewed distributions as exponential and lognormal distributions are widely used in the analysis of survival data.

3.3 Different Kurtosis Environments

In 1905, Karl Pearson originally defined the "degree of kurtosis" $\eta = \beta_2 - 3$, where $\beta_2 = \mu_4/m_2^2$ and μ_i is the i^{th} moment with respect to the mean, as the measure of peakedness in order to compare the distribution of a real-valued random variable to the normal distribution [24]. Balanda and Mac Gillivray [1] pointed out that kurtosis should not only be related to peakedness but also tails of the distribution by saying that kurtosis could be understood as "the location- and scale-free movement of probability mass from the shoulders of a distribution into its center and tails". Now, the representation of kurtosis as Pearson's coefficient β_2 for both peak and tails is more broadly accepted by statisticians and widely used in various statistics books [2]. Statisticians have defined several measures to quantify kurtosis and proposed different approaches of studying kurtosis [24]. The understanding of kurtosis is not restricted to β_2. For example, one simple way for introducing kurtosis to students is visualizing the peak and tails of an unimodal distribution to a uniform distribution with the same median and variance proposed [15]. Kurtosis can affect the performance of inferential tools, especially with regard to inference about the variance. Also the median is a more efficient estimator of center than the mean for symmetric distributions when the kurtosis is high. For small samples, the behavior of some tests will be weakened by high kurtosis of the distributions. Results obtained by simulation with regard to the power of tests in the presence of high kurtosis are included in the results section.

4 PROBABILITY DISTRIBUTIONS

The following eight distributions are considered in this study.

4.1 Normal Distribution

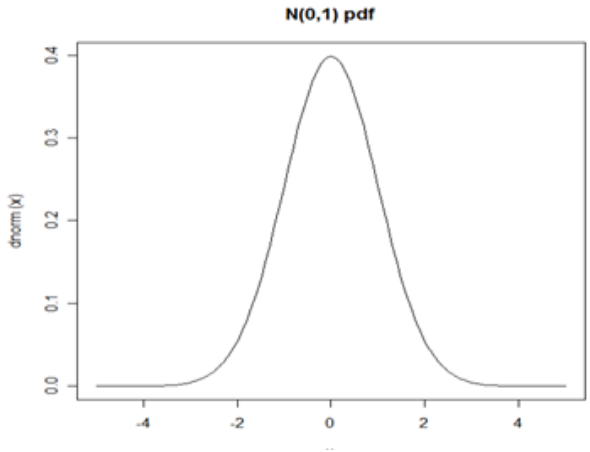


Figure 2: Normal Distribution N(0,1)

The normal (or Gaussian) distribution plays an extremely important part in statistics. The revolution of theoretical statistics started at the beginning of the twentieth century, however, the normal distribution appears earlier in history [2]. First it appeared in connection to the binomial distribution and later it was used to represent the distribution of errors [13]. Furthermore, Central Limit Theorem according to which the normal distribution serves as the basis of practical statistical work. It is also widely used as an approximation to other distributions. If the random variable X has a normal distribution, then the probability density function (pdf) is

given by:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad for \;\; x \in (-\infty, \infty) \tag{6}$$

For an arbitrary normal distribution N(μ, σ), the mean and variance are μ and σ^2. The standard normal distribution N(0,1) (Fig. 2) is defined as a specific normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$.

4.2 Uniform Distribution

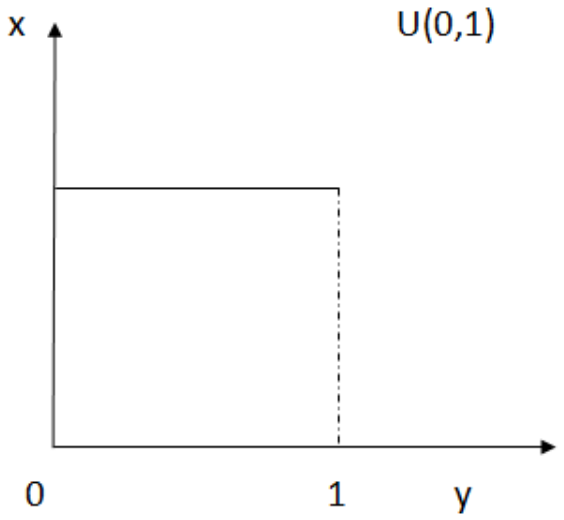


Figure 3: Uniform Distribution U(0,1)

The uniform distribution, also called rectangular distribution, refers to both the continuous and discrete cases. In this study, only the continuous uniform distribution is considered. The probability density function of the uniform distribution U(a,b) is

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a \leq x \leq b \\ 0 & otherwise \end{cases} \tag{7}$$

For an arbitrary uniform distribution U(a,b), the mean and variance are $\dfrac{a+b}{2}$ and $\dfrac{(b-a)^2}{12}$.

In particular, the *standardized* uniform distribution is the uniform distribution which has mean 0 and standard deviation 1 and the *standard* uniform distribution (Fig. 3) is the uniform distribution over (0,1). The relationship between standardized uniform distribution and standard uniform distribution is that if X has a standard uniform distribution, then $Y = \sqrt{3}(2X - 1)$ has a standardized uniform distribution.

4.3 Lognormal Distribution

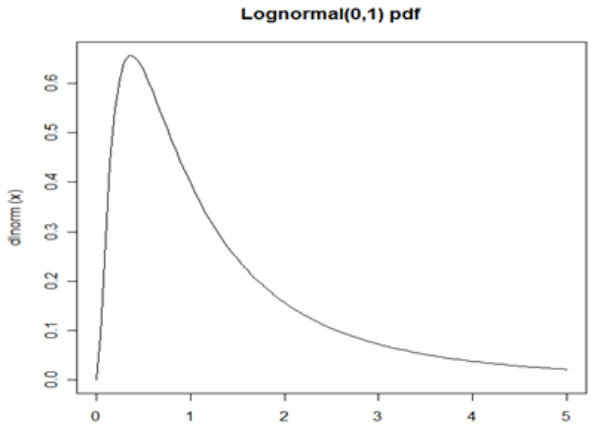


Figure 4: Lognormal Distribution Ln(0,1)

The normal distribution is considered as the logarithmic transformation of the lognormal distribution, that is, if a random variable Y has a lognormal distribution with parameters μ and σ, then the variable $X = log(Y)$ has a normal distribution

with mean μ and standard deviation σ. The notation Ln(μ,σ) or Log(μ, σ) is used to represent the lognormal distribution. Another way of expressing this relationship is to say that if X $\sim N(\mu, \sigma^2)$, then Y $= e^X$ has a lognormal distribution, $Y \sim Ln(\mu, \sigma^2)$. The parameters of the lognormal distribution are μ and σ. The lognormal distribution is extremely skewed to the right and it is widely used as a typical model in survival analysis. The mean, median and mode of the lognomal distribution Ln(μ, σ^2) are $\exp(\mu + 0.5\sigma^2)$, $\exp(\mu)$, and $\exp(\mu - \sigma^2)$, respectively. However, their estimators are biased and inefficient and this motivates statisticians to seek for different ways of estimating the parameters [17]. Usually, the logarithmic transformation is applied in the generalized linear model context in order to fulfill the normality assumption. The specific lognormal distribution, Ln(0,1) (Fig. 4), will be used in the simulations.

4.4 Exponential Distribution

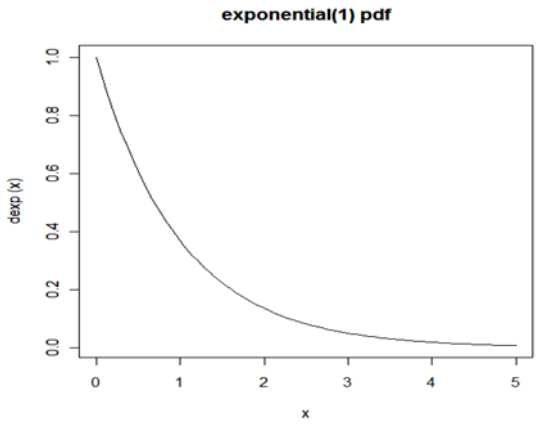


Figure 5: Exponential Distribution Exp(1)

A random variable X has an exponential distribution if its probability density

function is of the form:

$$f(x) = \frac{1}{\sigma} e^{-\frac{(x-\theta)}{\sigma}}, x > \theta; \sigma > 0 \tag{8}$$

When $\theta = 0$ and $\sigma = 1$, we call the exponential distribution 'standard exponential distribution' (Fig. 5) which has the mean 1 with probability density function:

$$f(x) = e^{-x}, x > 0. \tag{9}$$

4.5 SU Johnson Distribution

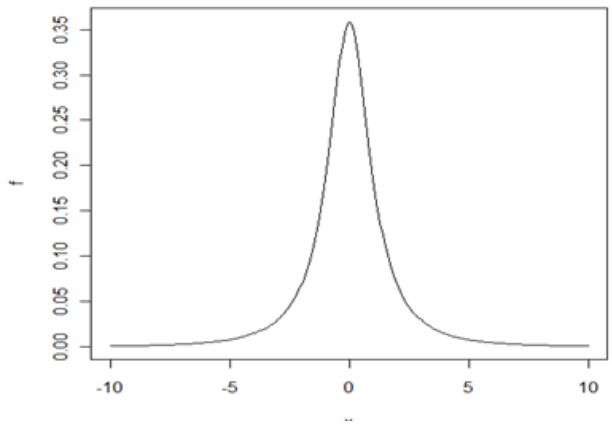


Figure 6: SU Johnson Distribution SU(0.9)

Johnson et al. [14] described the transformations

$$Z = \gamma + \delta log(X - \xi), \qquad X \geq \xi, \tag{10}$$

$$Z = \gamma + \delta log(\frac{X - \xi}{\xi + \lambda - X}), \qquad \xi \leq X \leq \xi + \lambda, \tag{11}$$

$$Z = \gamma + \delta sinh^{-1}(\frac{X - \xi}{\lambda}). \tag{12}$$

The distribution of Z is the standard normal distribution. Equation (10) corresponds to the family of lognormal distributions. For the other of two, the type of distribution depends on the range of X. If X is bounded, then the family of distributions in equation (11) is called S_B, otherwise, the symbol S_U is used. The four parameters of the S_U distribution are γ, δ, ξ and λ. In this study simulations are done with the SU(0.9) (Fig. 6) which has high kurtosis and has parameters $\gamma = 0$, $\delta = 0.9$, $\xi = 0$ and $\lambda = 1$.

4.6 Laplace Distribution

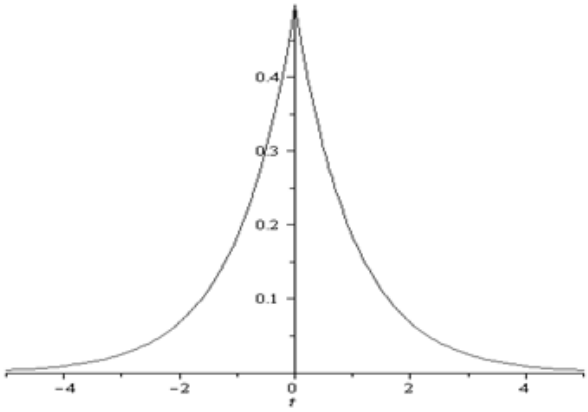


Figure 7: Laplace Distribution Laplace(0,1)

The Laplace distribution was defined by Pierre Laplace (1774) and is known under several names [14]: two-tailed exponential, bilateral exponential and the most common one, the double exponential distribution. The Laplace distribution is symmetric and has higher kurtosis than the normal distribution. The general version of the probability density function of Laplace distribution is

$$f(x) = \frac{1}{2\theta} e^{-\frac{|x-\phi|}{\theta}} \tag{13}$$

The Laplace distribution with $\phi = 0$ and $\theta = 1$ (Fig. 7) will be used in the simulation.

4.7 Tukey(λ) Distribution

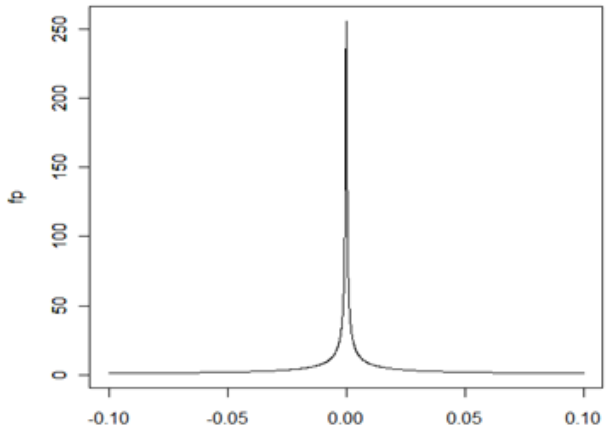


Figure 8: Tukey Distribution Tukey(10)

The Tukey(λ) family of distributions, sometimes also called Tukey distributions, are defined as transformed distributions. Let the variable U be standard uniformly distributed with the following probability density function

$$f(u) = 1, 0 < u < 1 \tag{14}$$

It is said that the variable X has a Tukey(λ) distribution if it is defined as

$$X = \frac{U^\lambda - (1-U)^\lambda}{\lambda}, \qquad -\lambda^{-1} \leq X \leq \lambda^{-1}, \qquad \lambda > 0 \tag{15}$$

and

$$X = log(\frac{U}{1-U}), \qquad \lambda = 0 \tag{16}$$

Tukey(10) (Fig. 8) denotes the symmetrical Tukey lambda distribution with $\lambda = 10$. This is the distribution to be used in the simulations because it is a challenging environment for inferential tools due to its short range of values for the variable and extreme peakedness.

4.8 Scale Contaminated Distribution

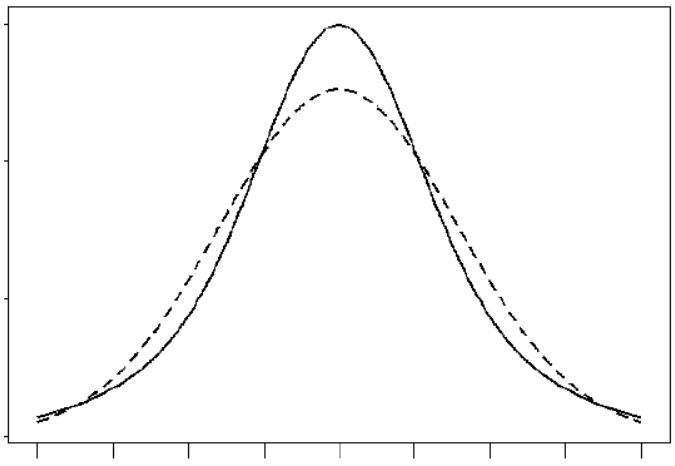


Figure 9: Scale Contaminated Distribution ScCon(5,0.1)

The scale contaminated distribution ScCon(a,p) denotes the mixture of a standard normal distribution N(0,1) with probability (1-p) and a normal distribution N(0,a) with probability p. The probability density function has a very complicated

form. However, it is a symmetric distribution with $\mu = 0$ and $\sigma^2 = (1-p)+pa^2$. The scale contaminated distribution ScCon(5, 0.1) (Fig. 9) will be used in the simulation.

4.9 Summary of distributions

Table 1 displays the summary of the eight distributions which will be used for simulation purposes.

Table 1: Summary of Distributions

Distribution	Mean	Standard Deviation	Skewness	Kurtosis
N(0,1)	0	1	0	3
U(0,1)	0.5	0.29	0	1.8
Tukey(10)	0	0.031	0	5.38
Lapalce(0,1)	0	1.414	0	6
SU(0.9)	0	2.328	0	82.1
ScCon(0.1,5)	0	1.844	0	16.5
Exp(1)	1	1	2	9
Ln(0,1)	1.65	2.16	6.18	113.9

5 SIMULATION RESULTS

5.1 Summary of Data

The null hypothesis and alternative statistical hypothesis for two-sided tests are

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

In this study, only small samples with equal or unequal sizes are considered. Therefore, the first sample size is fixed to be 10 whereas the second sample has the size 5, 10 and 15. The effect sizes for power comparison are set as 0.5, 1 and 2. For the randomization test, bootstrap confidence intervals and the bootstrap test, the most simple versions that can be taught in an introductory statistics course is used. For each sample, 1000 bootstrap subsamples were obtained or 1000 random re-groupings were done. Ten thousand simulations were used to obtain the estimated ASL and the statistical power for each method.

5.2 Summary of Comparisons

Twelve pairs of the tests are to be compared and all the results of pairwise comparisons are shown in the following nine sections.

1. Randomization test using the difference of means vs. randomization test using the t-statistic.

2. Randomization test vs. two sample t-test.

3. Bootstrap test vs. two sample t-test.

4. Bootstrap test vs. randomization test.

5. Overlapping t confidence intervals vs. two sample t-test.

6. Overlapping bootstrap confidence intervals vs. randomization test.

7. Overlapping bootstrap confidence intervals vs. bootstrap test.

8. Overlapping bootstrap confidence intervals vs. overlapping t confidence intervals.

9. Overlapping bootstrap percentile confidence intervals vs. overlapping bootstrap t confidence intervals.

5.3 Randomization Test Using the Difference of Means and Using T-Statistics

As we mentioned in Section 2, the randomization test can be conducted using different statistics and some of them are equivalent. For example, the difference between the sample means or the pooled t-test statistic could be used. According to Ernst [6], the randomization test, Ran(d), using the difference in means always agrees with the randomization test using the pooled t-test statistic. Instead of using the pooled t-test statistic, the randomization test using Welch's t statistic (Ran(tw)) will be compared in the simulations with the randomization test using the difference of means. Figure 10 displays the results for the estimated significance level and the power of those two types of randomization tests.

The simulation results show that the two randomization tests always agree when sample sizes are the same. However, this is not necessarily true when the two samples

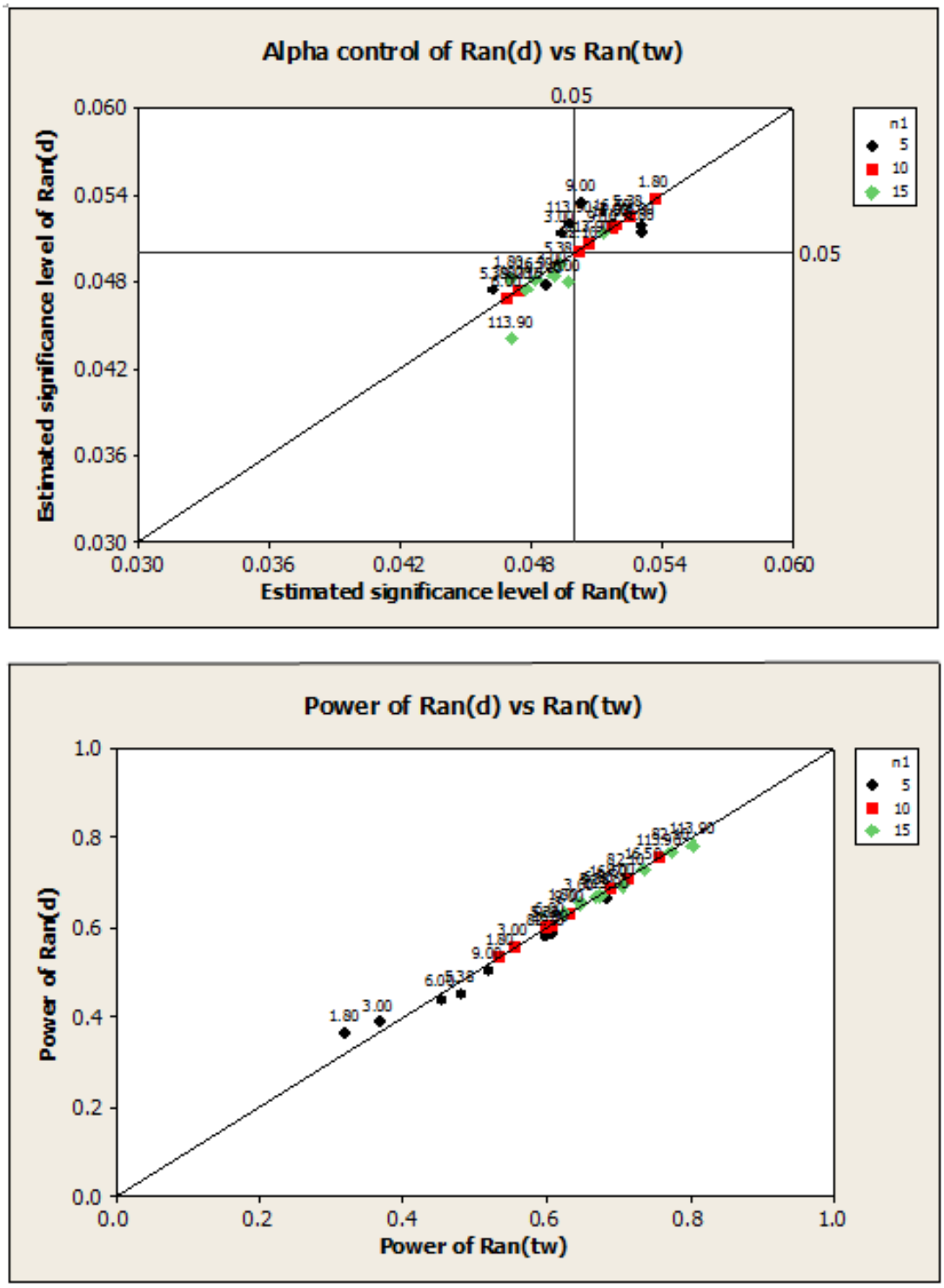


Figure 10: Randomization Tests with Different Statistics

are simulated with different sample sizes. On the other hand, the empirical significance levels are not far away from the nominal value 0.05. It seems that when one of the sample sizes is extremely small, the randomization test that uses the difference of the means as the test statistic works better. The lognormal distribution, that has kurtosis 113.9 is an outlier in Figure 10 with regard to alpha control when n_1=15. That is, its empirical significance level is relatively far away from 0.05 for both versions of the randomization test. The randomization test may have problems with strongly skewed data. The two versions of the randomization test almost agree in terms of power except for some rare cases, i.e. when at least one of the samples is too small. Overall for the randomization test, using the difference of means works better than using Welch's t statistic as the test statistic. Therefore, for the following analysis, we will focus on performing the randomization test using the difference of means as the test statistic.

5.4 Randomization Test and Two Sample T-Test

The Central Limit Theorem about the approximately normal distribution of the sample mean is an asymptotic result. In introductory statistics courses, it is frequently said that an approximately normal distribution can be assumed for sample means when sample sizes are larger than 15 or 30. The t-test will be relatively robust in most situations. However, we are interested in the analysis of alpha control and power for sample sizes smaller than 15, where the t-test may not be an optimal method. Figure 11 indicates that although the randomization test does not show a distinct advantage in terms of power, it has a better alpha control. It should be remembered

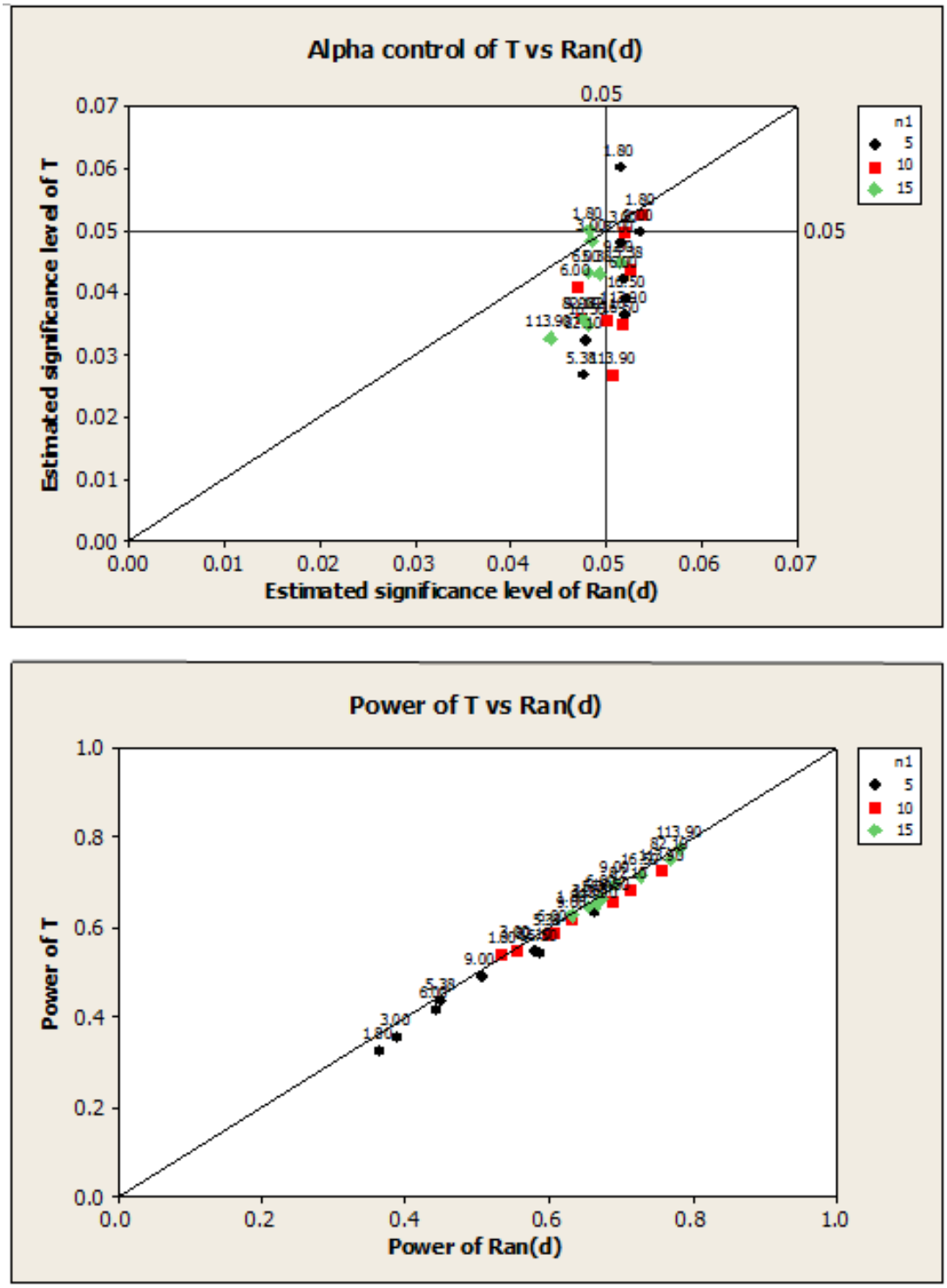


Figure 11: Randomization Test and Two Sample T-Test

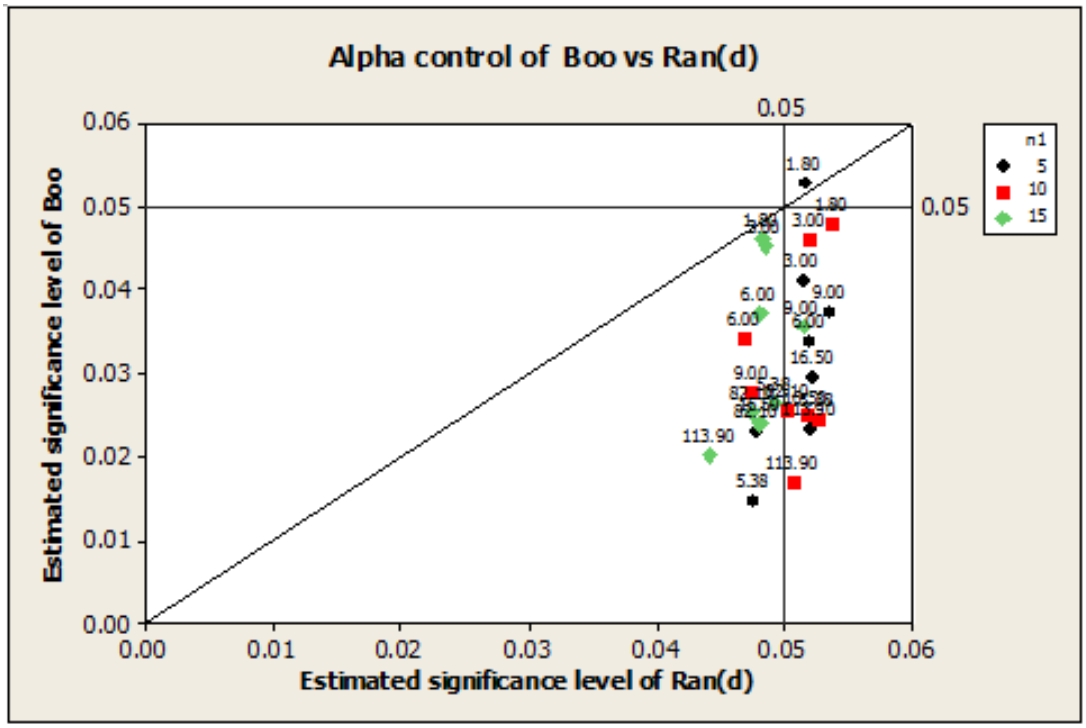
that in this study the exact version of the randomization test is not being applied, only 1000 re-groupings of the original samples were done. The exact randomization test (when all the possible re-groupings are done) always has its empirical significance level equal to the nominal significance level. Therefore, if all the possible permutations are included, the estimated significance level will be exactly 0.05.

5.5 Randomization Test and Bootstrap Test

There has been discussion in the statistical literature that the randomization test is not so much a test for equal means as a test for equal distributions [5]. A bootstrap test was defined by Efron and Tibshirani [5] as a possible replacement for the randomization test. However, the bootstrap test is more complicated to apply and has not made its way into introductory statistics courses yet. Figure 12 indicates that, although the bootstrap test and the randomization test can both be considered Monte Carlo methods, there is an overwhelming advantage in using the randomization test. The randomization test has both better alpha control and higher power than the bootstrap test. Randomization tests are exact when all the possible re-groupings of the samples are considered, which is not true for bootstrap tests.

5.6 Bootstrap Test and Two Sample T-Test

The limitations of the bootstrap test can be clearly seen from Figure 13. The t-test is not always appropriate in dealing with small samples when the data come from a non-normal distribution. However, the performance of the bootstrap test is even worse.



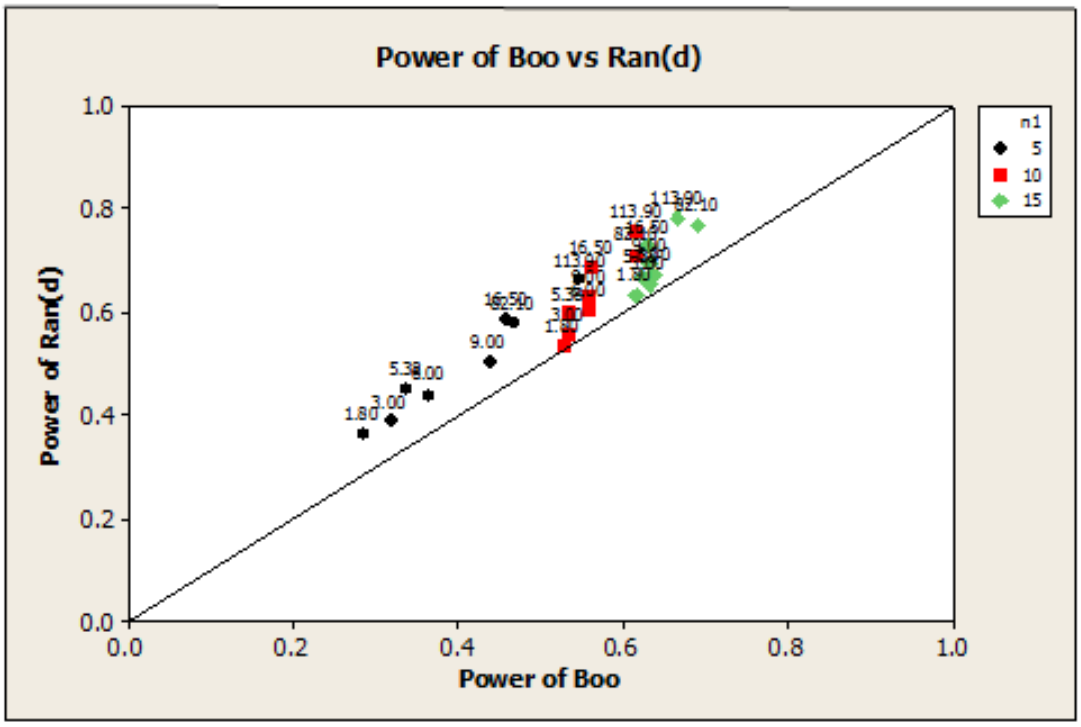


Figure 12: Randomization Test and Bootstrap Test

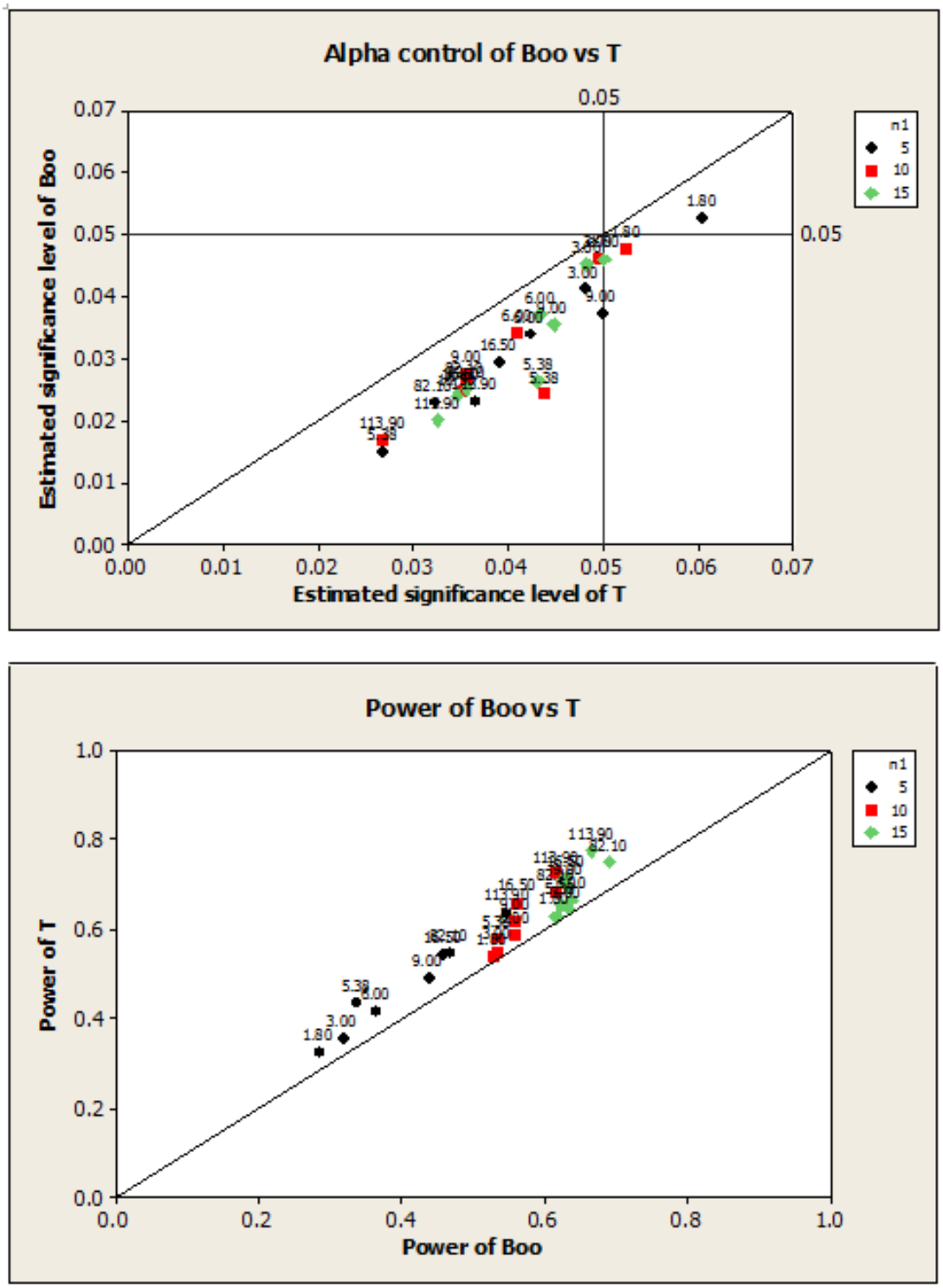


Figure 13: Bootstrap Test and Two Sample T-Test

5.7 Overlapping T Confidence Intervals and Two Sample T-Test

The comparison done in Figure 14 verifies Schenker's idea in judging the significance using overlapping confidence intervals and the corresponding test of hypothesis method [23]. The overlapping t confidence intervals method tends to be very conservative based on the same significance level as the corresponding two sample t-test. Statisticians have explored if adjusting the confidence level for each confidence interval helps in achieving the nominal significance level 0.05. For example, Payton, Greenstone and Schenker [20] proposed to adjust the confidence level of each confidence interval to 84% for large samples, instead of the usual 95%. According to Payton, Greenstone and Schenker [20], the adjusted significance level for each t confidence interval is associated with the ratio of the two standard errors.

Table 2 shows the results for the estimated significance levels when using 84% overlapping t confidence intervals in the case of small samples. The achieved significance levels are smaller than the expected value 0.05. However, this can be an starting point to look for more appropriate confidence levels in order to achieve the desired significance level.

In the statistical literature, only the t-test and the t-confidence intervals have been compared. We wanted to extend this comparison to other types of tests and confidence intervals. The 84% overlapping percentile and t bootstrap confidence intervals have been checked. The average empirical significance level for overlapping bootstrap t confidence intervals is around 0.06. For the percentile bootstrap confidence interval the empirical significance level is much larger.

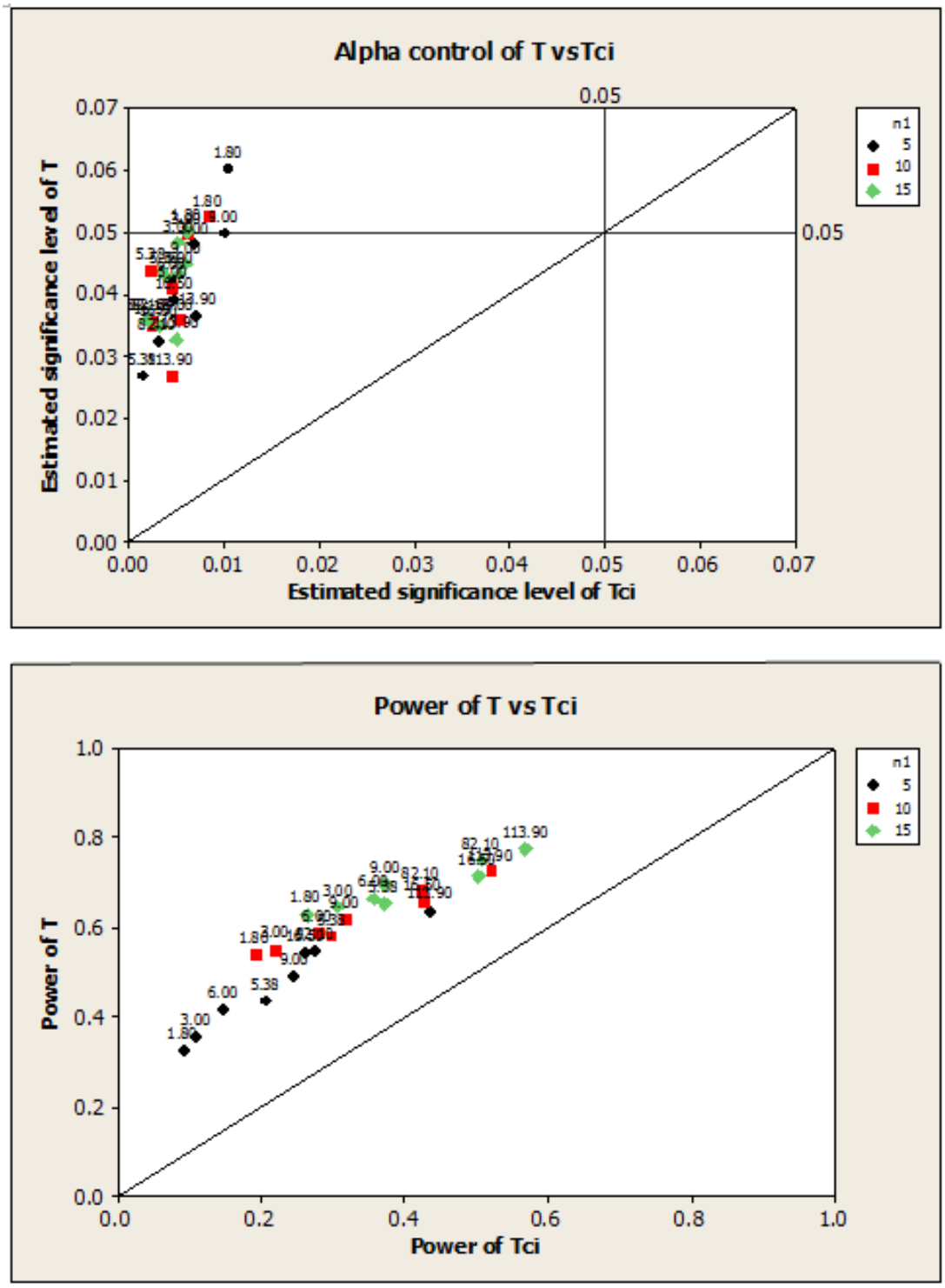


Figure 14: Overlapping T Confidence Intervals and Two Sample T-Test

Table 2: Simulation Results for Adjust Significance Levels

Distribution	n1	n2	84% t CIs	84%B(p)	84%B(t)
normal	10	10	0.0486	0.0817	0.0601
uniform	10	10	0.0496	0.0808	0.0613
lognormal	10	10	0.0533	0.1078	0.0667
exponential	10	10	0.0502	0.0890	0.0622
laplace	10	10	0.0416	0.0840	0.0539
tukey10	10	10	0.0486	0.1004	0.0625
SU0.9	10	10	0.0408	0.0897	0.0553
ScCon(0.1,5)	10	10	0.0387	0.0922	0.0531
normal	5	10	0.0488	0.1109	0.0699
uniform	5	10	0.0596	0.1098	0.0763
lognormal	5	10	0.0604	0.1408	0.0796
exponential	5	10	0.0659	0.1355	0.0851
laplace	5	10	0.0455	0.1192	0.0666
tukey10	5	10	0.0366	0.1256	0.0559
SU0.9	5	10	0.0375	0.1143	0.0562
ScCon(0.1,5)	5	10	0.0426	0.1203	0.0607
normal	15	10	0.0467	0.0719	0.0564
uniform	15	10	0.0480	0.0707	0.0573
lognormal	15	10	0.0547	0.0965	0.0646
exponential	15	10	0.0572	0.0903	0.0686
laplace	15	10	0.0444	0.0779	0.0536
tukey10	151	10	0.0464	0.0866	0.0569
SU0.9	15	10	0.0405	0.0835	0.0510
ScCon(0.1,5)	15	10	0.0388	0.0829	0.0493

5.8 Overlapping Bootstrap Percentile Confidence Intervals and Overlapping Bootstrap T Confidence Intervals

These two bootstrap confidence intervals are popular in statistics courses, however, there is a difference in their construction. The percentile method selects as the endpoints of a $(1-\alpha)100\%$ confidence interval, two quantiles of the empirical distribution (obtained by re-sampling) for the sample mean. The quantiles selected are those that occupy the $m \times \alpha/2$ and $m \times (1-\alpha/2)$ (where m is the number of bootstrap samples generated by resampling) positions, once the values of the bootstrap sample means have been ordered. The bootstrap t confidence interval is

$$\bar{x} \pm t_{\alpha/2} \times boostrap \;\; standard \;\; error \quad .$$

The standard error is calculated as the standard deviation of the means of the bootstrap samples.

The simulation results summarized in Figure 15 indicate that with regard to alpha control, both of them are far away from the nominal value 0.05 when 95% confidence intervals are used. However, the percentile method looks relatively better and has higher power than the overlapping bootstrap t confidence intervals in the case of small samples.

5.9 Overlapping Bootstrap Confidence Intervals and Overlapping T Confidence Intervals

In this section both types of bootstrap confidence intervals (percentile and t-bootstrap) are being compared with the traditional t-student confidence intervals. The construction of the bootstrap confidence intervals is described in the previous

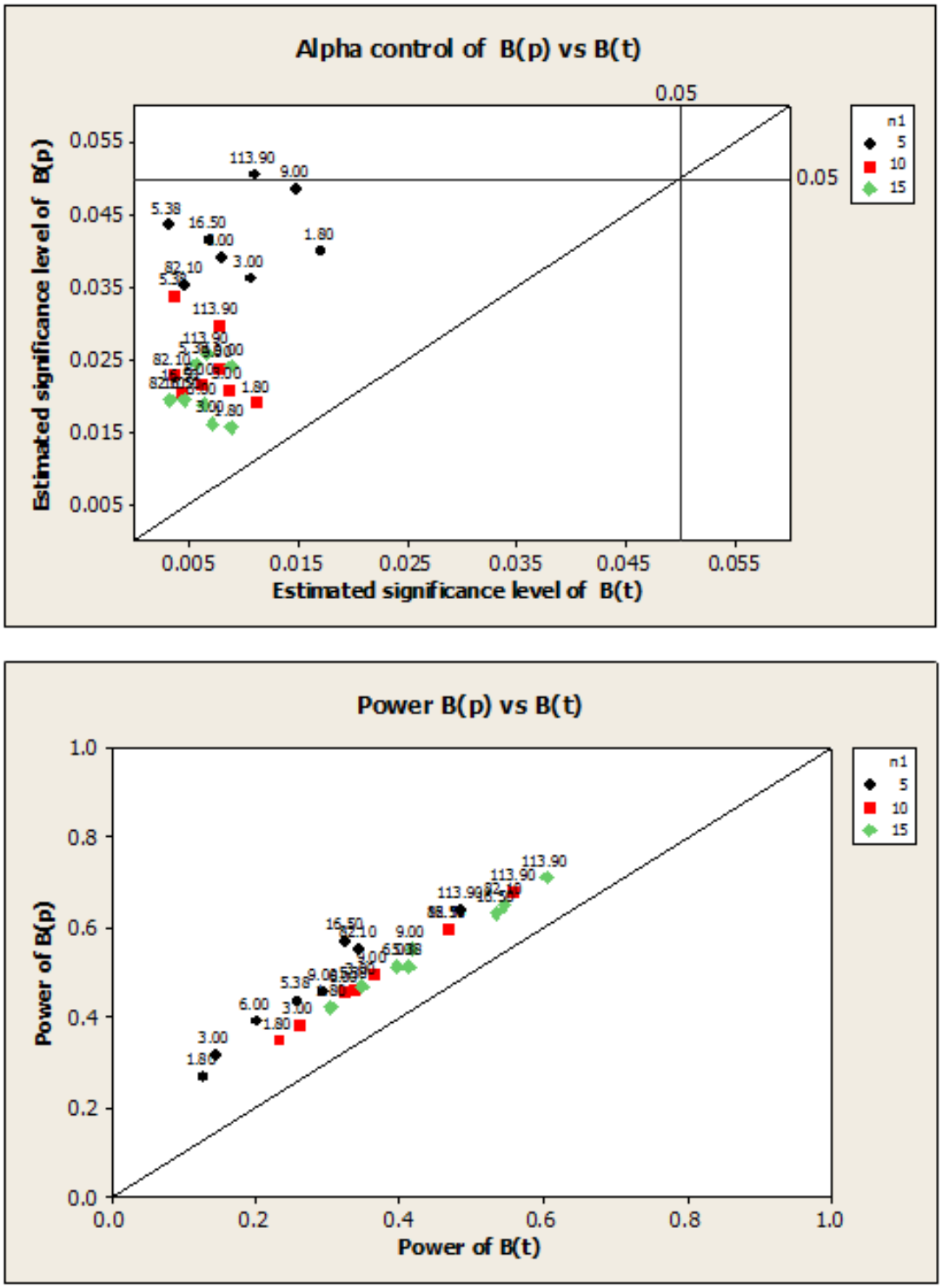


Figure 15: Overlapping Percentile and t Bootstrap Confidence Intervals

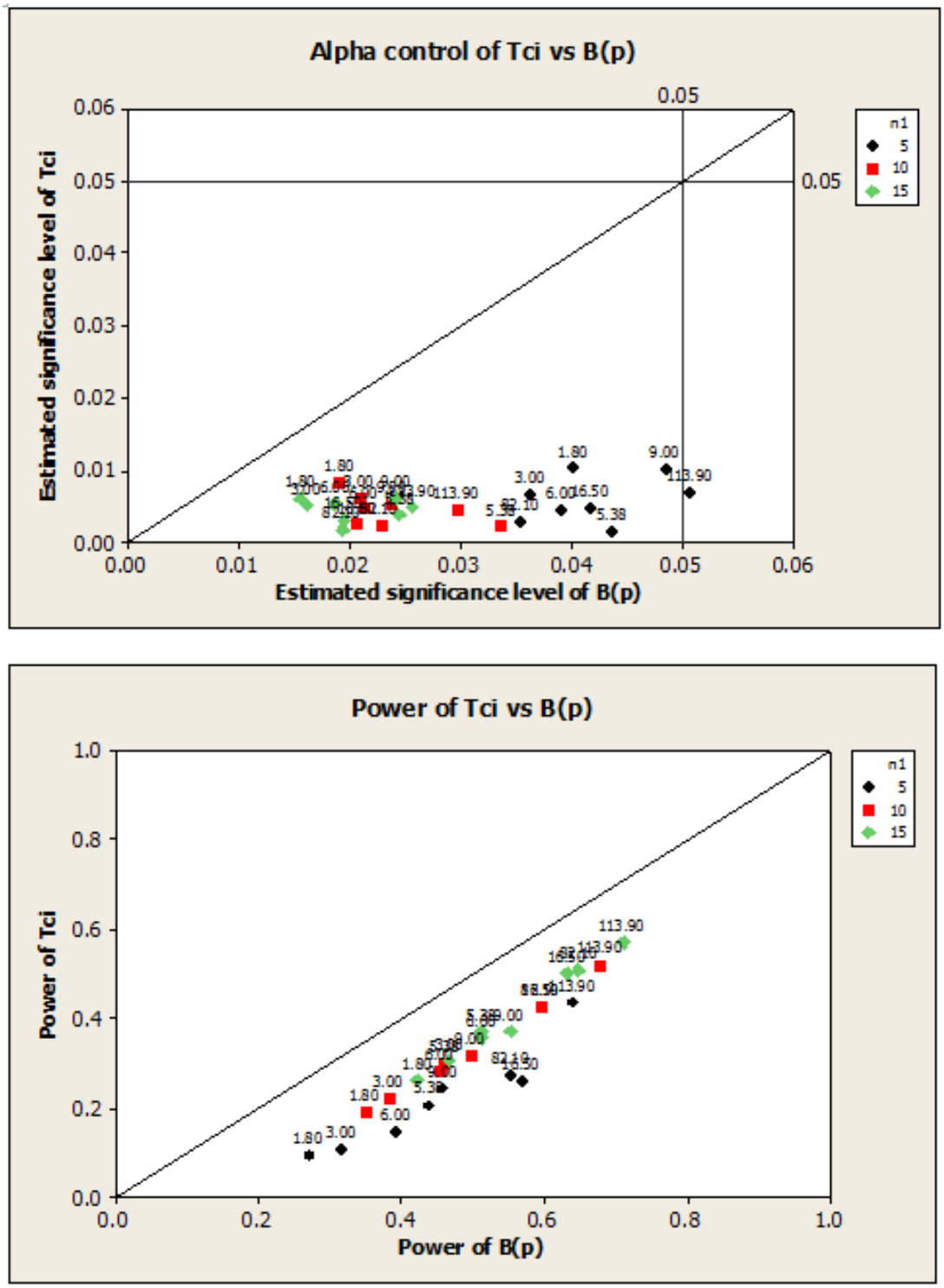


Figure 16: Overlapping Bootstrap Percentile and T Confidence Intervals

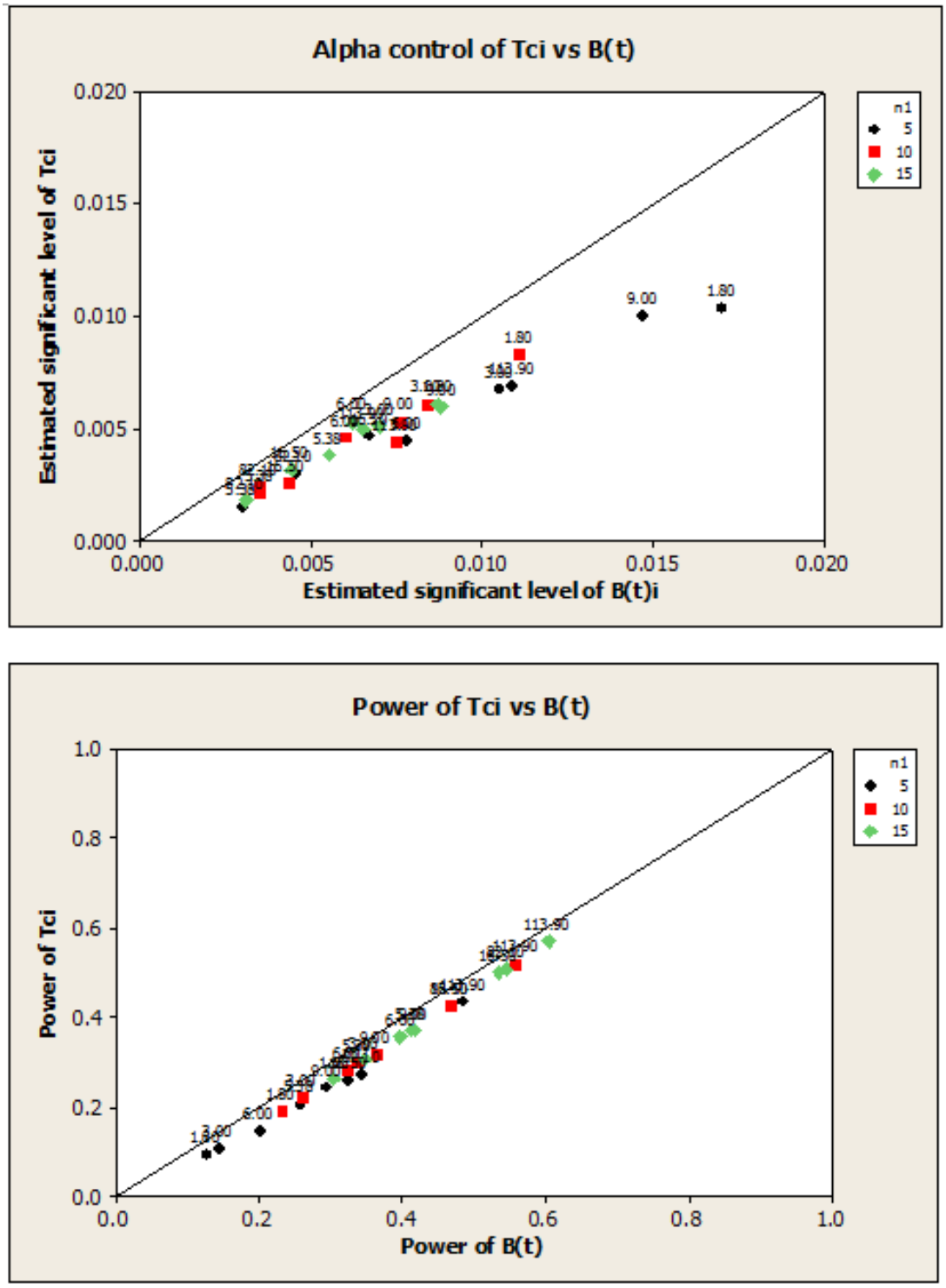


Figure 17: Overlapping Bootstrap T and T Confidence Intervals

section. The t-student confidence interval is calculated as

$$\bar{x} \pm t_{\alpha/2} \times \frac{s}{\sqrt{n}}$$

where s is the standard deviation of the sample. When the confidence intervals for the two samples do not overlap, the null hypothesis of equal population means is rejected.

Figures 16 and 17 summarize the simulation results. The conclusion is that the overlapping bootstrap t-confidence intervals work better than the classic t-confidence intervals, both alpha control and power. Among the three overlapping confidence intervals, the overlapping bootstrap percentile confidence intervals method is best while the overlapping classic t-confidence intervals method produces the worse results.

5.10 Overlapping Bootstrap Confidence Intervals vs Randomization Test

In this section the overlapping bootstrap confidence intervals (both percentile and bootstrap-t) are compared to the randomization tests. The simulation results are summarized in Figures 18 and 19. The results indicate that the method of overlapping bootstrap confidence intervals is conservative (the empirical α is below the nominal α). This is the same pattern observed for overlapping t confidence intervals as compared to the t-test. Unfortunately, the results also show that Schenker's idea [23] of reducing the confidence level of the overlapping intervals in order to achieve the nominal value of α is not applicable to the overlapping bootstrap confidence intervals in order to achieve the same α than the randomization test. Although both bootstrap and randomization tests are based on the idea of resampling, they use different strategies to obtain new samples - sampling with replacement for the first one and

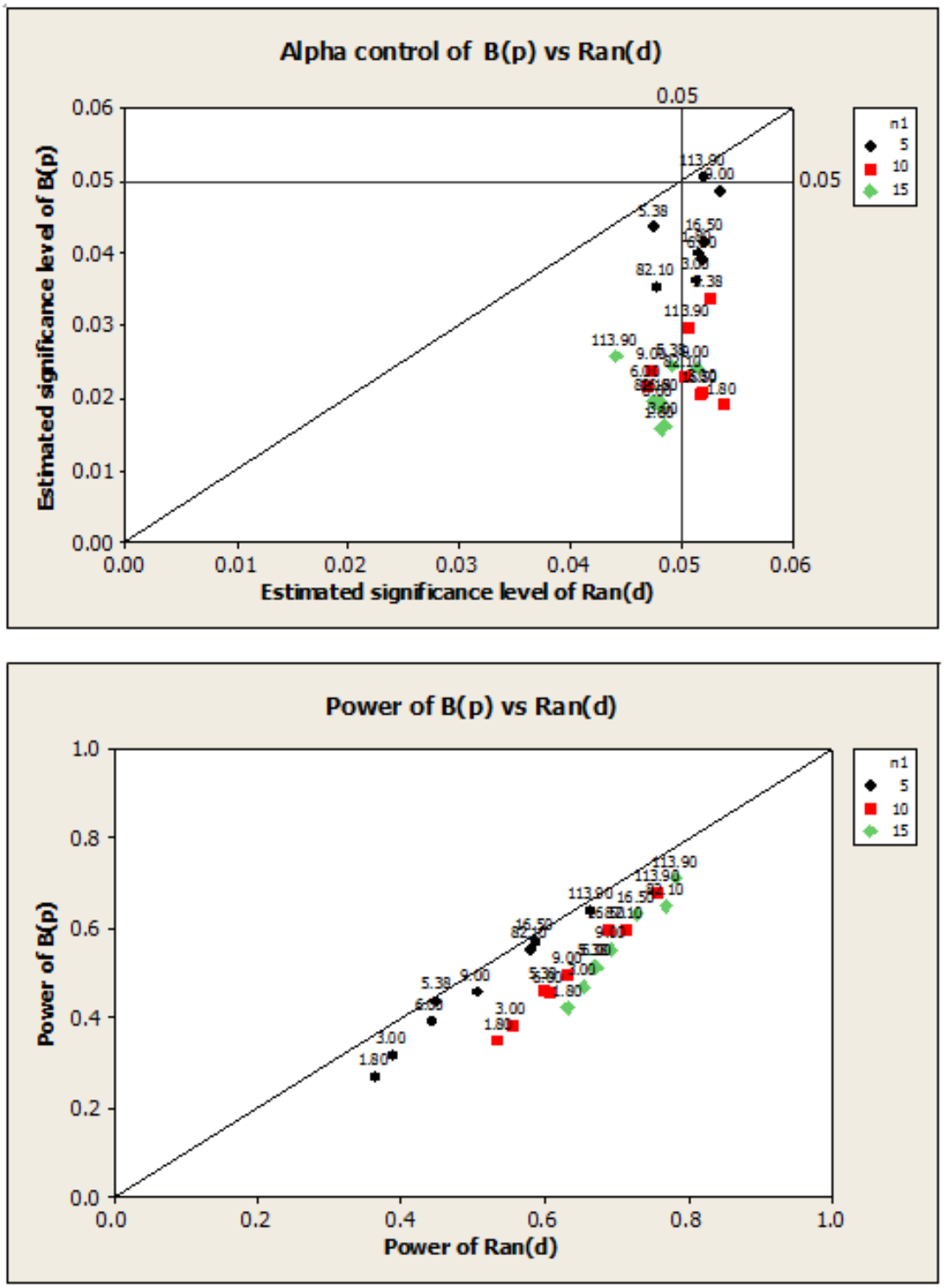


Figure 18: Overlapping Bootstrap Percentile Confidence Intervals and Randomization Test

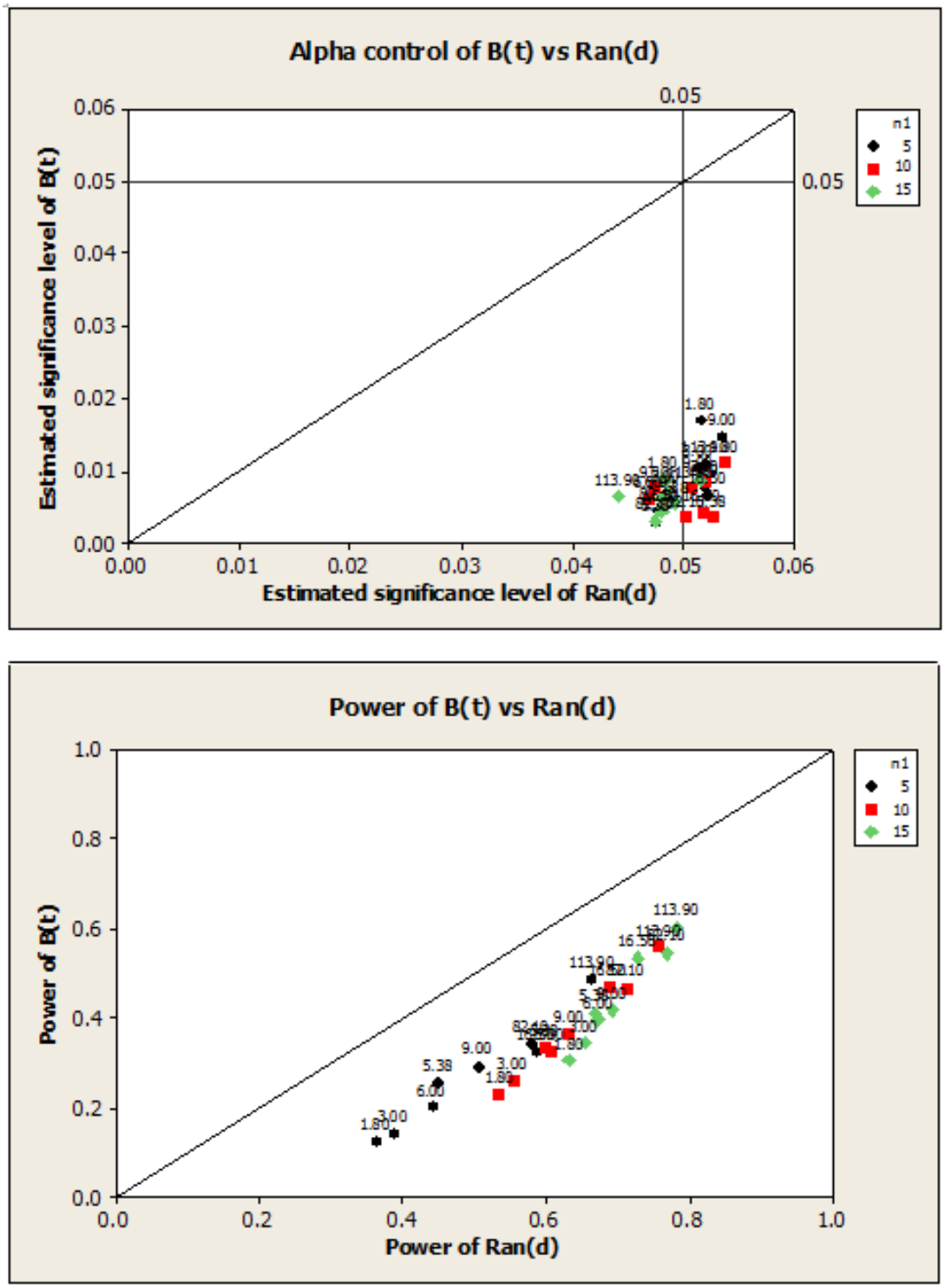


Figure 19: Overlapping Bootstrap T Confidence Intervals and Randomization Test

sampling without replacement for the latter one.

5.11 Overlapping Bootstrap Confidence Intervals vs Bootstrap Test

In this section the bootstrap test is being compared to the method of using overlapping bootstrap confidence intervals to arrive at a conclusion about the null hypothesis of equal means. In section 5.5 the bootstrap test was found to be more conservative and have lower power than the randomization test for small samples. However, according to the simulation results summarized in Figures 20 and 21, the bootstrap has a better performance than the method that uses overlapping bootstrap confidence intervals to judge whether to reject the null hypothesis or not. Although both bootstrap confidence intervals and the bootstrap test use exactly the same re-sampling methods, the bootstrap test works better. Our simulation results also indicate that Schenker's idea [23] of reducing the confidence of the intervals to achieve the desired value of α when testing hypotheses is not applicable to the bootstrap test.

5.12 Other Simulation Results

In addition to the results explained in the previous sections, the agreement or disagreement between different methods was also studied. For example, if the two t-confidence intervals do not overlap, then the two bootstrap percentile confidence intervals will not overlap either. For two population means, $\mu_1 < \mu_2$, the upper bound of the bootstrap percentile confidence interval for μ_1 is always smaller than the upper bound of the t confidence interval. On the other hand, the lower bound

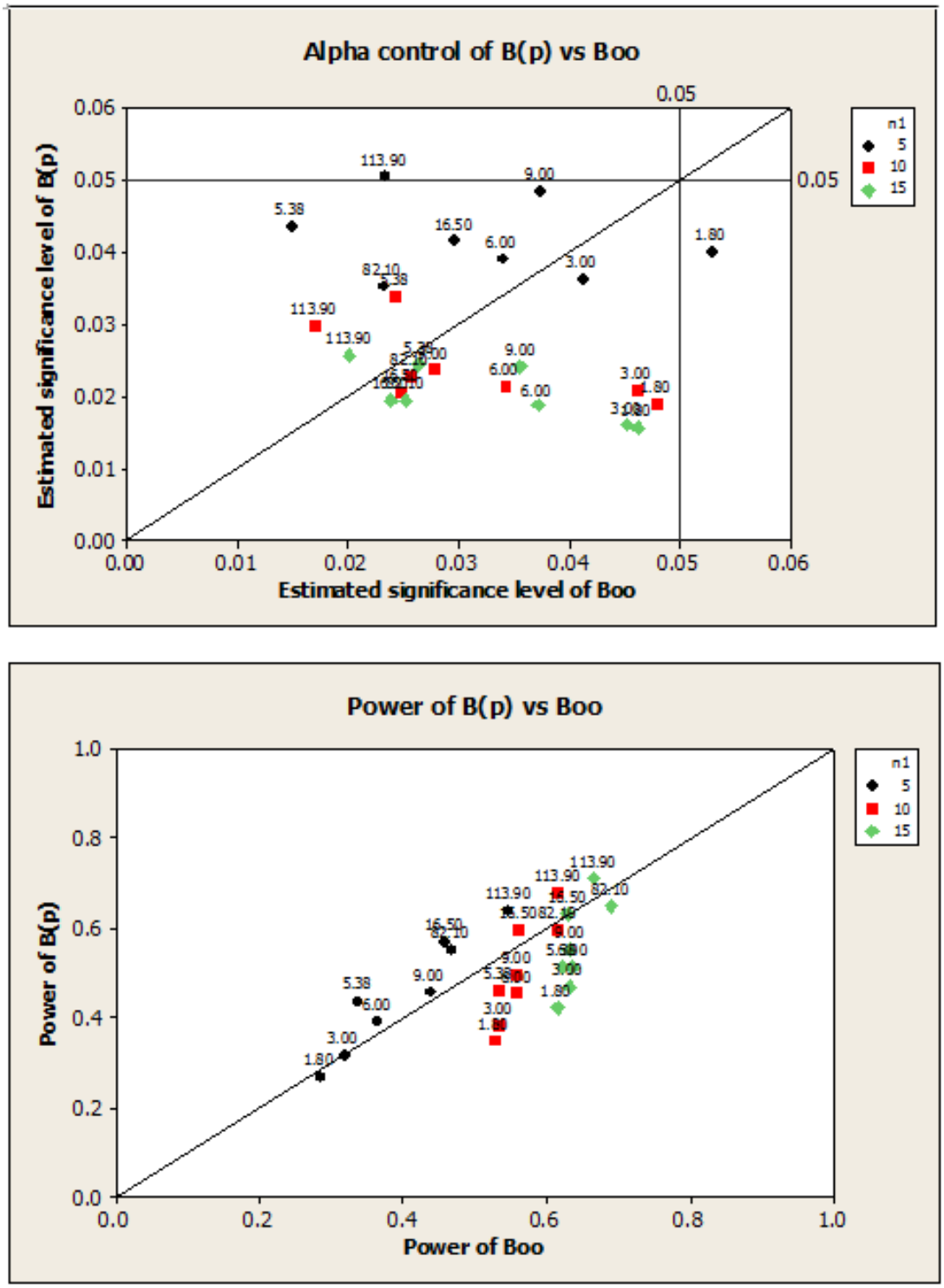


Figure 20: Overlapping Bootstrap Percentile Confidence Intervals and Bootstrap Test

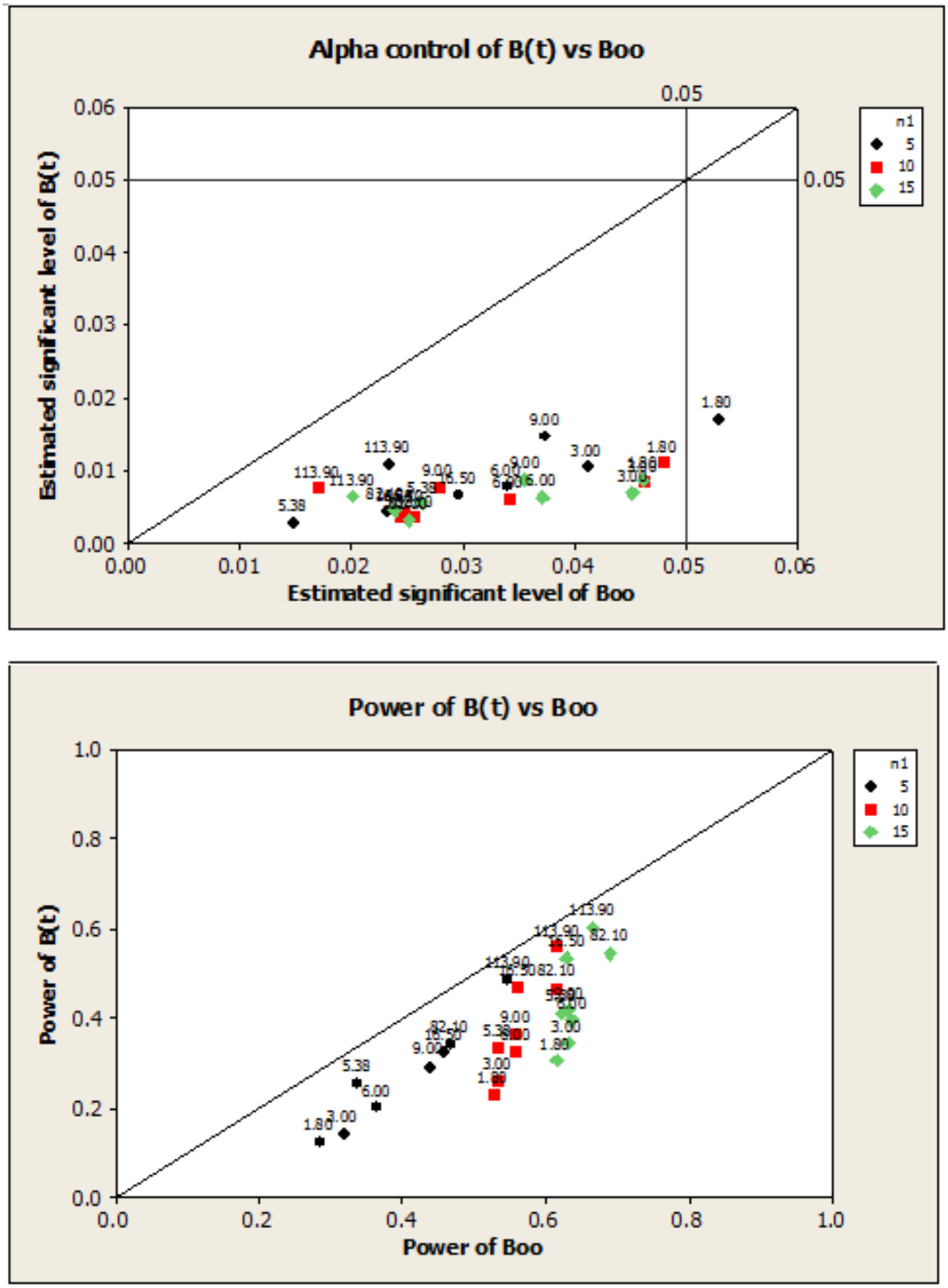


Figure 21: Overlapping Bootstrap T Confidence Intervals and Bootstrap Test

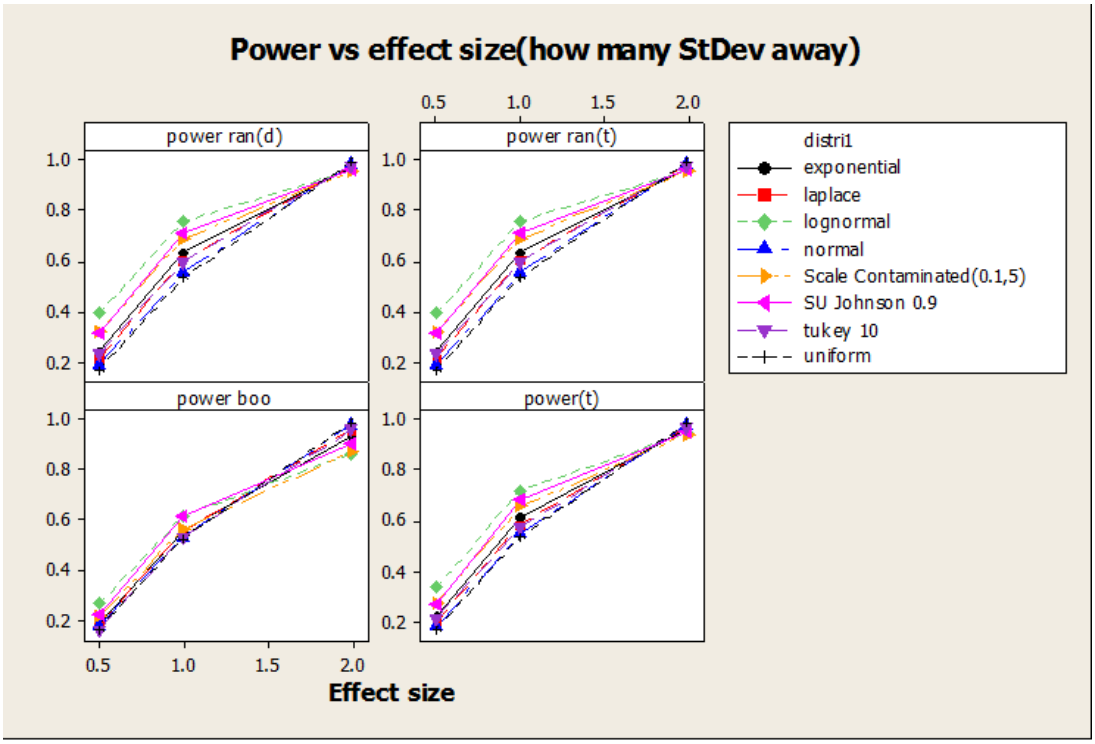


Figure 22: Power vs Effect Sizes

of the bootstrap percentile confidence interval for μ_2 is always greater than the lower bound of the t confidence interval.

The fact that power increases as the sample size increases is a widely known fact. It is also interesting to compare the power in relation to the effect size or true difference between the two population means when the null hypothesis is not true. Figure 22 shows the change in power in terms of different effect sizes for four different tests and equal sample sizes ($n_1 = n_2 = 10$) from both populations. The effect size is in terms of the standard deviation of each distribution. The tests being compared are the classical t-test, the two versions of the randomization test (using the difference of means and Welch't statistic), and the bootstrap test. The comparison is being

done for a wide range of distributions with regard to skewness and kurtosis. There is no difference in terms of power between the two versions of the randomization test because equal sample sizes were used in the simulations. The bootstrap test has lower power than other tests for some distributions. However, something interesting about the bootstrap test is that it seems to be pretty robust with regard to the shape of the distribution. For a fixed effect size of half or one standard deviations, there is less difference in power among the different distributions for the bootstrap test than for the other tests. However, when the difference between the two population means is equal to two standard deviations, the other three tests achieve a power of 1 for all the distributions, but not the bootstrap test. The average power for the bootstrap tests tends to be lower than for the other tests for all effect sizes.

- In this study, different methods to test the null hypothesis of equality of means for two populations have been compared, in terms of α control and power, by simulation using samples generated with eight distributions with different degrees of skewness and kurtosis.

- For small samples, the randomization test works better in terms of alpha control and power than other methods for a wide spectrum of distributions. Therefore, it should be considered appropriate to teach it in introductory statistics courses.

- One important reason for preferring the randomization test when some of the assumptions of the t-test are not held is that it has the best alpha control. For the exact randomization test the probability of type I error always equals the significance level. The approximate randomization test can still give the estimated significance level close to the nominal value provided that a large enough number (at least 1,000) of re-groupings is done.

- There is more than one version of the randomization test with respect to the test statistic to be used, i.e difference between sample means, t-test statistic, sum of one sample, difference of medians, etcetera. The main question in an introductory statistics course is whether to use the difference between the two means or to use the t-statistic. The Welch's t-test statistic and the difference of means will produce the same results only if the two sample sizes are equal. It is not necessary to calculate the Welch's t-statistic while doing randomization test since the difference between means is more simple to calculate and produces

better results when the sample sizes are not equal.

- The t-test is not considered an appropriate test when the data are very skewed because the assumption of normality is not being fulfilled. However, when small samples are simulated from strongly skewed distributions or distributions with high kurtosis, the power of the test may not be lower than when samples are simulated with the normal distribution or the uniform distribution when the effect size is fixed in terms of the standard deviation of the distribution. The reason is that for highly skewed distributions such as the lognormal, the standard deviation is relatively large and it is known that the larger the "effect size", the higher the power.

- The adjusted significance level for overlapping t confidence intervals proposed by Payton, Greenstone and Schenker [20] is preferred to being used in large samples. The 84% overlapping t confidence interval method for small samples is a little bit conservative.

- Schenker's idea [23] about the relationship between overlapping confidence intervals can only be applied to the t-test and t-confidence intervals but not to overlapping bootstrap confidence intervals, randomization tests or the bootstrap test.

- There is a relationship between overlapping t confidence interval methods and overlapping bootstrap percentile confidence intervals. If the two t confidence intervals are not overlapping, then the two bootstrap percentile confidence intervals also will not overlap.

- The bootstrap test is not as popular in data analysis as the randomization test. In this study, we verified by simulation that the bootstrap test is not efficient enough in dealing with small samples.

- Among all the overlapping confidence interval methods, bootstrap percentile confidence intervals work relatively better than the bootstrap-t and t-confidence intervals. When samples are so small, the normal theory based methods do not have a good performance.

BIBLIOGRAPHY

[1] K. P. Balanda and H. L. MacGillivray, Kurtosis and Spread, *The Canadian Journal of Statistics*, **18** (1990) 17-30.

[2] G. Casella and R. L. Berger, Statistical Inference, Second Edition, Duxbury, Thousand Oaks, CA (2002).

[3] M. Donegani, An Adaptive and Powerful Randomization Test, *Biometrika* **78** (1991) 930-933.

[4] S. Edgington, Randomization Test, Third Edition, Marcel Dekker, New York (1995).

[5] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, Boca Raton, FL (1997).

[6] M. D. Ernst, Permutation Methods: A Basis for Exact Inference, *Statistical Science* **19** (2004) 676-685.

[7] R. A. Fisher, Statistical Methods for Research Workers, Eleventh Edition, Hafner Publishing Company, New York (1951).

[8] R. A. Fisher, The Design of Experiments, Sixth Edition, Hafner Publishing Company, New York (1951).

[9] R. A. Fisher and P. Hall, On Bootstrap Hypothesis Testing, *Australian Journal of Statistics* **32** (1990) 177-190.

[10] Gauss Command Reference Manual, Aptech Systems, Inc. Vol II (1997).

[11] R. A. Groeneveld and G. Meeden, Measuring Skewness and Kurtosis, *The Statistician* **33** (1984) 391-399.

[12] P. Hall and S.R. Wilson, Two Guidelines for Bootstrap Hypothesis Testing, *Biometrics* **47** (1991) 757-762.

[13] R. V. Hogg and E. A. Tanis, Probability and Statistical Inference, Seventh Edition., Pearson Prentice Hall, Upper Saddle River, NJ (2006).

[14] N. L. Johnson, S. Kotz and N. Balakrishnan, Continuous Univariate Distributions Volume I and II, 2nd edition. John Wiley and Sons, NY (1994).

[15] S. Kotz and E. Seier, Visualizing Peak and Tails to Introduce Kurtosis, *The American Statistician* **62** (2008) 346-352.

[16] E. L. Lehmann, Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, San Francisco, CA (1975).

[17] N. T. Longford, Inference with the Lognormal Distribution, *Journal of Statistical Planning and Inference* **139** (2009) 2329-2340.

[18] B. F. J. Manly, Randomization Bootstrap and Monte Carlo Methods in Biology, Second Edition. Chapman & Hall, Boca Raton, FL (1997).

[19] D. S. Moore, Essential Statistics, W. H. Freeman and Company, New York (2010).

[20] M. E. Payton, M. H. Greenstone, and N. Schenker, Overlapping Confidence Intervals or Standard Error Intervals: What do they mean in terms of statistical significance?, *Journal of Insect Science* **3** (2003) 34-40.

[21] M. E. Payton, A. E. Miller, and W. R. Raun, Testing Statistical Hypotheses Using Standard Error Bars and Confidence Intervals, *Communications in Soil Science and Plant Analysis* **31** (2000) 547-552.

[22] J. G. Pitmane, Significance Tests Which May Be Applied to Samples from Any Populations, *Journal of the Royal Statistical Society* **4** (1937) 119-130.

[23] N. Schenker and J. F. Gentleman, On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals. *The American Statistician* **55** (2001) 182-186.

[24] E. Seier, Kurtosis-An Overview. *International Encyclopedia of Statistical Science* Miodrag Lovric (ed), Springer, Berlin, (2011) 328-330.

[25] E. Seier and K. Joplin, Introduction to Statistics in a Biological Context, Create Space Pub. *In press.*

[26] Student, The Probable Error of a Mean, *Biometrika* **6** (1908) 1-25.

APPENDIX

0.1 Gauss code for calculating empirical significance levels

This is the program to calculate the empirical significance level. The program to calculate power (not included here) is similar, only that the samples are generated from populations with different means. The difference between the population means is indicated by the selected effect size.

```
         /*INPUT TO BE CHANGED*/
n1 = 15;                /*  first sample size */
n2= 10 ;                /*  second sample size */
tcrit1=2.145;           /* t critical value for n1-1 */
tcrit2=2.262  ;         /* t critical value for n2-1 */
tcrit3 = 1.484;         /*for overlapping CIs*/
tcrit4 = 1.532;         /*for overlapping CIs*/
rep= 10000;             /* number of simulations */
sim = 1000;             /*number of regroupings */
mboo=1000;              /* number of bootstrap subsamples*/
              /*REJECT   Ho */
rejt=0;
rejci=0;
rejcib = 0;
rejcibt = 0;
rejcran = 0;
rejcrant=0;
rejcboo = 0;
rejcia=0;
rejciba = 0 ;
rejcibta = 0;
               /*INDICATOR FOR REJECTION*/
rejct =0;
rejcci = 0;
rejccran = 0;
rejccrant =0;
rejccib = 0;
rejccibt = 0;
rejccboo = 0;
```

```
rejccia = 0;
rejcciba = 0;
rejccibta = 0;
             /* CHECK AGREEMENT*/
  /*t vs overelapping t CIs*/
sagreetci=0;
gret = 0;
smalt = 0;
/* randomization d vs boostrap p*/
sagreeranb = 0;
grecibr=0;
smacibr = 0;
/*randomization t vs boostrap p*/
sagreeranbtt = 0;
grecibtt = 0;
smacibtt =0;
/* randomization t vs boostrap t*/
sagreeranbbt = 0;
grecibt = 0;
smacibt = 0;
/*randomization t vs boostrap p*/
sagreeranbt = 0;
grecibrt = 0;
smacibrt =0;
/* randomization d vs randomization t*/
sagreerans = 0;
greaterr= 0;
smallerr = 0;
/* bootstrap p vs bootstrap t */
sagreeboos = 0;
greaterb = 0;
smallerb =0;
/* randomization d vs boostrap test*/
sagreeranboo = 0;
greranboo  =0;
smaranboo =0;
/*randomization t vs boostrap test*/
sagreeranboot=0;
greranboot=0;
smaranboot=0;
/*bootstrap test vs t test*/
```

```
sagreeboot = 0;
greboot = 0;
smaboot = 0;
/*bootstrap test vs bootstrap percentile CI*/
sagreebooc =0;
grebooc=0;
smabooc = 0;
/*bootstrap test vs bootstrap t CI*/
sagreebooct = 0;
grebooct = 0;
smabooct = 0;
/* overlapping t CI vs boostrap percentile CI*/
sagreecib = 0;
grecib =0;
smacib = 0;
/*overlapping t CI vs boostrap t CI*/
sagreecibt = 0;
greacibt =0;
smalcibt= 0;
/* randomization(d) vs t test*/
sagreerant =0;
grerant=0;
smarant=0;
/* randomization t vs t test*/
sagreerantt=0;
grerantt=0;
smarantt=0;
/*sum of the p value for diff test*/
sumpt = 0;
sumpran=0;
sumprant=0;
sumpboo = 0;
                /*MAIN PROCEDURES*/
n = n1+n2;
n1seq=seqa(1,1,n1);
n2seq=seqa(n1+1,1,n2);
c1=0;
print " Normal vs Lognormal" ;
/*GENERATE DATA FROM CERTAIN DISTRIBUTION*/
do while c1 < rep;
  c1 = c1 + 1;
```

```
y11 = rndn(n1,1);
y1=y11+0.5;

prox2 = rndn(n2,1);
y2 = exp(prox2);
y=y1|y2;
y1m=meanc(y1);    /* calculates mean  in the first sample */
y2m=meanc(y2);    /* mean second sample */
y1s=stdc(y1);     /* standard deviation in first sample */
y2s=stdc(y2);     /* standard deviation in second sample */
ym= meanc(y);
y1star = y1-y1m+ym;  /* adjusted mean for bootstrap test, sample1*/
y2star = y2-y2m+ym;  /* adjusted mean for bootstrap test, sample2*/
torig =   (y2m-y1m)/ sqrt(y1s^2/n1+y2s^2/n2);
truedif=y1m-y2m;
tb=zeros(mboo,1);
cr=0;
cb =0;
numrejr = 0;   /* greater than true  for randomization test*/
nurejrt=0;    /* greater than t   for randomization t test*/
numrejrb = 0;  /* greater than t for bootstrap test*/

  /*RANDOMIZATION TEST FOR THE DIFFERENCE OF MEANS*/
do while cr<sim;
cr=cr+1;
hx=rndn(n,1);
hr=rankindx(hx,1);
scry=submat(y,hr,1);
rg1=submat(scry,n1seq,1);
rg2=submat(scry,n2seq,1);
meanrg1=meanc(rg1);
meanrg2=meanc(rg2);
difmeanrg = meanrg1-meanrg2;
if abs(difmeanrg) > abs(truedif);
numrejr = numrejr+1;
endif;
                     /* RANDOMIZATION T TEST */
y1sr=stdc(rg1);
y2sr=stdc(rg2);
trand =   (difmeanrg)/ sqrt(y1sr^2/n1+y2sr^2/n2);
if abs(trand)>abs(torig);
```

```
nurejrt=nurejrt+1;
endif;
endo;
pvalran = numrejr/sim;
pvalrant=nurejrt/sim;
sumpran = sumpran+pvalran;
sumprant=sumprant+pvalrant;
if pvalran<0.05;
rejcran = rejcran+ 1;
endif;
if pvalran<0.05;
rejccran = 1;
else; rejccran = 0;
endif;
if pvalrant<0.05;
rejcrant = rejcrant+ 1;
endif;
if pvalrant<0.05;
rejccrant = 1;
else; rejccrant = 0;
endif;

 /*BOOTSTRAP CI AND BOOTSTRAP TEST*/
        /*BOOTSTRAP SAMPLE 1*/
whob1 = rndu(n1,mboo);
whosb1 = n1*whob1;
whosib1 = ceil(whosb1);
py1 = submat(y1,whosib1,0);
yb1 = reshape(py1,n1,mboo);
yvar1=meanc(yb1);
sovar1 = sortc(yvar1,1);
seb1=stdc(yvar1);
pyb1 = submat(y1star,whosib1,0);
ybb1 = reshape(pyb1,n1,mboo);
yvarb1=meanc(ybb1);
sebb1=stdc(ybb1);
k1 = (mboo+1)*0.025;
k2 = (mboo+1)*0.975;
k11 = (mboo+1)*0.08;
k22 = (mboo+1)*0.92;
Lper1 = sovar1[k1,.];
```

```
Uper1 = sovar1[k2,.];
Lper11 = sovar1[k11,.];
Uper11 = sovar1[k22,.];
Lbt1=y1m-tcrit1*seb1;
Ubt1=y1m+tcrit1*seb1;
Lbt11=y1m-tcrit3*seb1;
Ubt11=y1m+tcrit3*seb1;
    /*BOOTSTRAP SAMPLE 2*/
whob2 = rndu(n2,mboo);
whosb2 = n2*whob2;
whosib2 = ceil(whosb2);
py2 = submat(y2,whosib2,0);
yb2 = reshape(py2,n2,mboo);
yvar2=meanc(yb2);
sovar2 = sortc(yvar2,1);
seb2=stdc(yvar2);
pyb2= submat(y2star,whosib2,0);
ybb2 = reshape(pyb2,n2,mboo);
yvarb2=meanc(ybb2);
sebb2=stdc(ybb2);
Lper2 = sovar2[k1,.];
Uper2 = sovar2[k2,.];
Lper22 = sovar2[k11,.];
Uper22 = sovar2[k22,.];
Lbt2=y2m-tcrit2*seb2;
Ubt2=y2m+tcrit2*seb2;
Lbt22=y2m-tcrit4*seb2;
Ubt22=y2m+tcrit4*seb2;
     /*CHECK WHETER OVERLAPPING FOR PERCENTILE METHOD*/
if (Lper2>Uper1);
 rejcib=rejcib+1;
endif;
if (Lper2>Uper1);
 rejccib = 1;
else; rejccib = 0;
endif;
if(Lper1>Uper2);
 rejcib = rejcib+1;
endif;
if (Lper1>Uper2);
 rejccib = 1;
```

```
else; rejccib = 0;
endif;

/* for 84% individual CI*/
if (Lper22>Uper11);
 rejciba=rejciba+1;
endif;
if (Lper22>Uper11);
 rejcciba = 1;
else; rejcciba = 0;
endif;
if(Lper11>Uper22);
 rejciba = rejciba+1;
endif;
if (Lper11>Uper22);
 rejcciba = 1;
else; rejcciba = 0;
endif;
    /* CHECK WHETHER OVERLAPPING FOR  T METHOD*/
if(Lbt1>Ubt2);
 rejcibt = rejcibt+1;
endif;
if (Lbt1>Ubt2);
 rejccibt = 1;
else; rejccibt = 0;
endif;
if (Lbt2>Ubt1);
 rejcibt=rejcibt+1;
endif;
if (Lbt2>Ubt1);
 rejccibt = 1;
else; rejccibt = 0;
endif;
if(Lbt11>Ubt22);
 rejcibta = rejcibta+1;
endif;
if (Lbt11>Ubt22);
 rejccibta = 1;
else; rejccibta = 0;
endif;
if (Lbt22>Ubt11);
```

```
 rejcibta=rejcibta+1;
endif;
if (Lbt22>Ubt11);
 rejccibta = 1;
else; rejccibta = 0;
endif;
    /*BOOTSTRAP TEST USING T*/
do while cb<mboo;
cb=cb+1;
tb[cb] = (yvarb1[cb]-yvarb2[cb])/sqrt((sebb1[cb])^2/n1+(sebb2[cb])^2/n2);
if abs(tb[cb] )> abs(torig);
numrejrb = numrejrb+1;
endif;
endo;
pvalboo = numrejrb/mboo;
sumpboo = sumpboo+pvalboo;
if pvalboo<0.05;
rejcboo = rejcboo+ 1;
endif;
if pvalboo<0.05;
rejccboo =1;
else; rejccboo=0;
endif;
       /*   TWO SIDED T TEST FOR UNEQUAL VARIANCES */
t =   (y2m-y1m)/ sqrt(y1s^2/n1+y2s^2/n2);
dft=  (y1s^2/n1+y2s^2/n2)^2/(y1s^4/(n1^2*(n1-1))+y2s^4/(n2^2*(n2-1)));
at=abs(t);
pvalt=2*cdftc(at,dft);
sumpt=sumpt+pvalt;
if pvalt<0.05;
rejt=rejt+1 ;
endif ;
if pvalt<0.05;
rejct=1;
else;
rejct=0;
endif;
      /*  OVERLAPPING CONFIDENCE INTERVALS */
le1=y1m-tcrit1 * y1s/sqrt(n1);
ue1= y1m+tcrit1 * y1s/sqrt(n1);
le2=y2m-tcrit2* y2s/sqrt(n2);
```

```
ue2=y2m+tcrit2 * y2s/sqrt(n2);
 if (le2>ue1);
 rejci=rejci+1;
endif;
if(le2>ue1);
rejcci=1;
else; rejcci=0;
endif;
if (le1>ue2);
 rejci=rejci+1;
endif;
if (le1>ue2);
rejcci=1;
else; rejcci=0;
endif;
le11=y1m-tcrit3 * y1s/sqrt(n1);
ue11= y1m+tcrit3 * y1s/sqrt(n1);
le22=y2m-tcrit4* y2s/sqrt(n2);
ue22=y2m+tcrit4 * y2s/sqrt(n2);
 if (le22>ue11);
 rejcia=rejcia+1;
endif;
if(le22>ue11);
rejccia=1;
else; rejccia=0;
endif;
if (le11>ue22);
 rejcia=rejcia+1;
endif;
if (le11>ue22);
rejccia=1;
else; rejccia=0;
endif;
      /*AGREEMENT FOR REJECTION */
                /*agreement for two randomization tests*/
if rejccran == rejccrant;
sagreerans = sagreerans +1;
endif;
if rejccran > rejccrant;
greaterr = greaterr+1;
endif;
```

```
if rejccran < rejccrant;
smallerr = smallerr+1;
endif;
           /*agreement for two bootstrap CIs*/
if rejccib ==rejccibt;
sagreeboos = sagreeboos +1;
endif;
if rejccib > rejccibt;
greaterb = greaterb+1;
endif;
if rejccib< rejccibt;
smallerb = smallerb+1;
endif;
    /*for percentile bootstrap CI v.s. randomization d*/
 if (rejccib==rejccran);
sagreeranb=sagreeranb+1;
endif;
if rejccib>rejccran;
grecibr = grecibr +1;
endif;
if rejccib<rejccran;
smacibr = smacibr+1;
endif;
if(rejccib==rejccrant);
sagreeranbt= sagreeranbt+1;
endif;
if rejccib>rejccrant;
grecibrt = grecibrt+1;
endif;
if rejccib<rejccrant;
smacibrt = smacibrt+1;
endif;
    /* for t bootstrap CI v.s. randomization*/
 if (rejccibt==rejccran);
sagreeranbbt=sagreeranbbt+1;
endif;
if rejccibt>rejccran;
grecibt =grecibt+1;
endif;
if rejccibt<rejccran;
smacibt=smacibt+1;
```

```
endif;
if(rejccibt==rejccrant);
sagreeranbtt= sagreeranbtt+1;
endif;
if rejccibt>rejccrant;
grecibtt = grecibtt+1;
endif;
if rejccibt<rejccrant;
smacibtt = smacibtt+1;
endif;
if (rejcci==rejct);
sagreetci=sagreetci+1;
endif;
if rejcci > rejct;
gret = gret+1;
endif;
if rejcci<rejct;
smalt = smalt +1;
endif;
  /*agreement  randomization t test v.s t test */
if rejccrant == rejct;
sagreerantt = sagreerantt+1;
endif;
if rejccrant>rejct;
grerantt = grerantt+1;
endif;
if rejccrant<rejct;
smarantt = smarantt+1;
endif;
    /*agreement overlap CI and bootstrap percentile CI*/
if rejccib == rejcci;
sagreecib = sagreecib+1;
endif;
if rejccib>rejcci;
grecib = grecib+1;
endif;
if rejccib<rejcci;
smacib = smacib+1;
endif;
   /*agreement overlap CI and bootstrap t CI*/
if rejccibt == rejcci;
```

```
sagreecibt = sagreecibt+1;
endif;
if rejccibt>rejcci;
greacibt = greacibt+1;
endif;
if rejccibt<rejcci;
smalcibt = smalcibt+1;
endif;
     /*agreement bootstrap test vs t test*/
 if rejccboo == rejct;
sagreeboot = sagreeboot+1;
endif;
if rejccboo>rejct;
greboot = greboot+1;
endif;
if rejccboo<rejct;
smaboot = smaboot+1;
endif;
   /*agreement boostrap test vs randomization test*/
 if (rejccboo==rejccran);
sagreeranboo=sagreeranboo+1;
endif;
if rejccboo>rejccran;
greranboo =greranboo+1;
endif;
if rejccboo<rejccran;
smaranboo=smaranboo+1;
endif;
    /*agreement boostrap test and bootstrap percentile CI*/
if rejccboo == rejccib;
sagreebooc= sagreebooc+1;
endif;
if rejccboo>rejccib;
grebooc = grebooc+1;
endif;
if rejccboo<rejccib;
smabooc = smabooc+1;
endif;
   /*agreement bootstrap test and bootstrap t CI*/
if rejccboo== rejccibt;
sagreebooct = sagreebooct+1;
```

```
endif;
if rejccboo>rejccibt;
grebooct = grebooct+1;
endif;
if rejccboo<rejccibt;
smabooct = smabooct+1;
endif;
       /*agreement randomization d vs t test*/
if rejccran== rejct;
sagreerant = sagreerant+1;
endif;
if rejccran>rejct;
grerant = grerant+1;
endif;
if rejccran<rejct;
smarant = smarant+1;
endif;
    /* agreement bootstrap test vs randomization t test*/
 if (rejccboo==rejccrant);
sagreeranboot=sagreeranboot+1;
endif;
if rejccboo>rejccrant;
greranboot =greranboot+1;
endif;
if rejccboo<rejccrant;
smaranboot=smaranboot+1;
endif;
endo;
/* THE LOOP ENDS */
      /* CALCULATE SUMMARIES */
avepvalt=sumpt/rep;
avepvalr= sumpran/rep;
avepvalrt = sumprant/rep;
avepvalboo= sumpboo/rep;
alphat=rejt/rep;
alphaover=rejci/rep;
alphaovera=rejcia/rep;
alpharan = rejcran/rep;
alpharant= rejcrant/rep;
alphaboo = rejcboo/rep;
alphacib = rejcib/rep;
```

```
alphacibt = rejcibt/rep;
alphaciba = rejciba/rep;
alphacibta= rejcibta/rep;
agreetci=sagreetci/rep;
agreeranb = sagreeranb/rep;   /* rand.  v.s bootstrap perc. CI*/
agreeranbt = sagreeranbt/rep;     /* rand.t  v.s bootstrap perc. CI*/
agreeranbbt = sagreeranbbt/rep;  /* rand. v.s bootstrap t CI*/
agreeranbtt  = sagreeranbtt/rep;    /* rand.t  v.s bootstrap t CI*/
agreerans = sagreerans/rep;   /*rand. vs rand. t*/
agreeboos = sagreeboos/rep;  /*bootstrap perc. CI v.s bootstrap t CI*/
agreerantt = sagreerantt/rep;  /*rand.t vs t test*/
agreecib = sagreecib/rep;    /* bootstrap perc. CI vs overlapping CI*/
agreecibt = sagreecibt/rep; /* bootstrap t CI vs overlapping CI*/
agreeboot = sagreeboot/rep;  /* bootstrap test vs t test*/
agreeranboo = sagreeranboo/rep; /*bootstrap test vs rand. test*/
agreebooc = sagreebooc/rep;    /*bootstrap test vs boostrap perc. CI*/
agreebooct = sagreebooct/rep;   /*bootstrap test vs boostrap t CI*/
agreeranboot = sagreeranboot/rep; /*bootstrap test vs rand. t test*/
agreerant = sagreerant/rep; /*rand. d vs t test*/
         /* PRINT RESULTS */
print "TWO SIDED TEST " ;
print "Sample sizes:"    n1~n2;
print "  ALPHA-t    ALPHA-T CI  ALPHA RAN(DIF)  ALPHA RAN(T)  ";
print    alphat~alphaover~alpharan~alpharant;
print BOO CI(PER)  BOO T CI   BOO tci(84%) b(p)(84%) b(t)(84%)";
print alphacib~alphacibt~alphaboo~alphaovera~alphaciba~alphacibta;
print "TvsCI RAN(D)vsB(P)RAN(D)vsB(T)RAN(T)vsB(P) RAN(T)vsB(T)";
print "RANs    BOOs    RAN(T)vsT   RAN(D) vs T  B(P)vsCI";
print "B(T)vsCI BOOvsT BOOvsRAN(D) BOOvsRAN(T) BOOvsB(P)  BOOvsB(T)";
print   agreetci~agreeranb~agreeranbt~agreeranbbt~agreeranbtt;
print agreerans~agreeboos~agreerantt~agreerant~agreecib;
print agreecibt~agreeboot~agreeranboo~agreeranboot~agreebooc~agreebooct;
print "  AVE P (T)    AVE P(RAN DIF)    AVE P(RAN T)    AVE BOO";
print avepvalt~avepvalr~avepvalrt~avepvalboo;
print "     Agreement       Rej1not2          Rej2not1";
print "CI test Vs T";
print sagreetci~gret~smalt;
print "Bootstrap percentile CI vs Randomization test";
print  sagreeranb~grecibr~smacibr;
print "Bootstrap percentile CI vs Randomization t test";
print  sagreeranbt~grecibrt~smacibrt;
```

```
print "Bootstrap t CI vs Randomization test";
print  sagreeranbbt~grecibt~smacibt;
print "Boostrap t CI vs Randomization t test";
print  sagreeranbtt~grecibtt~smacibtt;
print "Randomization vs randomization t test";
print sagreerans~greaterr~smallerr;
print "Bootstrap percentile CI vs Bootstrap t CI";
print  sagreeboos~greaterb~smallerb;
print "Randomization t test vs t test";
print  sagreerantt~grerantt~smarantt;
print "Randomization d test vs t test";
print  sagreerant~grerant~smarant;
print "Boostrap percentile CI vs CI";
print  sagreecib~grecib~smacib;
print "Boostrap t CI vs CI";
print  sagreecibt~greacibt~smalcibt;
print "Bootstrap test vs t test";
print  sagreeboot~greboot~smaboot;
print "Boostrap test vs randomization d test";
print sagreeranboo~greranboo~smaranboo;
print "Boostrap test vs randomization t test";
print sagreeranboot~greranboot~smaranboot;
print "Boostrap test vs boostrap percentile CI";
print sagreebooc~grebooc~smabooc;
print "Boostrap test vs boostrap t CI";
print sagreebooct~grebooct~smabooct;
end;
```

VITA

HAIYIN LI

Education:	B.S. Information and Computing Science, Shandong Normal University, Jinan, China; 06/2008
	B.S.(Summa Cum Laude) Mathematics (Statistics track), East Tennessee State University, Johnson City, Tennessee; 05/2009
	M.S. Mathematics, East Tennessee State University; Johnson City, Tennessee; 05/2011
Professional Experience:	Data Analyst, Anatomical Collaborative Research, College of Public Health, East Tennessee State University, Johnson City, Tennessee; 04/2008-12/2008
	Math Lab Monitor, Department of Mathematics and Statistics, East Tennessee State University, Johnson City, Tennessee; 09/2008-05/2009
	Graduate Assistant-Teaching, Department of Mathematics and Statistics, East Tennessee State University, Johnson City, Tennessee; 08/2009-05/2011