The Pennsylvania State University

The Graduate School

Eberly College of Science

AN ALTERNATIVE APPROACH TO BOOTSTRAP HYPOTHESIS TESTING

A Thesis in

Statistics

by

Amanda Tomlinson

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2011

The thesis of Amanda Tomlinson was reviewed and approved* by the following:

Trent Gaugler
Assistant Professor of Statistics
Thesis Advisor

Michael Akritas
Professor of Statistics

Bruce Lindsay
Willaman Professor of Statistics
Head of the Department of Statistics

*Signatures are on file in the Graduate School

ABSTRACT

We propose a bootstrap procedure for hypothesis testing that avoids the conflicting nature of existing bootstrap hypothesis testing methods which employ a single observed dataset for representing the distribution of the test statistic under the null hypothesis as well as for providing evidence for rejecting the null hypothesis. Our proposed procedure, the quantile bootstrap, avoids this conflict by testing under the alternative hypothesis, which lends itself to improved power. A second level of iterations within the standard resampling scheme common to bootstrap procedures gives the procedure accurate type I error. By way of several simulations under several distributions, we show that the procedure is appropriate for many basic hypothesis tests, such as the 1-sample test of the mean, the 2-sample test of equality of means, and the test of equality of *r* means.

TABLE OF CONTENTS

LIST OF FIGURES v

LIST OF TABLES viii

ACKNOWLEDGEMENTS ix

Chapter 1 Introduction 1

Chapter 2 One-Sample Test 2

- The Quantile Bootstrap 2
- Simulation Methodology 4
- Simulation Results for Standard Normal 4
- Simulation Results for Uniform 6
- Simulation Results for Beta 8
- Simulation Results for Discrete 10

Chapter 3 Two-Sample One-Sided Test 13

- Extension of Method to Two-Sample One-Sided Test 13
- Difficulties arising from skewness 17

Chapter 4 Tests of Two or More Means 23

- Extension of Method for Tests Comparing Equality of Two or More Means 23
- Simulation Results for r = 2 25
- Three or More Groups 36
- Simulation Results for r = 3 38
- Simulation Results for r = 5 45
- Simulation Results for r = 10 48

Chapter 5 Conclusions 51

Bibliography 52

Appendix A R Code 53

LIST OF FIGURES

Figure 2-1. Level for 1-Sample Test, X ~ Normal(0,1). Dashed lines added for scale. 5

Figure 2-2. Power for 1-Sample Test, X ~ Normal(0,1) 6

Figure 2-3. Level for 1-Sample Test, X ~ Uniform(-1, 1) 7

Figure 2-4. Power for 1-Sample Test, X ~ Uniform(-1,1) 8

Figure 2-5. Power for 1-Sample Test, X ~ Beta(5,1) 10

Figure 2-6. Probability Mass Function for Discrete Distribution 11

Figure 2-7. Power for 1-Sample Test, X ~ Discrete 11

Figure 3-1. Power for 2-Sample, 1-Sided Test for Normal/Normal, homoscedastic, balanced cases. Panel labels represent σ_1^2, σ_2^2 / n_1, n_2 16

Figure 3-2. Power for 2-Sample, 1-Sided Test for Normal/Normal, heteroscedastic, balanced cases. Panel labels represent σ_1^2, σ_2^2 / n_1, n_2.....16

Figure 3-3. Power for 2-Sample, 1-Sided Test for Normal/Normal, heteroscedastic, unbalanced cases. Panel labels represent σ_1^2, σ_2^2 / n_1, n_2.....17

Figure 3-4. Probability Mass Function of Distributions "Left 4.55" and "Right 4.55".....20

Figure 3-5. Probability Mass Function of Distributions "Left 4.55" and "Right 11.03".....20

Figure 3-6. Power for 2-Sample, 1-Sided Test, X_1 ~ "Left 4.55", X_2 ~ "Right 4.55". Panel labels represent n_1, n_2 22

Figure 3-7. Power for 2-Sample, 1-Sided Test, X_1 ~ "Left 4.55", X_2 ~ "Right 11.03". Panel labels represent n_1, n_2 22

Figure 4-1. Power for r = 2, X_1 ~ Normal, X_2 ~ Normal; σ_1^2, σ_2^2 = 1, 1; n_1, n_2 = 10, 10.....28

Figure 4-2. Power for r = 2, X_1 ~ Normal, X_2 ~ Normal; σ_1^2, σ_2^2 = 1, 1; n_1, n_2 = 30, 30.....28

Figure 4-3. Power for r = 2, X_1 ~ Normal, X_2 ~ Normal; σ_1^2, σ_2^2 = 9, 1; n_1, n_2 = 10, 10.....29

Figure 4-4. Power for r = 2, X_1 ~ Normal, X_2 ~ Normal; σ_1^2, σ_2^2 = 9, 1; n_1, n_2 = 30, 30.....29

Figure 4-5. Power for r = 2, X_1 ~ Normal, X_2 ~ Normal; σ_1^2, σ_2^2 = 9, 1; n_1, n_2 = 10, 30.....30

Figure 4-6. Power for r = 2, X_1 ~ Normal, X_2 ~ Normal; σ_1^2, σ_2^2 = 9, 1; n_1, n_2 = 30, 10.....30

Figure 4-7. Power for r = 2, X_1 ~ Chi-square, X_2 ~ Normal; σ_1^2, σ_2^2 = 4, 4; n_1, n_2 = 10, 10 31

Figure 4-8. Power for r = 2, X_1 ~ Chi-square, X_2 ~ Normal; σ_1^2, σ_2^2 = 4, 4; n_1, n_2 = 30, 30 32

Figure 4-9. Power for r = 2, X_1 ~ Chi-square, X_2 ~ Normal; σ_1^2, σ_2^2 = 2, 1; n_1, n_2 = 10, 10 32

Figure 4-10. Power for r = 2, X_1 ~ Chi-square, X_2 ~ Normal; σ_1^2, σ_2^2 = 2, 1; n_1, n_2 = 30, 30 ..33

Figure 4-11. Power for r = 2, X_1 ~ Chi-square, X_2 ~ Normal; σ_1^2, σ_2^2 = 2, 1; n_1, n_2 = 30, 10 ..33

Figure 4-12. Power for r = 2, X_1 ~ "Left 4.55", X_2 ~ "Right 4.55"; n_1, n_2 = 10, 10 34

Figure 4-13. Power for r = 2, X_1 ~ "Left 4.55", X_2 ~ "Right 4.55"; n_1, n_2 = 30, 30 35

Figure 4-14. Power for r = 2, X_1 ~ "Left 4.55", X_2 ~ "Right 11.03"; n_1, n_2 = 10, 30 35

Figure 4-15. Power for r = 2, X_1 ~ "Left 4.55", X_2 ~ "Right 11.03"; n_1, n_2 = 30, 10 36

Figure 4-16. Power for r = 3, X1 ~ Chi-square, X2 and X3 ~ Normal; Homoscedastic, Balanced,.....39

Figure 4-17. Power for r = 3, X_1 ~ Chi-square, X_2 and X_3 ~ Normal; Homoscedastic, Balanced,.....39

Figure 4-18. Power for r = 3, X_1 ~ Chi-square, X_2 and X_3 ~ Normal; Heteroscedastic, Unbalanced, Large/Large.....40

Figure 4-19. Power for r = 3, all groups Normal; Homoscedastic, Balanced, n = 10.....42

Figure 4-20. Power for r = 3, all groups Normal; Homoscedastic, Balanced, n = 30.....42

Figure 4-21. Power for r = 3, all groups Normal; Heteroscedastic, Balanced, n = 10.....43

Figure 4-22. Power for r = 3, all groups Normal; Heteroscedastic, Balanced, n = 30.....43

Figure 4-23. Power for r = 3, all groups Normal; Heteroscedastic, Unbalanced, Large/Large.....44

Figure 4-24. Power for r = 3, all groups Normal; Heteroscedastic, Unbalanced, Small/Large.....44

Figure 4-25. Power for r = 5, all groups Normal; Homoscedastic, Balanced.....46

Figure 4-26. Power for r = 5, all groups Normal; Heteroscedastic, Balanced.....46

Figure 4-27. Power for r = 5, all groups Normal; Heteroscedastic, Unbalanced Large/Large.....47

Figure 4-28. Power for r = 5, all groups Normal; Heteroscedastic, Unbalanced, Small/Large.....47

Figure 4-29. Power for r = 10, all groups Normal; Homoscedastic, Balanced.....49

Figure 4-30. Power for r = 10, all groups Normal; Heteroscedastic, Balanced 49

Figure 4-31. Power for r = 10, all groups Normal; Heteroscedastic, Unbalanced, Large/Large 50

Figure 4-32. Power for r = 10, all groups Normal; Heteroscedastic, Unbalanced, Small/Large 50

LIST OF TABLES

Table 2-1. Level for 1-Sample Test, X ~ Normal(0,1) 6

Table 2-2. Level for 1-Sample Test, X ~ Uniform(-1, 1) 8

Table 2-3. Level for 1-Sample Test, X ~ Beta(5,1) 9

Table 2-4. Level for 1-Sample Test, X ~ Discrete 11

Table 3-1. Level for 2-Sample 1-Sided Test, X_1, X_2 ~ Normal 15

Table 3-2. Level for 2-Sample 1-Sided Test, X_1 ~ Chi-square, X_2 ~ Normal 18

Table 3-3. Level for 2-Sample 1-Sided Test, X_1 ~ Chi-square, X_2 ~ Exponential 19

Table 3-4. Level for 2-Sample 1-Sided Test, X_1, X_2 ~ Discrete 21

Table 4-1. Level for r = 2, X_1 ~ Normal, X_2 ~ Normal 27

Table 4-2. Level for r = 2, X_1 ~ Chi-square, X_2 ~ Normal 31

Table 4-3. Level for r = 2, X_1 ~ Chi-square, X_2 ~ Exponential 34

Table 4-4. Level for r = 2, X_1 ~ Discrete, X_2 ~ Discrete 34

Table 4-5. Details of r = 3 simulation cases, X_1 ~ Chi-square, X_2 and X_3 ~ Normal 38

Table 4-6. Level for r = 3, X_1 ~ Chi-square, X_2 and X_3 ~ Normal 38

Table 4-7. Details of r = 3 simulation cases, all groups drawn from Normal distribution 41

Table 4-8. Level for r = 3, all groups drawn from Normal distribution 41

Table 4-9. Details of r = 5 simulation cases, all groups drawn from Normal distribution 45

Table 4-10. Level Error for r = 5, all groups drawn from Normal distribution 45

Table 4-11. Details of r = 10 simulation cases, all groups drawn from Normal distribution 48

Table 4-12. Level Error for r = 10 48

ACKNOWLEDGEMENTS

First and foremost, I'd like to thank my advisor, Trent Gaugler, for agreeing to do this in the first place and for his patient guidance and explanations. I'm sad to see Penn State lose such a genuine educator.

I'd like to thank my parents, Craig and Susan Tomlinson, for imparting the value of ~~argument~~ inquiry, for never doubting that I'd ever actually finish this thing, and for keeping me grounded through all of my academic pursuits.

I'd like to thank my employer Minitab for supporting this endeavor, especially all of my enthusiastic coworkers in Tech Support who helped me with homework problems (despite the violation of support policy).

Finally, I'd like to thank my partner Amber Miller, for her endless supply of patience, understanding, pencils, witticisms, and love. "That's good! One less thing."

Chapter 1 Introduction

One of the most common and basic of hypothesis tests is the comparison of means. The standard parametric answer to these types of problems is the t-test or one-way ANOVA, depending on the number of means being compared. As with all parametric tests, these tests have parametric assumptions that, when violated, can lead to inaccurate type-I errors. One (non-parametric) solution to violations of the assumptions is the set of bootstrap methods. Bootstrap methods have been devised for testing the equality of group means in two or more groups, without the homoscedastic assumption, and have shown improved level over the standard F-test specifically when the homoscedastic assumption is violated (Fisher and Hall, 1990).

Fisher and Hall (1990) contend that, when testing hypotheses, it is essential that the data be transformed appropriately so that any departures from the null hypothesis represented in the observed data do not detract from the accuracy of the test. However, we suggest that such transformations, while preserving the accuracy of level error, lead to significant reductions in power. Consider that, when testing a hypothesis, the goal of the researcher is usually to collect enough evidence to reject the null hypothesis. It seems counter-productive to utilize the same observed dataset to represent the null hypothesis *and* to provide compelling evidence to reject the null hypothesis. Alternatively, we propose a method that does not attempt to use the observed dataset to represent the null hypothesis: Rather than use the observed dataset to represent the null hypothesis and determine the extremity of the observed test statistic, we propose using the observed dataset to represent an alternative hypothesis and determining the extremity of the null value. We suggest that this alternative paradigm results in greater power while maintaining accurate level error.

Chapter 2 One-Sample Test

The Quantile Bootstrap

A naïve version of this approach is as follows. Assume that a random sample $X = \{x_j, 1 \leq j \leq n\}$ is drawn from a population with mean μ and variance σ^2. No assumption is made about the distribution of the population. The goal is to test the null hypothesis H_0: $\mu = \mu_0$ against the one-sided alternative hypothesis H_A: $\mu > \mu_0$. The procedure is as follows:

1. Sample, with replacement, from X, resulting in resample $X^* = \{x_1^*, \dots, x_n^*\}$.

2. Calculate $\bar{x}^* = \frac{1}{n} \sum_{j=1}^{n} x_j^*$

3. Repeat steps 1-2 B times. Denote $\bar{x}_b^*$ as the mean of the b^{th} resample of X.

Then, an approximate α-level test is given by first estimating the quantile $\bar{x}_\alpha$ by solving $P(\bar{x}^* < \hat{\bar{x}}_\alpha \mid X) = \alpha$ and rejecting H_0 if $\hat{\bar{x}}_\alpha > \mu_0$. Equivalently, we can compute the bootstrap p-value $P = \frac{1}{B} \sum_{b=1}^{B} (\bar{x}_b^* < \mu_0)$ and reject the null hypothesis if P is less than α.

Here, we use $\bigcup_{b=1}^{B} \bar{x}_b^*$ to approximate the distribution of the sample mean, possibly under an alternative hypothesis, and compare the α^{th} percentile to μ_0, instead of the standard bootstrap procedure of transforming X to represent the null hypothesis and then comparing the (1- α)th percentile to the observed test statistic.

Simulations under various distributions show great power for the naïve approach, although level is shown to be liberal. To correct for the liberal test, we propose an alternative approach utilizing two stages of bootstrapping. The first stage estimates the probability of observing a sample mean less than or equal to μ_0 under the alternative hypothesis. The second stage estimates the probability of obtaining that probability's corresponding quantile of the sample mean's sampling distribution under the null hypothesis. For a one-sample test, the quantile bootstrap procedure is as follows:

Stage 1

1. Sample, with replacement, from X, resulting in resample $X^* = \{x_1^*, \dots, x_n^*\}$.

2. Calculate $\bar{x}^* = \frac{1}{n} \sum_{j=1}^{n} x_j^*$

3. Repeat steps 1-2 B times. Denote $\bar{x}_b^*$ to be the mean of the b^{th} resample of X.

4. Calculate $p = \frac{1}{B}\sum_{b=1}^{B}(\bar{x}_b^* < \mu_0)$

Stage 2

1. Transform the data to represent the null hypothesis. Let $y_j = x_j - \bar{x} + \mu_0$, resulting in

$Y = \{y_1, \dots, y_n\}$.

2. Sample, with replacement, from Y.

3. Repeat step 2 M times. Denote the m^{th} resample of Y as $Y^{*m} = \{y_1^{*m}, \dots, y_n^{*m}\}$.

4. Sample, with replacement, fromY^{*m}, $\forall m$.

5. Repeat step 4 B times. Denote the b^{th} resample of Y^{*m} as $Y^{**mb} = \{y_1^{**mb}, \dots, y_n^{**mb}\}$.

6. Calculate $\bar{y}^{**mb} = \frac{1}{n}\sum_{j=1}^{n} y_j^{**mb}$, $\forall m, b$.

7. Let $\bar{Y}^{*m} = \bigcup_{b=1}^{B} \bar{y}_b^{**mb}$, $m = 1, \dots, M$.

8. Calculate y_p^{*m}, the $(100 * p)^{th}$ quantile of $\bar{Y}^{*m}$, $\forall m$.

9. Calculate the p-value as $\frac{1}{M}\sum_{m=1}^{M}(y_p^{*m} > \mu_0)$.

Notice that Stage 1 is equivalent to the naïve method; Stage 1 results in p, an estimate of the probability of obtaining the null value under an alternative hypothesis, as represented by the resampled observed data. However, Stage 2 goes further by asking: If the null hypothesis is true, what is the probability of obtaining as small a probability as p? In other words, Stage 2 estimates the distribution of p under the null hypothesis.

To illustrate this method, we performed a simulation study that compared the performance of the naïve method and the quantile bootstrap method to the standard parametric test, the 1-sample t-test. It should be noted that there are countless other 1-sample nonparametric tests of the sample mean that we could have included for comparison as well. There are even more if we opened the field to all tests of central tendency. However, the goal of this study was not to compare the performance of our proposed methods to the entire set of 1-sample options, but to show that it is a viable method upon which more complicated hypotheses can be tested.

Simulation Methodology

To assess the performance of the naïve method and the quantile bootstrap method compared to the standard t-test, we produced 2,000 random independent samples from a specific distribution and applied the t-test, the naïve method test, and the quantile bootstrap method test to the data in order to test the null hypothesis $H_0: \mu = 0$. To estimate level, we calculated the proportion of times each test produced a statistically significant result (p-value < α) when H_0 was true. To estimate power, we produced 2,000 random independent samples, added an increment, $\Delta > 0$, applied the three tests to the data, setting B=900 and M=200, and calculated the proportion of times each test produced a statistically significant result (p-value < α) when the alternative hypothesis $H_A: \mu = \Delta > 0$ was true. For all tests, α was set at 0.05.

Simulation Results for Standard Normal

We began with the standard normal distribution. As we expected, the naïve method produced improved power over the t-test, but it also produced very liberal level, especially at smaller sample sizes. However, at N = 15, level for the naïve method was estimated at 0.0635,

which is not as liberal as one might expect. And at N = 30, level was estimated at 0.055, which is very reasonable. As we expected, the quantile bootstrap method produced excellent level at all sample sizes, performing as well as the t-test.

As would be expected with a liberal test, the naïve method produced excellent power, although with level error of 7.9% and higher for N ≤ 10, it is an unusable test. The quantile bootstrap performed nearly as well as the t-test in terms of power, with the notable exception of N = 5, where it performed much worse than the t-test. This is unsurprising, given what is an extremely small dataset for a bootstrap procedure.

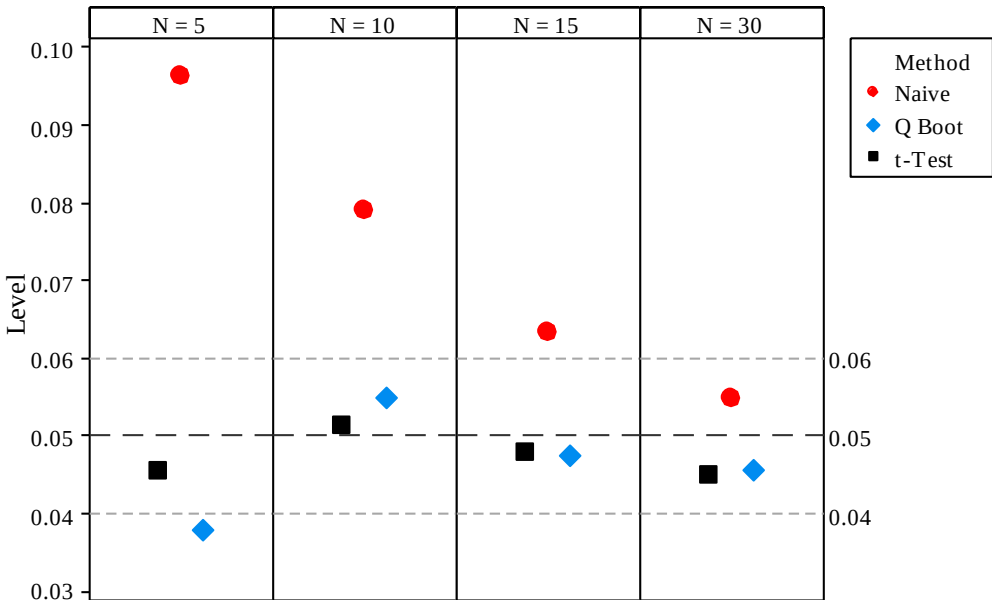

Figure 2-1. Level for 1-Sample Test, X ~ Normal(0,1). Dashed lines added for scale.

Table 2-1. Level for 1-Sample Test, X ~ Normal(0,1)

N	Naïve	Q boot	t-Test
5	9.65%	3.80%	4.55%
10	7.90%	5.50%	5.15%
15	6.35%	4.75%	4.80%
30	5.50%	4.55%	4.50%

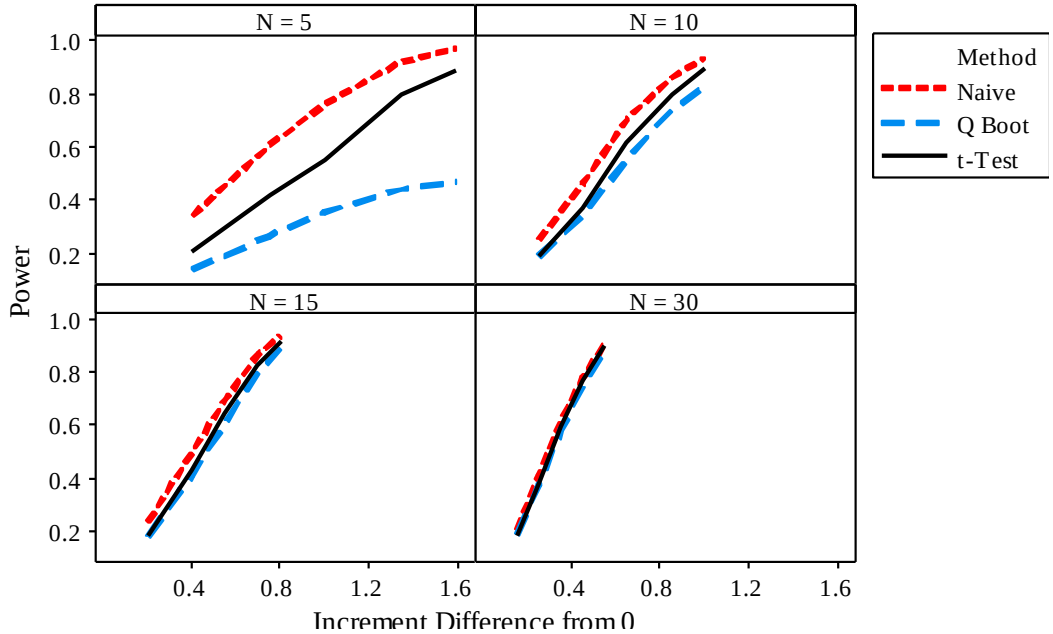

Figure 2-2. Power for 1-Sample Test, X ~ Normal(0,1)

Simulation Results for Uniform

Next, we assessed the performance of the naïve and quantile bootstrap methods against the t-test while breaking the assumptions of the t-test with increasing severity. First, we simulated random samples from the Uniform(-1,1) distribution, following the same procedure as with the standard normal distribution.

As with the standard normal distribution, the naïve method produced liberal level at N = 5 and N = 10. At N = 15, it produced marginal level, but by N = 30, it produced level approximately equal to α. The t-test and quantile bootstrap both produced acceptable level at all sample sizes, with the quantile bootstrap producing very conservative level at the lower sample sizes N = 5 and N = 10.

Most notable about the three methods, in terms of power, is the strong performance of the quantile bootstrap, especially considering the very conservative level it produced. Notice that at N = 10, despite a very conservative level of 0.0245, the quantile bootstrap method produced power almost identical to the t-test, and in fact, slightly exceeds the t-test in power as the increment increases.

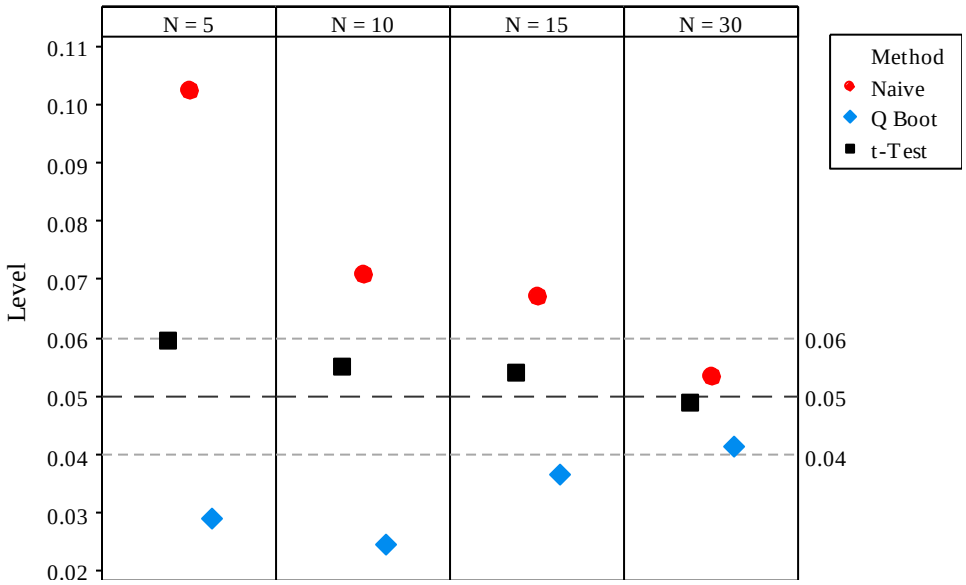

Figure 2-3. Level for 1-Sample Test, X ~ Uniform(-1, 1)

Table 2-2. Level for 1-Sample Test, X ~ Uniform(-1, 1)

N	Naïve	Q boot	t-Test
5	10.25%	2.90%	5.95%
10	7.10%	2.45%	5.50%
15	6.70%	3.65%	5.40%
30	5.35%	4.15%	4.90%

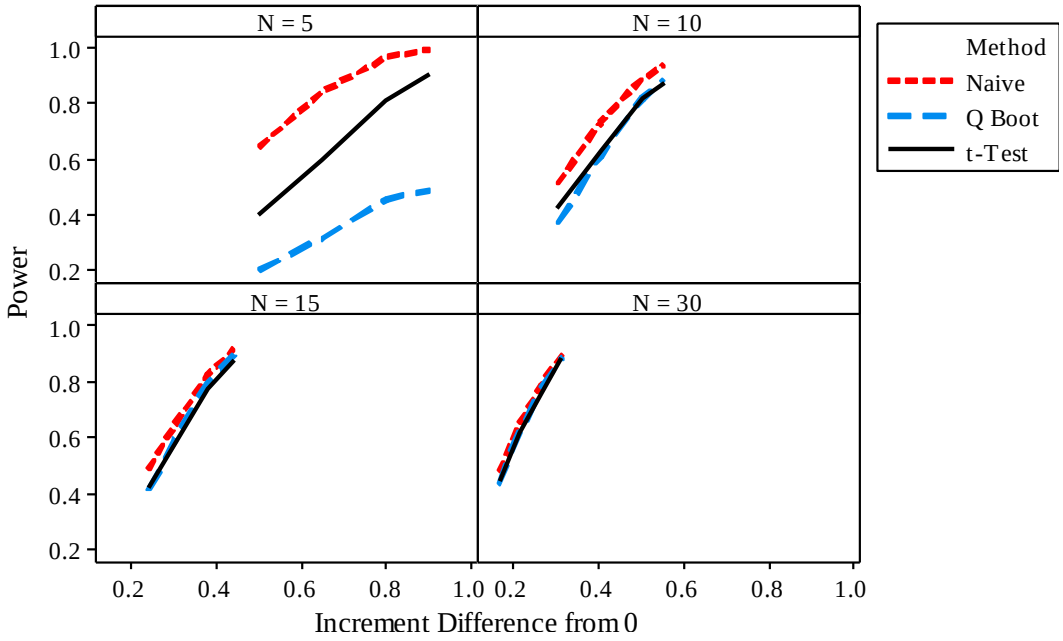

Figure 2-4. Power for 1-Sample Test, X ~ Uniform(-1,1)

Simulation Results for Beta

Next, we simulated data in the same manner as before from a skewed distribution, Beta(5,1), which is skewed left and bounded between 0 and 1. Both the naïve method and the t-test produced levels that were liberal enough to make either test unusable for practical purposes. The quantile bootstrap method, by contrast, performed very well in terms of level.

Unsurprisingly, given the excessively liberal level of both the naïve method and the t-test, the quantile bootstrap method lags behind both tests in power at the lower sample sizes. But perhaps most impressively, the quantile bootstrap method produces almost identical power to the two liberal tests at N = 30.

Table 2-3. Level for 1-Sample Test, X ~ Beta(5,1)

N	Naïve	Q boot	t-Test
5	17.60%	5.70%	12.40%
10	11.30%	5.30%	9.35%
15	10.00%	5.60%	8.65%
30	8.10%	5.40%	7.95%

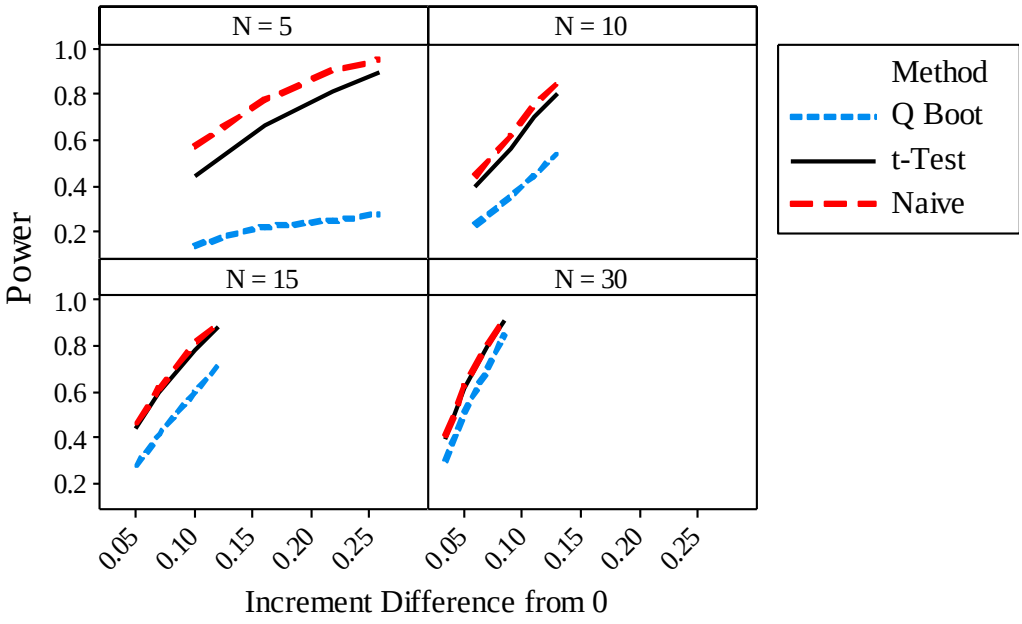

Figure 2-5. Power for 1-Sample Test, X ~ Beta(5,1)

Simulation Results for Discrete

Lastly, we simulated data from a skewed-right discrete distribution of our own invention, with expected value 3.18 and variance 5.09. For this simulation, we found conservative level for both the quantile bootstrap method and the t-test at N = 10 and conservative level for all three methods at N = 30. In terms of power, the quantile bootstrap method outperformed the t-test, and surprisingly, even slightly outperformed the naïve method.

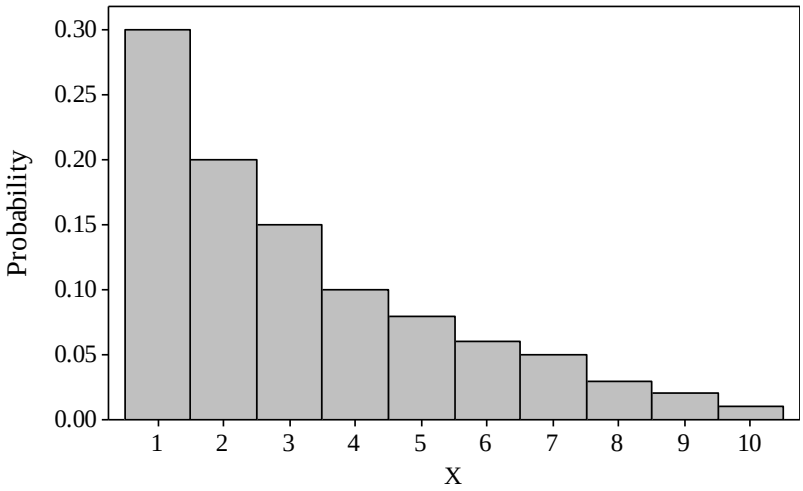

Figure 2-6. Probability Mass Function for Discrete Distribution

Table 2-4. Level for 1-Sample Test, X ~ Discrete

N	Naïve	Q boot	t-Test
10	5.10%	3.85%	2.35%
30	3.50%	3.65%	3.00%

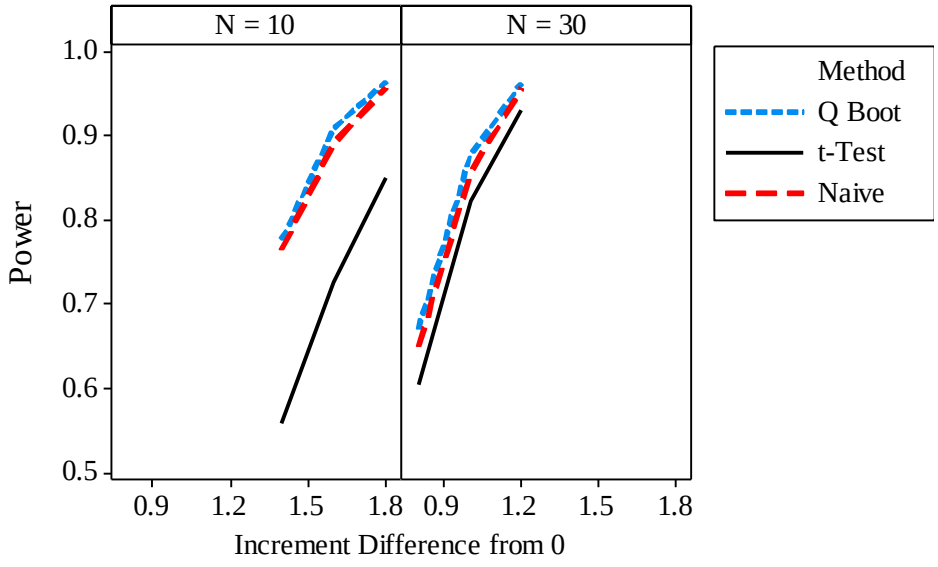

Figure 2-7. Power for 1-Sample Test, X ~ Discrete

Across all simulations, the quantile bootstrap produced consistently acceptable level, even erring on the conservative side for some distributions. As well, the test delivered comparable power, with the only exception being at very small sample sizes, N = 5, and when the t-test and naïve method resulted in excessively liberal level.

Chapter 3 Two-Sample One-Sided Test

Extension of Method to Two-Sample One-Sided Test

Next, we examined the performance of the quantile bootstrap method in the 2-sample case. Assume we have two independent random samples, $X_1 = \{x_{1j},\ 1 \le j \le n_1\}$ from a population with mean μ_1 and variance σ_1^2, and $X_2 = \{x_{2j},\ 1 \le j \le n_2\}$ from a population with mean μ_2 and variance σ_2^2. No assumptions are made about the distributions of either population. We aim to test the null hypothesis H_0: $\mu_1 = \mu_2$ against the 1-sided alternative H_A: $\mu_1 < \mu_2$.

The naïve approach of testing under the alternative hypothesis for the 2-sample 1-sided case is a natural extension of the 1-sample procedure outlined above. Specifically, the naïve procedure is as follows:

1. Sample, with replacement, from X_1, resulting in resample $X_1^* = \{x_{11}^*, \dots, x_{1n_1}^*\}$.

   Sample, with replacement, from X_2, resulting in resample $X_2^* = \{x_{21}^*, \dots, x_{2n_2}^*\}$.

2. Calculate the difference in sample means, $d^* = \bar{x}_2^* - \bar{x}_1^*$, where $\bar{x}_i^* = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}^*$ for $i = 1,2$.

3. Repeat steps 1-2 B times. Denote d_b^* as the difference in sample means of the b^{th} resample of X_1 and X_2.

Just as we did in the 1-sample case, we compute the bootstrap p-value, $P = \frac{1}{B}\sum_{b=1}^{B}(d_b^* < 0)$. Note that a non-zero difference, ∂, could also be tested by computing $P = \frac{1}{B}\sum_{b=1}^{B}(d_b^* < \partial)$. We reject the null if P is less than α.

As expected, just like the naïve method for the 1-sample test, power was very good, but the test was liberal. Again, to correct for the liberal test, we propose the quantile bootstrap method. For the 2-sample 1-sided test, the procedure is as follows:

Stage 1

1. Sample, with replacement, from X_i, resulting in resample $X_i^* = \{x_{i1}^*, \dots, x_{in_i}^*\}$, for $i = 1,2$.
2. Calculate $d^* = \bar{x}_2^* - \bar{x}_1^*$.
3. Repeat steps 1-2 B times. Denote d^{*b} as the difference in sample means of the b^{th} resample of X_1 and X_2.
4. Calculate $p = \frac{1}{B}\sum_{b=1}^{B}(d^{*b} < 0)$

Stage 2

1. Transform the data to represent the null hypothesis. Let $y_{ij} = x_{ij} - \bar{x}_i$, resulting in $Y_i = \{y_{i1}, \dots, y_{in_i}\}$ for $i = 1,2$.
2. Sample, with replacement, from Y_i for $i = 1,2$.
3. Repeat step 2 M times. Denote the m^{th} resample of Y_i as $Y_i^{*m} = \{y_{i1}^{*m}, \dots, y_{in_i}^{*m}\}$.
4. Sample, with replacement, from Y_i^{*m}, $\forall m, i$.
5. Repeat step 4 B times. Denote the b^{th} resample of Y_i^{*m} as $Y_i^{**mb} = \{y_{i1}^{**mb}, \dots, y_{in_i}^{**mb}\}$.
6. Calculate $d^{**mb} = \bar{y}_2^{**mb} - \bar{y}_1^{**mb}$ $\forall m, b$, where $\bar{y}_i^{**mb} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}^{**mb}$.
7. Let $D^{*m} = \bigcup_{b=1}^{B} d^{**mb}$, $m = 1, \dots, M$.
8. Calculate d_p^{*m}, the $(100 * p)^{th}$ quantile of D^{*m}, $\forall m$.
9. Calculate the p-value as $P = \frac{1}{M}\sum_{m=1}^{M}(d_p^{*m} > 0)$.

To assess the performance of the naïve method and quantile bootstrap method in the 2-sample 1-sided test, we performed simulations as described in the 1-sample section, with the obvious exception that we simulated two independent samples at a time instead of one.

When we simulated two random samples from Normal distributions, the quantile bootstrap produced level that was consistent with the appropriate t-test. So, when $\sigma_1^2 = \sigma_2^2$, the pooled variance was used in calculating the test statistic t (notated as "t-Test (eq)"), and when $\sigma_1^2 \neq \sigma_2^2$, the separate variance test statistic was used (notated as "t-Test (ne)"). As expected, the naïve method produced liberal level, especially in the smaller sample sizes ($n_1, n_2 = 10$). The quantile bootstrap also produced comparable power, especially in the larger sample sizes ($n_1, n_2 = 30$).

Table 3-1. Level for 2-Sample 1-Sided Test, X_1, X_2 ~ Normal

σ_1^2	σ_2^2	n_1	n_2	Naïve	Q boot	t-Test
1	1	10	10	6.45%	4.30%	4.45%
1	1	30	30	5.70%	4.95%	5.10%
9	1	10	10	7.10%	4.55%	4.65%
9	1	30	30	5.40%	4.35%	4.25%
9	1	30	10	6.15%	4.85%	5.10%
9	1	10	30	8.00%	4.95%	5.60%

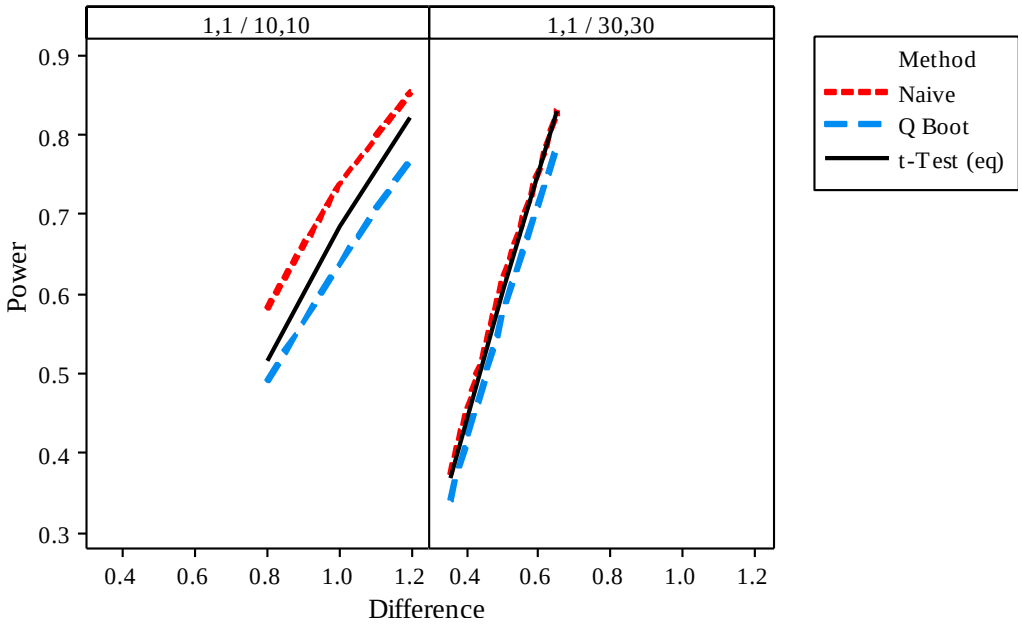

Figure 3-1. Power for 2-Sample, 1-Sided Test for Normal/Normal, homoscedastic, balanced cases. Panel labels represent σ_1^2, σ_2^2 / n_1, n_2

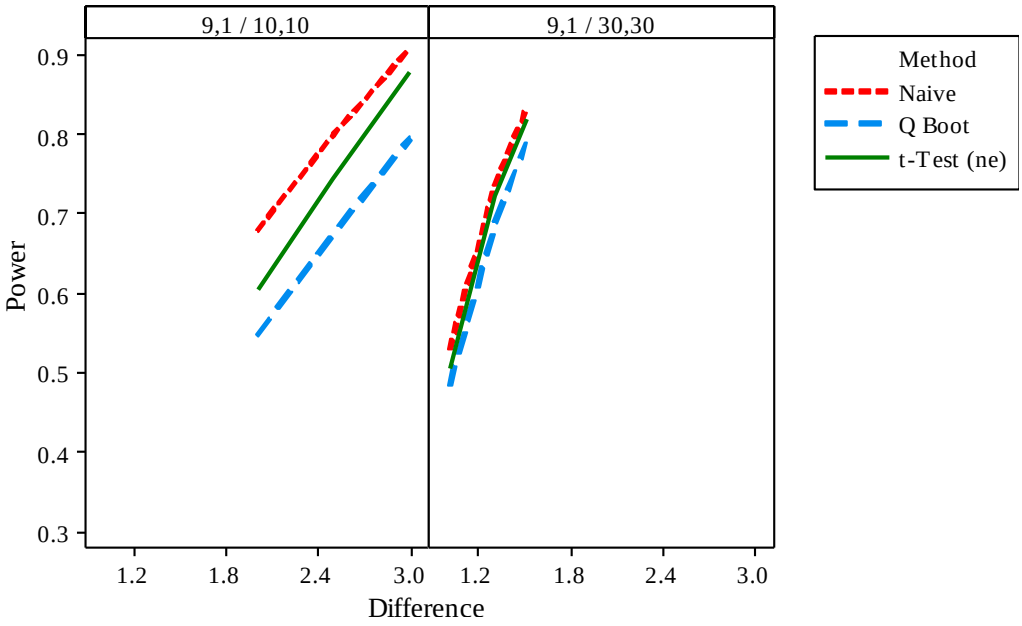

Figure 3-2. Power for 2-Sample, 1-Sided Test for Normal/Normal, heteroscedastic, balanced cases. Panel labels represent σ_1^2, σ_2^2 / n_1, n_2

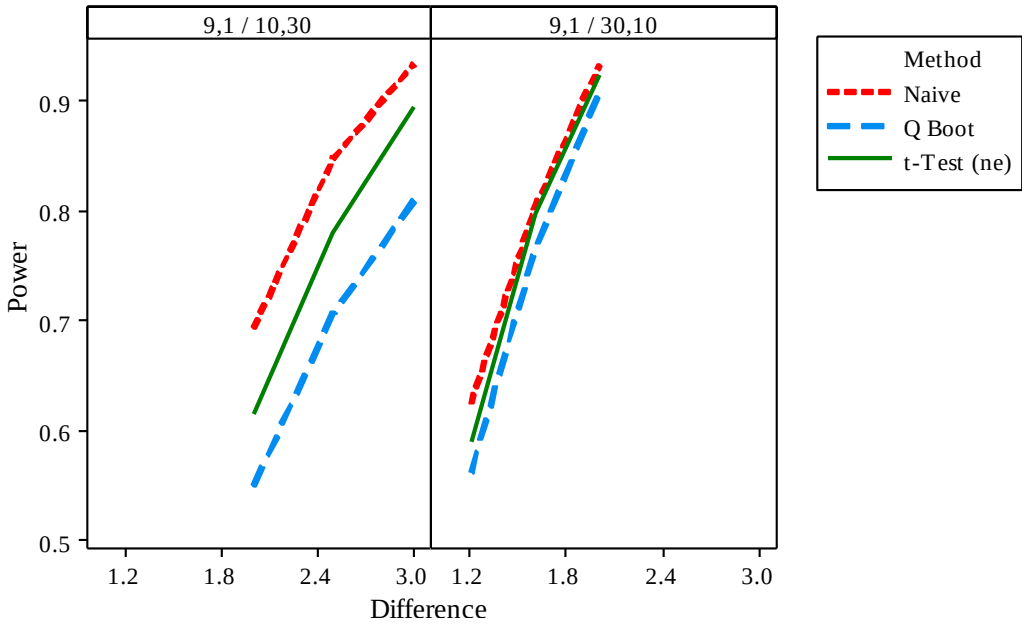

Figure 3-3. Power for 2-Sample, 1-Sided Test for Normal/Normal, heteroscedastic, unbalanced cases. Panel labels represent σ_1^2, σ_2^2 / n_1, n_2

Difficulties arising from skewness

To test the viability of the quantile bootstrap method in non-standard scenarios, we introduced skew into one or both of the samples. Just as we did with the 1-sample method, 2,000 simulations were run with M=200, B=200, α=0.05, and we estimated level/power by the proportion of rejections out of 2,000.

Here we encountered disappointing results. It has been shown that the t-test results in higher level when larger variance is paired with smaller sample sizes (see, for a thorough treatment of the topic, Zimmerman, 2004). Unfortunately, the quantile bootstrap method is not immune to this same problem when one of the distributions is strongly skewed.

When both samples were drawn from a very skewed distribution (Chi-Square(0.5) with skew γ = 4 and Exponential(0.5) with skew γ = 2), the quantile bootstrap method performed just as poorly as the parametric tests, producing unacceptable level error. When only one of the

samples was drawn from a strongly-skewed distribution, the quantile bootstrap method produces accurate level error, except in the case where the sample with the larger variance also had the smaller sample size. For example, see the simulation where $X_1 \sim$ Chi-Square(1), $n_1 = 10$, $\sigma_1^2 = 2$, $\gamma_1 = \sqrt{8} \cong 2.83$ and $X_2 \sim$ Normal(1,1), $n_2 = 30$, $\sigma_1^2 = 1$, $\gamma_2 = 0$, which produced level error of 9.65%. Finally, when only one of the samples is drawn from a skewed distribution and the skew is not very severe, the quantile bootstrap is able to produce accurate level error. For example, see the case where $X_1 \sim$ Chi-Square(8), $n_1 = 10$, $\sigma_1^2 = 16$, $\gamma_1 = 1$ and $X_2 \sim$ Normal(8,3), $n_2 = 30$, $\sigma_1^2 = 9$.

Table 3-2. Level for 2-Sample 1-Sided Test, X_1 ~ Chi-square, X_2 ~ Normal

σ_1^2	σ_2^2	n_1	n_2	γ_1	γ_2	Q Boot	t-Test
4	4	10	10	2	0	6.15%	7.45%
4	4	30	30	2	0	5.45%	6.75%
2	1	10	10	2.83	0	7.25%	9.50%
2	1	30	30	2.83	0	6.35%	9.15%
2	1	30	10	2.83	0	4.55%	5.40%
2	1	10	30	2.83	0	9.65%	12.90%
16	9	10	30	1	0	6.40%	7.70%
16	9	30	10	1	0	5.05%	5.20%
16	25	10	30	1	0	5.80%	6.85%
16	25	30	10	1	0	4.90%	5.35%

Table 3-3. Level for 2-Sample 1-Sided Test, X_1 ~ Chi-square, X_2 ~ Exponential

σ_1^2	σ_2^2	n_1	n_2	γ_1	γ_2	Q Boot	t-Test
1	0.25	10	10	4	2	14.30%	13.40%
1	0.25	30	30	4	2	9.85%	11.95%
1	0.25	10	30	4	2	17.20%	22.15%
1	0.25	30	10	4	2	6.75%	5.40%

Note that when simulating skewed data from known distributions (Chi-square and Exponential), the skew is, in both cases, to the right. So, we devised several discrete distributions with both left and right skew to assess whether or not the direction of skewness resulted in any level error inaccuracy.

Distribution "Left 4.55" is a discrete, left-skewed distribution with variance $\sigma^2 = 4.55$ and skew $\gamma = -0.83$.

Distribution "Right 4.55" is a discrete, right-skewed distribution with variance $\sigma^2 = 4.55$ and skew $\gamma = 0.83$. Finally, distribution "Right 11.03" is a right-skewed distribution with variance $\sigma^2 = 11.03$ and skew $\gamma = 1.22$.

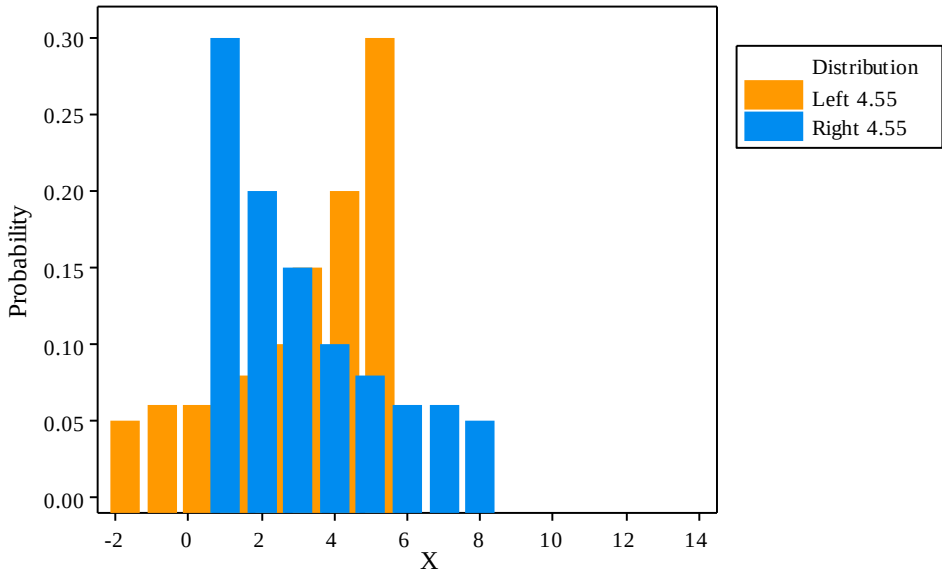

Figure 3-4. Probability Mass Function of Distributions "Left 4.55" and "Right 4.55"

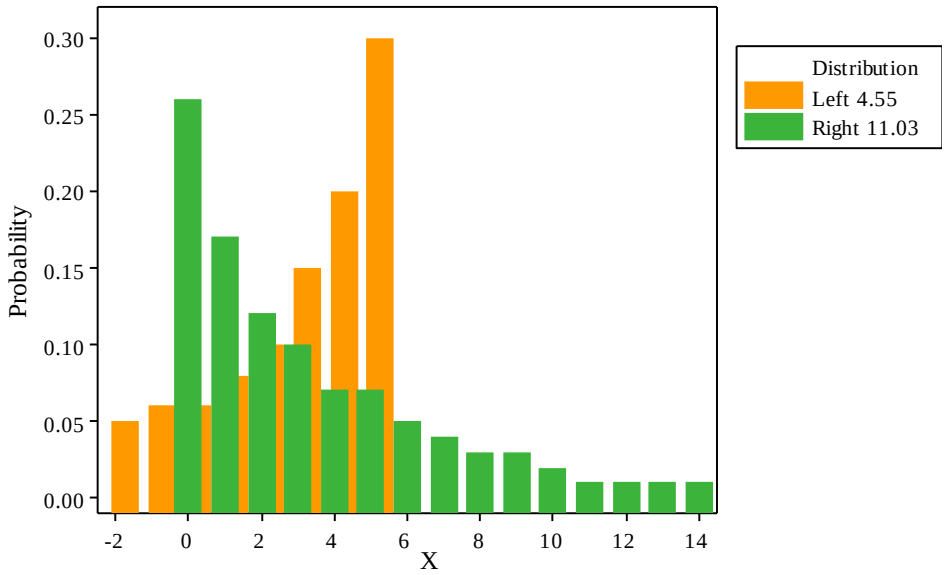

Figure 3-5. Probability Mass Function of Distributions "Left 4.55" and "Right 11.03"

In these simulations, we found that the quantile bootstrap produces satisfactory, if perhaps conservative, level, even in the large variance/small sample size scenario. Notice that the t-test is only liberal when testing whether X_1 pulled from a right-skewed distribution is less than or equal to X_2 pulled from a left-skewed distribution and not when testing whether X_1 pulled from a left-skewed distribution is less than or equal to X_2 pulled from a right-skewed distribution.

A power analysis for the discrete distributions shows that the quantile bootstrap method produces comparable power at the higher sample sizes (30,30), but is less powerful when one or more samples has size 10.

Table 3-4. Level for 2-Sample 1-Sided Test, X_1, X_2 ~ Discrete

γ_1	γ_2	σ_1^2	σ_2^2	n_1	n_2	Q Boot	t Test
L (-0.83)	R (0.83)	4.55	4.55	10	10	2.31%	3.65%
R (0.83)	L (-0.83)	4.55	4.55	10	10	3.06%	7.35%
L (-0.83)	R (0.83)	4.55	4.55	30	30	3.85%	3.45%
R (0.83)	L (-0.83)	4.55	4.55	30	30	4.40%	6.75%
L (-0.83)	R (1.22)	4.55	11.03	10	30	2.80%	3.25%
R (1.22)	L (-0.83)	11.03	4.55	30	10	2.70%	6.65%
L (-0.83)	R (1.22)	4.55	11.03	30	10	4.45%	2.30%
R (1.22)	L (-0.83)	11.03	4.55	10	30	4.90%	9.75%

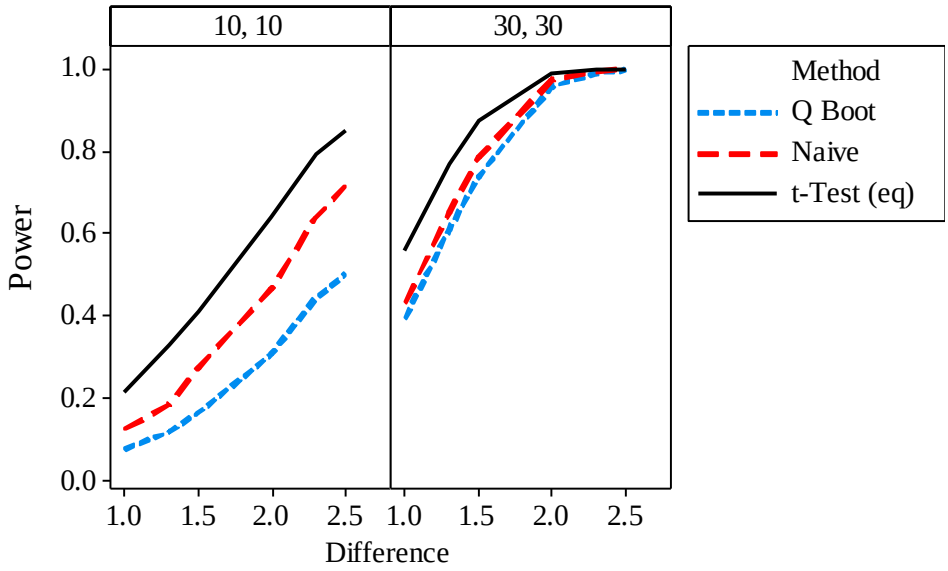

Figure 3-6. Power for 2-Sample, 1-Sided Test, X_1 ~ "Left 4.55", X_2 ~ "Right 4.55". Panel labels represent n_1, n_2

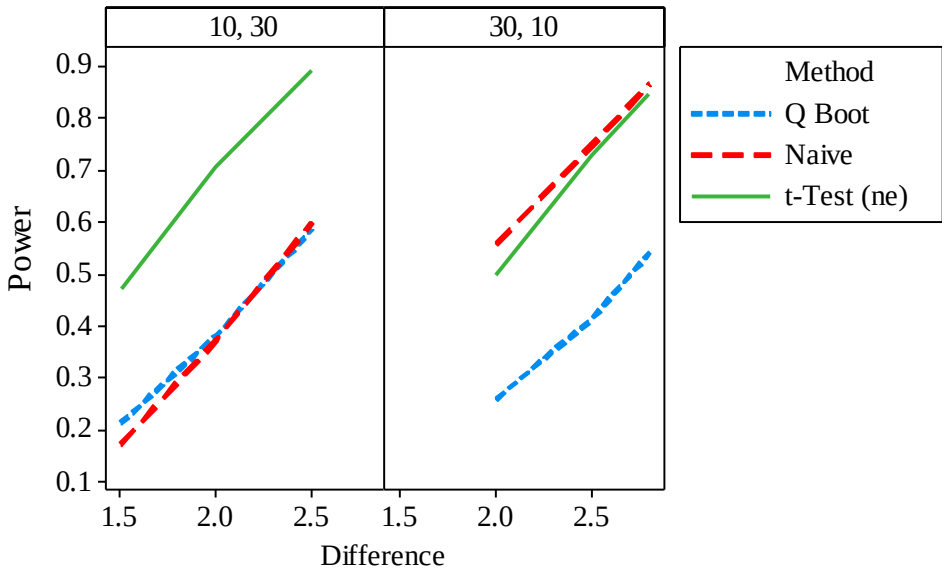

Figure 3-7. Power for 2-Sample, 1-Sided Test, X_1 ~ "Left 4.55", X_2 ~ "Right 11.03". Panel labels represent n_1, n_2

Chapter 4 Tests of Two or More Means

Extension of Method for Tests Comparing Equality of Two or More Means

Finally, we wished to generalize the quantile bootstrap method to the comparison of $r \geq 2$ group means. The model for this comparison is

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$$i = 1, \dots, r$$

$$j = 1, \dots, n_i$$

$$\sum_{i=1}^{r} \alpha_i = 0$$

$$E(\varepsilon_{ij}) = 0$$

$$Var(\varepsilon_{ij}) = \sigma_i^2$$

ε_{ij} independent of $\varepsilon_{i^* j^*}$ for all $i \neq i^*$ and $j \neq j^*$

Note that there are no distribution assumptions placed on the error term ε_{ij}, nor is there a homogeneous variance assumption. The null hypothesis for testing the equality of the group means is $H_0: \alpha_1 = \cdots = \alpha_r$.

While the one-sided two-sample test lends itself to an obvious test statistic, there isn't as obvious a choice when the test is generalized to the comparison of two or more group means. Fisher and Hall (1990) suggest a test statistic, T_2, that they define as

$$T_2 \equiv \sum_i \left\{ n_i(n_i - 1)(\bar{x}_{i\cdot} - x_{\cdot\cdot})^2 \Big/ \sum_j \left(x_{ij} - \bar{x}_{i\cdot}\right)^2 \right\}$$

Where $\bar{x}_{i\cdot} \equiv n_i^{-1} \sum x_{ij}$ and $\bar{x}_{\cdot\cdot}$ is the overall mean, defined as $\bar{x}_{\cdot\cdot} \equiv n^{-1} \sum \sum x_{ij}$ with $n = \sum n_i$. T_2 has the advantage of being asymptotically pivotal, and they show that resampling

methods that utilize the T_2 statistic have improved level error over the standard F statistic in situations where the homogeneous variance assumption is violated.

For our test statistic, we required a statistic that, like T_2 and the standard F statistic, compared the within-group variation to the between-group variation. But, unlike T_2 and F, we required that our statistic had expected value zero, since we do not transform our data to the null hypothesis.

One statistic that satisfies these two requirements is what we call T_3, with obvious reference to Fisher and Hall's T_2.

$$T_3 \equiv \sum_i (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2 - \frac{r-1}{r}\sum_i \left\{ \frac{1}{n_i(n_i-1)} \sum_j \left(x_{ij} - \bar{x}_{i\cdot}\right)^2 \right\}$$

Here, $\bar{x}_{i\cdot} \equiv n_i^{-1} \sum x_{ij}$ is defined the same way as in T_2. However, note that $\bar{x}_{\cdot\cdot}$ is defined here as the least squares mean, $\bar{x}_{\cdot\cdot} \equiv \frac{1}{r}\sum \bar{x}_{i\cdot}$.

The quantile bootstrap procedure for testing $H_0: \alpha_1 = \cdots = \alpha_r$ is an obvious extension of the 1-sample and 2-sample methods detailed previously. For completeness, it is specified here:

Stage 1

1. Let $X_i = \bigcup_j x_{ij}$. Sample, with replacement, from X_i, resulting in resample $X_i^* = \left\{x_{i1}^*, \ldots, x_{in_i}^*\right\}$, for $i = 1, \ldots, r$.

2. Calculate

$$T_3^* = \sum_i (\bar{x}_{i\cdot}^* - \bar{x}_{\cdot\cdot}^*)^2 - \frac{r-1}{r}\sum_i \left\{ \frac{1}{n_i(n_i-1)} \sum_j \left(x_{ij}^* - \bar{x}_{i\cdot}^*\right)^2 \right\}$$

3. Repeat steps 1-2 B times. Denote ${T_3}^{*b}$ as the T_3^* statistic of the b^{th} resample of $X_1, \ldots, X_r$.

4. Calculate $p = \frac{1}{B}\sum_{b=1}^{B}\left({T_3}^{*b} < 0\right)$

Stage 2

1. Transform the data to represent the null hypothesis. Let $y_{ij} = x_{ij} - \bar{x}_i$, resulting in

$Y_i = \{y_{i1}, \ldots, y_{in_i}\}$ for $i = 1, \ldots, r$.

2. Sample, with replacement, from Y_i, $\forall i$.

3. Repeat step 2 M times. Denote the m^{th} resample of Y_i as $Y_i^{*m} = \{y_{i1}^{*m}, \ldots, y_{in_i}^{*m}\}$.

4. Sample, with replacement, from Y_i^{*m}, $\forall m, i$.

5. Repeat step 4 B times. Denote the b^{th} resample of Y_i^{*m} as $Y_i^{**mb} = \{y_{i1}^{**mb}, \ldots, y_{in_i}^{**mb}\}$.

6. Calculate

$$T_3^{**mb} = \sum_i\left(\bar{y}_{i\cdot}^{**mb} - \bar{y}_{\cdot\cdot}^{**mb}\right)^2 - \frac{r-1}{r}\sum_i\left\{\frac{1}{n_i(n_i-1)}\sum_j\left(y_{ij}^{**mb} - \bar{y}_{i\cdot}^{**mb}\right)^2\right\}, \forall m, b,$$

where $\bar{y}_i^{**mb} = \frac{1}{n}\sum_{j=1}^{n} y_{ij}^{**mb}$.

7. Let $\mathbb{T}_3^{*m} = \bigcup_{b=1}^{B} T_3^{**mb}$, $m = 1, \ldots, M$.

8. Calculate ${T_3}_p^{*m}$, the p^{th} quantile of $\mathbb{T}_3^{*m}$, $\forall m$.

9. Calculate the p-value as $P = \frac{1}{M}\sum_{m=1}^{M}\left({T_3}_p^{*m} > 0\right)$.

As previously, the naïve approach is stage 1 alone.

Simulation Results for r = 2

We began assessing the performance of the quantile bootstrap method with the T_3 statistic by reconsidering the two-sample test ($r = 2$), but as a two-sided test. We also included the method proposed by Fisher and Hall, both because our test statistic is based on their T_2

statistic and because theirs represents a bootstrap procedure that also makes no distribution or homogeneity assumptions but utilizes a transformation in order to test under the null hypothesis. We expected that both the quantile bootstrap method and Fisher and Hall's method (noted as the T_2 method from this point forward) would produce accurate level (as their own simulation studies showed), but expected the quantile bootstrap method to outperform theirs in terms of power. Our expectations were mostly met, with two main exceptions. First, the quantile bootstrap method produced excessively liberal level when one or more of the groups was severely skewed and the group with larger variance also had smaller sample size. Second, the quantile bootstrap method did not outperform the T_2 method in power by any notable degree when all groups were Normally distributed.

In all other cases we simulated, the quantile bootstrap method produced accurate, if perhaps conservative level, while improving on the power produced by the T_2 method.

When both groups are normally distributed, all methods produce accurate level. Perhaps surprisingly, the F test produced accurate level even under strongly heterogeneous scenarios, although that would not be the case when the groups were heterogeneous and unbalanced. This result matches other studies; the F test produces excessively liberal level when larger variance is paired with smaller sample size and produces very conservative level when larger variance is paired with larger sample size. For example, see Zimmerman, 2004. Notably, the T_2 method also shows this pattern, although not to an excessive degree.

The primary limitation of the quantile bootstrap method is when one or both of the sample groups is strongly skewed, and the group with the larger variance is paired with the smaller sample size, as can be seen in Table 4.2, where the quantile bootstrap produced level error of 8.45%.

Although, just as we saw with the one-sided tests, the quantile bootstrap method did not have the same problems with level when both groups were skewed, but in opposite directions. In

that case, even when a larger variance was paired with a smaller sample size, the quantile bootstrap method produced accurate level.

In terms of power, the quantile bootstrap method outperforms or matches the T_2 method in all cases except in the Normal, heterogeneous, unbalanced case when the group with the larger variance also has the smaller sample size. As well, the T_2 method greatly outperforms the quantile method in all of the discrete simulations.

Table 4-1. Level for r = 2, X_1 ~ Normal, X_2 ~ Normal

σ_1^2	σ_2^2	n_1	n_2	Naïve	Q boot	F-Test	T_2
1	1	10	10	2.05%	3.95%	4.20%	2.75%
1	1	30	30	1.25%	5.00%	5.10%	4.85%
9	1	10	10	3.20%	4.70%	5.90%	4.15%
9	1	30	30	1.40%	4.75%	5.50%	4.95%
9	1	10	30	3.75%	4.95%	23.55%	6.65%
9	1	30	10	0.85%	3.90%	0.30%	2.00%

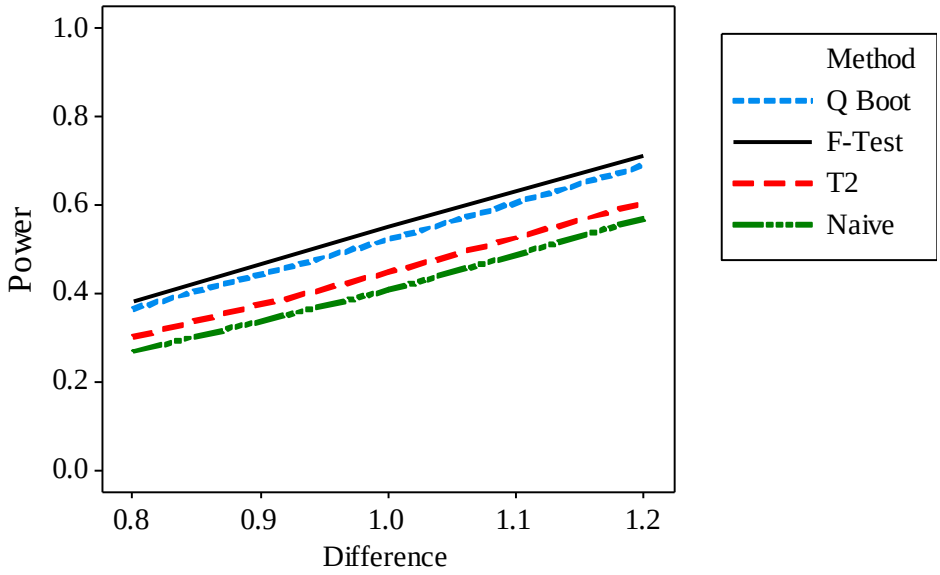

Figure 4-1. Power for r = 2, X_1 ~ Normal, X_2 ~ Normal; σ_1^2, σ_2^2 = 1, 1; n_1, n_2 = 10, 10

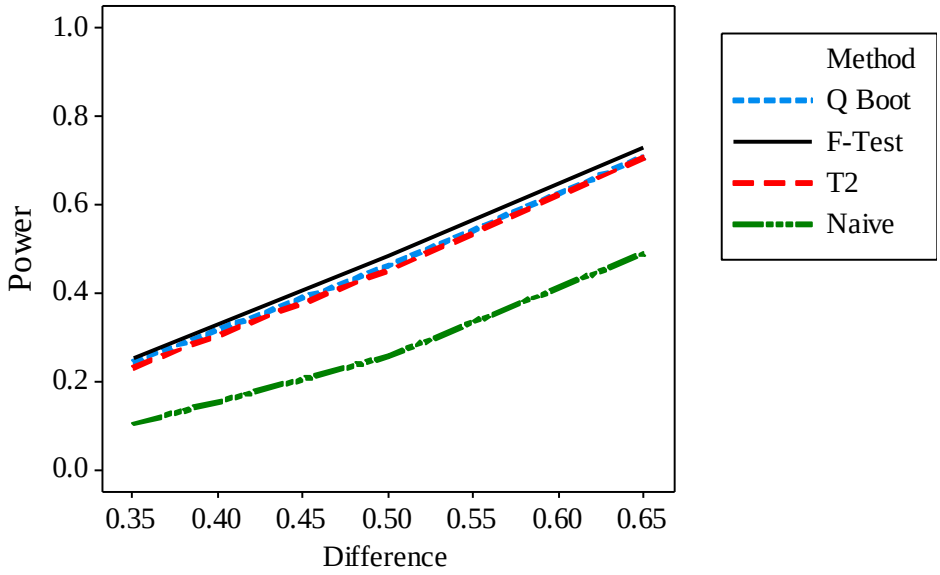

Figure 4-2. Power for r = 2, X_1 ~ Normal, X_2 ~ Normal; σ_1^2, σ_2^2 = 1, 1; n_1, n_2 = 30, 30

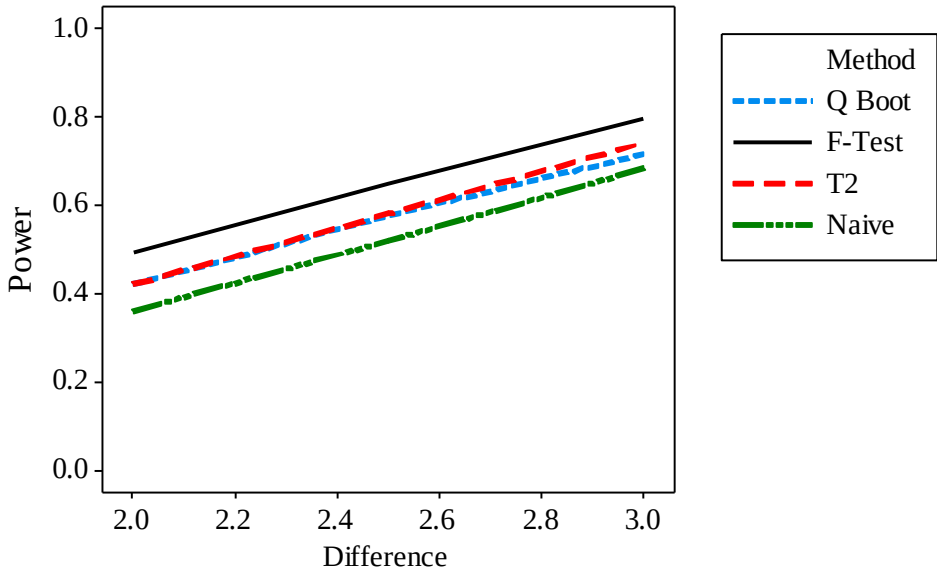

Figure 4-3. Power for r = 2, X_1 ~ Normal, X_2 ~ Normal; σ_1^2, σ_2^2 = 9, 1; n_1, n_2 = 10, 10

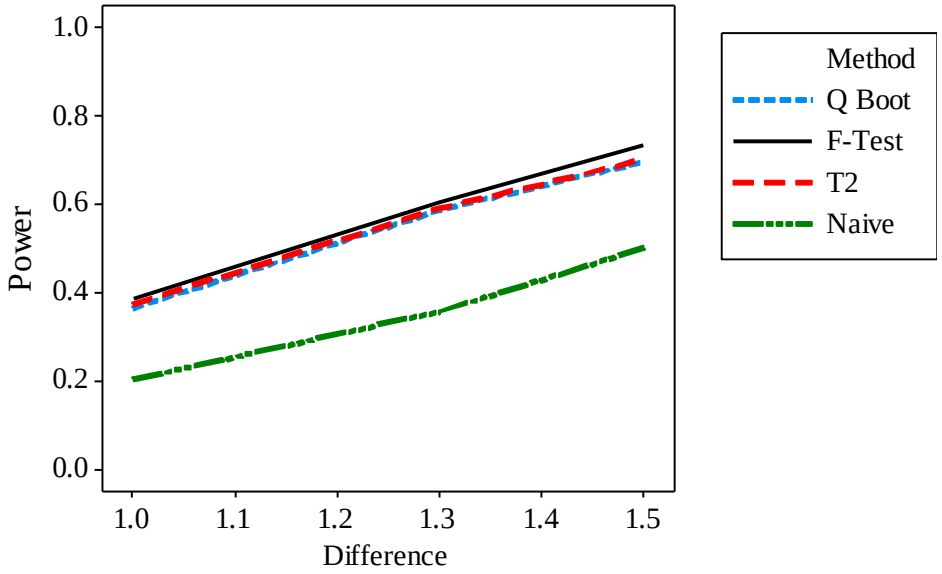

Figure 4-4. Power for r = 2, X_1 ~ Normal, X_2 ~ Normal; σ_1^2, σ_2^2 = 9, 1; n_1, n_2 = 30, 30

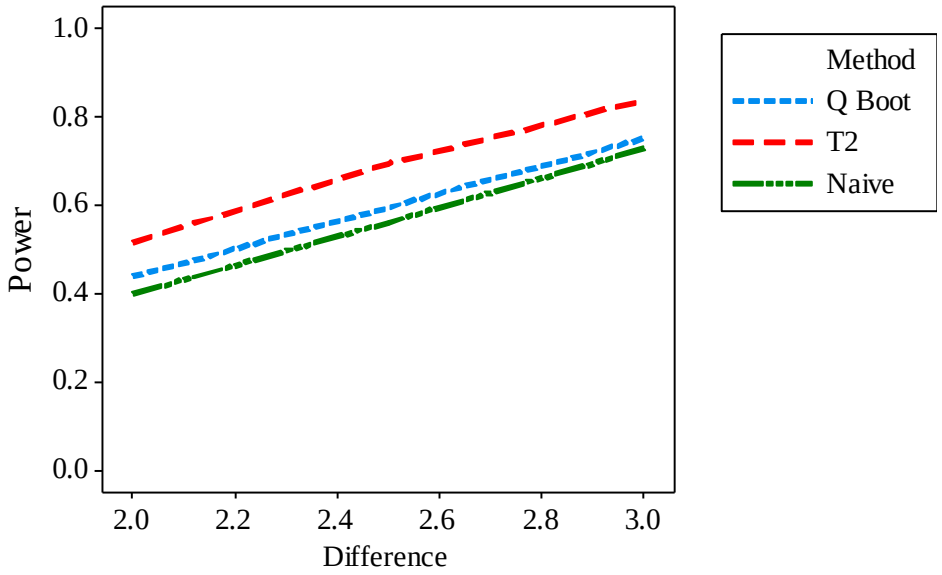

Figure 4-5. Power for r = 2, X_1 ~ Normal, X_2 ~ Normal; σ_1^2, σ_2^2 = 9, 1; n_1, n_2 = 10, 30

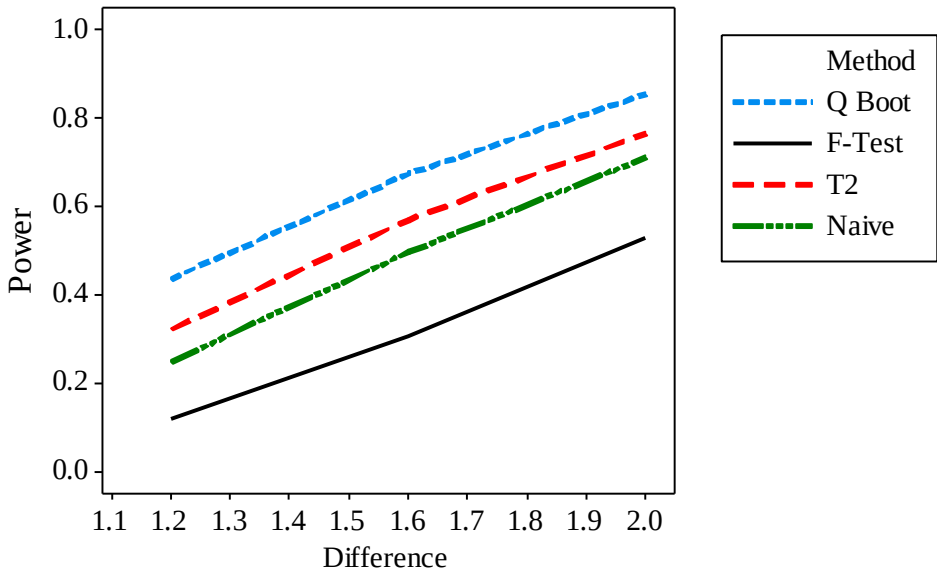

Figure 4-6. Power for r = 2, X_1 ~ Normal, X_2 ~ Normal; σ_1^2, σ_2^2 = 9, 1; n_1, n_2 = 30, 10

Table 4-2. Level for r = 2, X_1 ~ Chi-square, X_2 ~ Normal

σ_1^2	σ_2^2	n_1	n_2	Naïve	Q Boot	F-Test	T_2
4	4	10	10	2.60%	5.00%	5.20%	3.10%
4	4	30	30	1.40%	4.60%	5.25%	3.65%
2	1	10	10	3.65%	5.00%	6.35%	3.00%
2	1	30	30	1.80%	5.50%	5.95%	4.50%
2	1	10	30	5.65%	8.45%	8.60%	4.60%
2	1	30	10	2.10%	4.40%	3.90%	2.50%

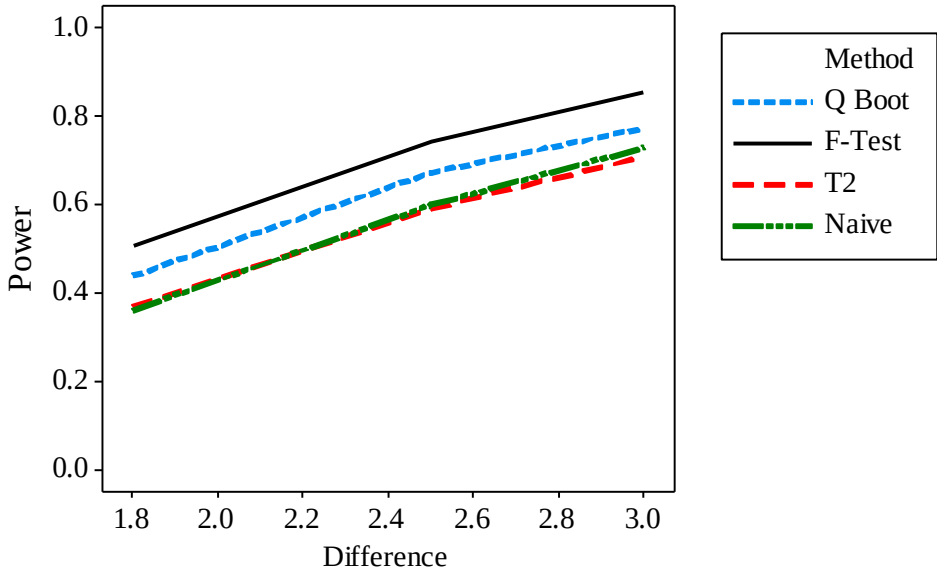

Figure 4-7. Power for r = 2, X_1 ~ Chi-square, X_2 ~ Normal; σ_1^2, σ_2^2 = 4, 4; n_1, n_2 = 10, 10

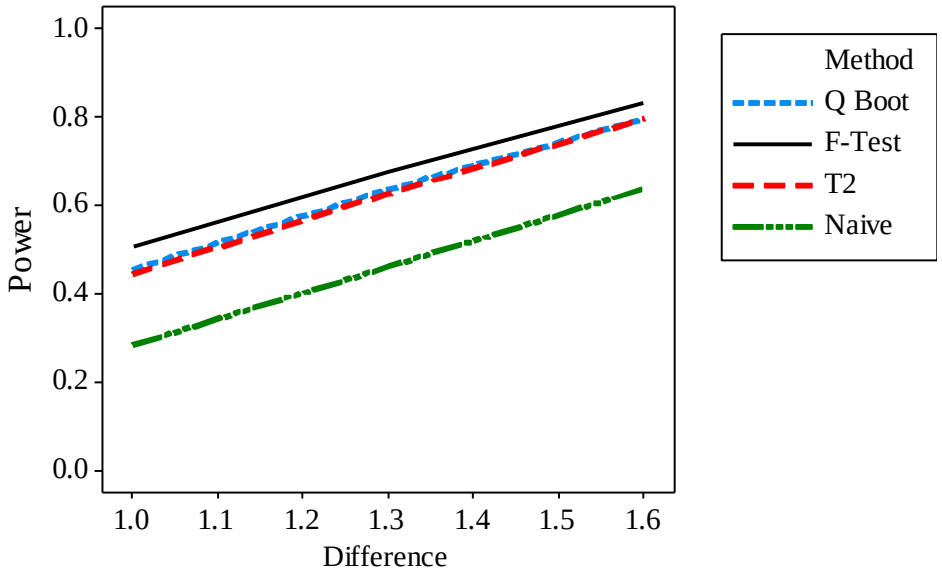

Figure 4-8. Power for r = 2, X_1 ~ Chi-square, X_2 ~ Normal; σ_1^2, σ_2^2 = 4, 4; n_1, n_2 = 30, 30

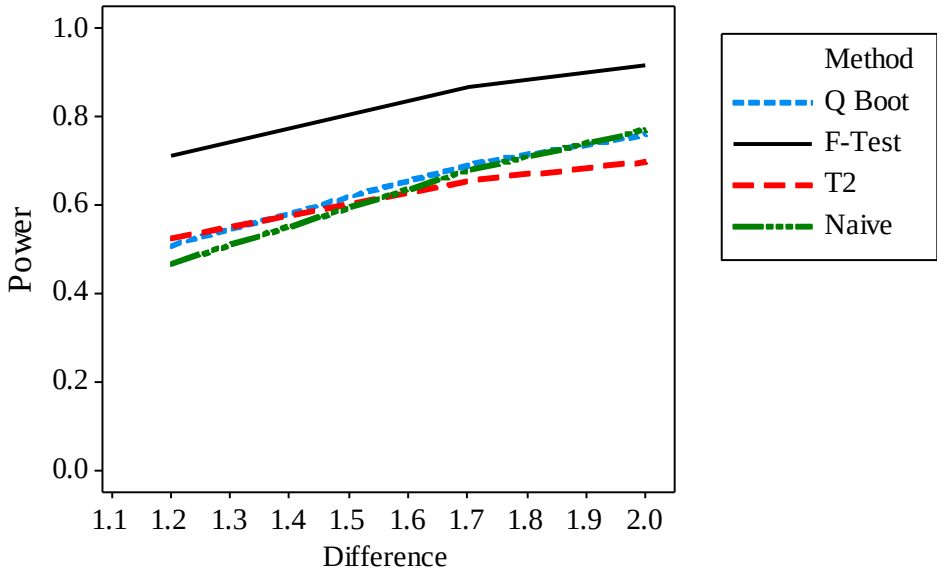

Figure 4-9. Power for r = 2, X_1 ~ Chi-square, X_2 ~ Normal; σ_1^2, σ_2^2 = 2, 1; n_1, n_2 = 10, 10

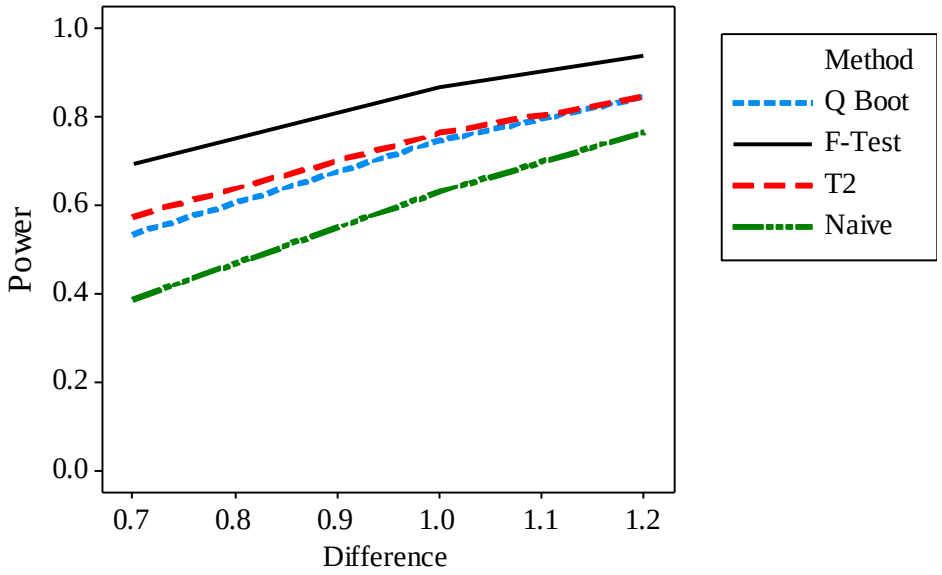

Figure 4-10. Power for r = 2, X_1 ~ Chi-square, X_2 ~ Normal; σ_1^2, σ_2^2 = 2, 1; n_1, n_2 = 30, 30

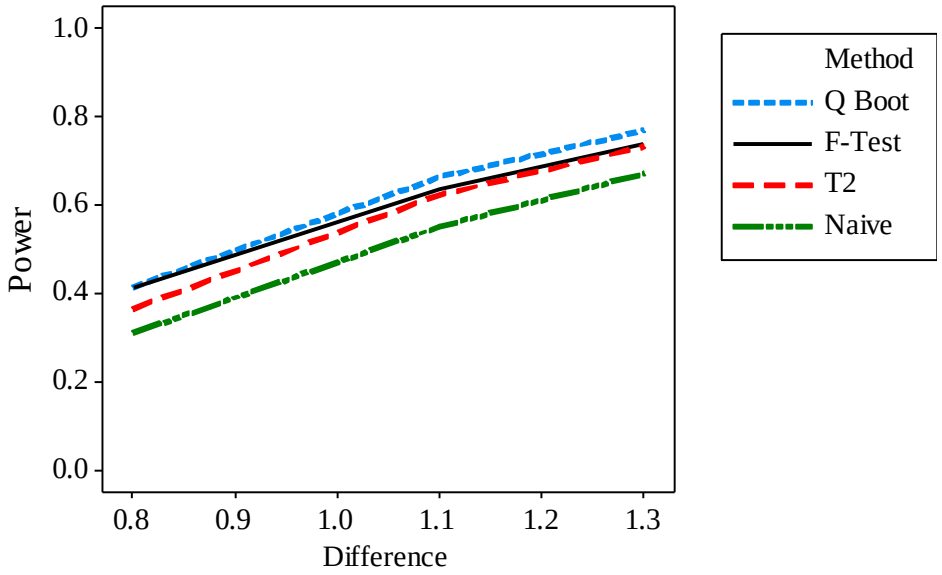

Figure 4-11. Power for r = 2, X_1 ~ Chi-square, X_2 ~ Normal; σ_1^2, σ_2^2 = 2, 1; n_1, n_2 = 30, 10

Table 4-3. Level for r = 2, X_1 ~ Chi-square, X_2 ~ Exponential

σ_1^2	σ_2^2	n_1	n_2	Naïve	Q Boot	F-Test	T_2
1	0.25	10	10	7.45%	7.80%	9.40%	2.70%
1	0.25	30	30	4.70%	8.00%	9.40%	6.00%
1	0.25	10	30	12.00%	15.50%	16.55%	6.15%
1	0.25	30	10	2.40%	3.60%	4.25%	1.80%

Table 4-4. Level for r = 2, X_1 ~ Discrete, X_2 ~ Discrete

γ_1	γ_2	σ_2^2	σ_1^2	n_2	n_1	Naïve	Q Boot	F-Test	T_2
L (-0.83)	R (0.83)	4.55	4.55	10	10	2.75%	4.70%	8.85%	4.20%
L (-0.83)	R (0.83)	4.55	4.55	30	30	1.20%	4.90%	6.35%	4.60%
L (-0.83)	R (1.22)	4.55	11.03	10	30	4.30%	6.25%	4.05%	3.80%
L (-0.83)	R (1.22)	4.55	11.03	30	10	2.05%	4.50%	4.85%	4.15%

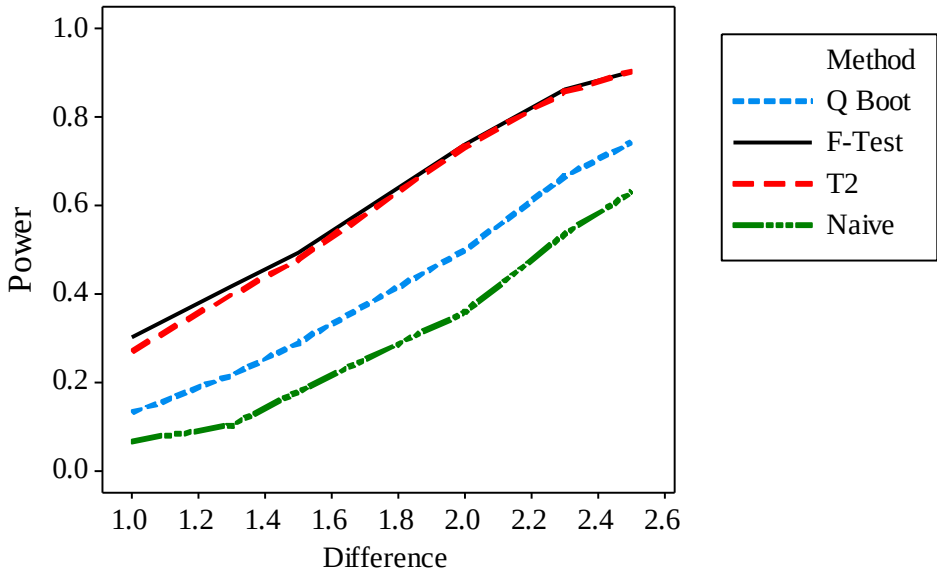

Figure 4-12. Power for r = 2, X_1 ~ "Left 4.55", X_2 ~ "Right 4.55"; n_1, n_2 = 10, 10

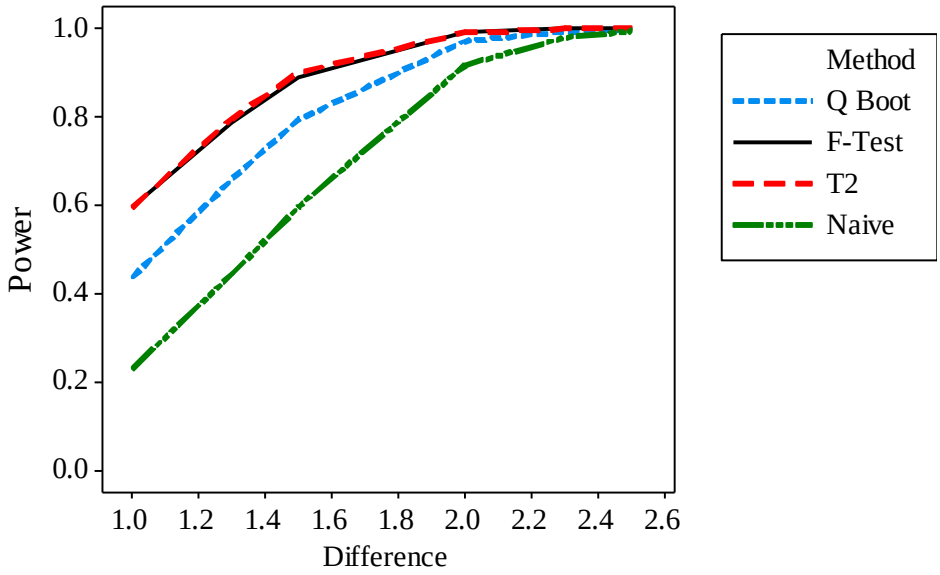

Figure 4-13. Power for r = 2, X_1 ~ "Left 4.55", X_2 ~ "Right 4.55"; n_1, n_2 = 30, 30

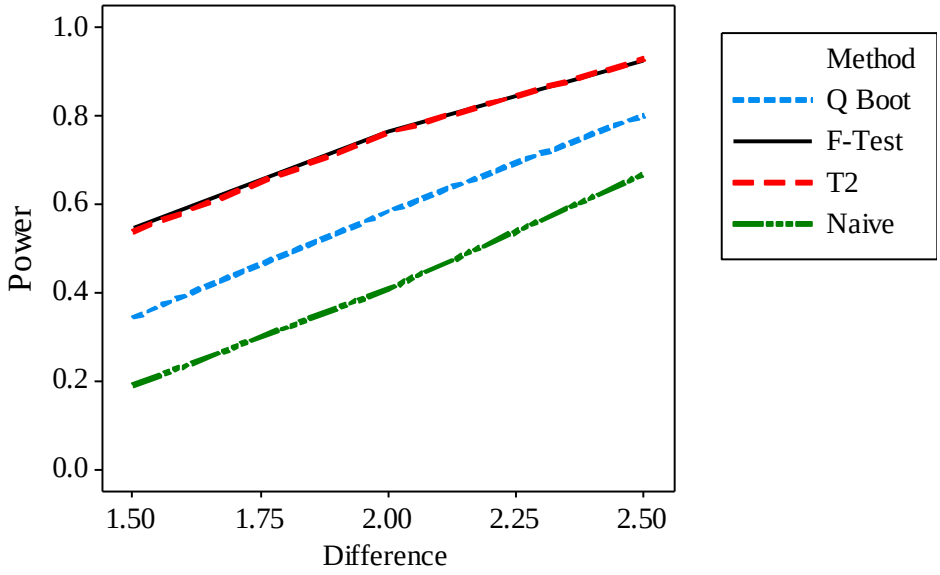

Figure 4-14. Power for r = 2, X_1 ~ "Left 4.55", X_2 ~ "Right 11.03"; n_1, n_2 = 10, 30

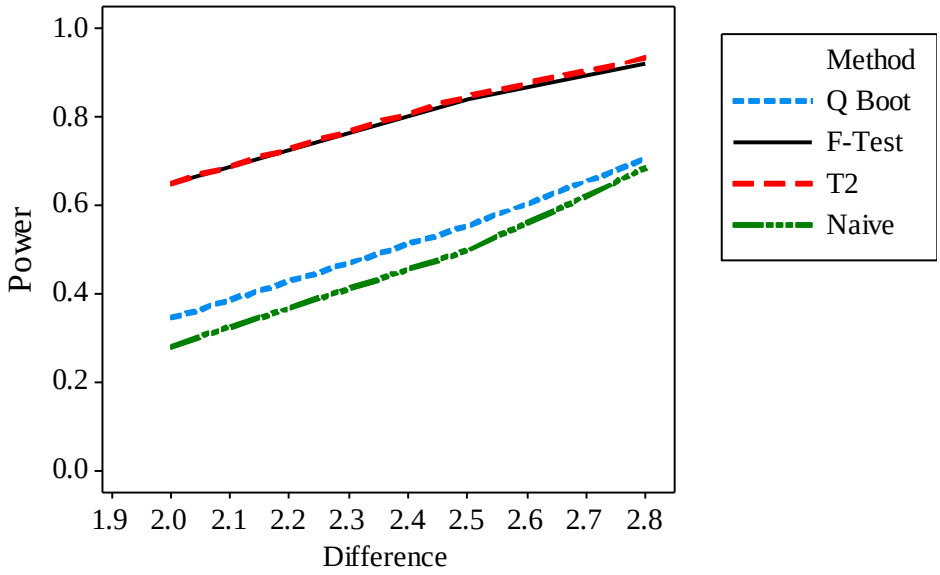

Figure 4-15. Power for r = 2, X_1 ~ "Left 4.55", X_2 ~ "Right 11.03"; n_1, n_2 = 30, 10

Three or More Groups

Simulations performed at 3, 5, and 10 levels produced results consistent with those shown in the 2-level (r = 2) cases.

The skewness problem exhibited in the r = 2 section remains a problem at higher numbers of levels. Specifically, simulations were performed at r = 3 where X_1 was drawn from Chi-square(2), while X_2 and X_3 were drawn from a Normal distribution. In the heteroscedastic, unbalanced cases where groups with larger variance also had smaller sample size ("Heteroscedastic, Unbalanced, Small/Large"), the quantile bootstrap method produced an inaccurate level error of 9.30%. No simulations were performed at 5 or 10 levels that included one or more skewed groups, but we expect the same behavior would be observable.

For simulations performed at 5 and 10 levels, all groups were drawn from the Normal distribution. Accurate level error was achieved by all methods (other than the naïve method) for all cases, as expected. In addition, the quantile bootstrap method outperformed the T_2 method in terms of power for all cases except the heteroscedastic, unbalanced, small/large case, where it

produced very inadequate power compared to the T_2 method. This was consistent across 3, 5 and 10 levels.

For these simulations, it should be noted that power was calculated by adding an increment difference to only one of the groups. So, for example, at r = 5, power at Difference = 1.0 was calculated by drawing X_1 from a normal distribution with mean $\mu = 1$, while drawing $X_2,\ldots, X_5$ from a normal distribution each with mean $\mu = 0$. In this way, the "Difference" reported in the power plots could be interpreted as maximum difference between means. As well, the power plots only include those methods that produce level error ≤ 7% for the respective case. So, for example, Figure 4-24 (r = 3, Heteroscedastic, Unbalanced, Small/Large) does not display power for the F test because the level error for that case is 18%.

Simulation Results for r = 3

Table 4-5. Details of r = 3 simulation cases, X_1 ~ Chi-square, X_2 and X_3 ~ Normal

		Homoscedastic, Balanced		Heteroscedastic, Unbalanced, Large/Large		Heteroscedastic, Unbalanced, Small/Large	
i	Distribution	n_i	σ_i^2	n_i	σ_i^2	n_i	σ_i^2
1	Chi-Square	10/30	1	15	4	5	4
2	Normal	10/30	1	10	1	10	1
3	Normal	10/30	1	5	1	15	1

Table 4-6. Level for r = 3, X_1 ~ Chi-square, X_2 and X_3 ~ Normal

Description	Naive	Q Boot	F-Test	T_2
Homoscedastic, Balanced, n = 10	5.60%	5.10%	5.40%	2.50%
Homoscedastic, Balanced, n = 30	2.75%	5.35%	5.35%	5.20%
Heteroscedastic, Unbalanced, Large/Large	7.35%	5.20%	4.65%	0.80%
Heteroscedastic, Unbalanced, Small/Large	12.25%	9.30%	13.50%	4.00%

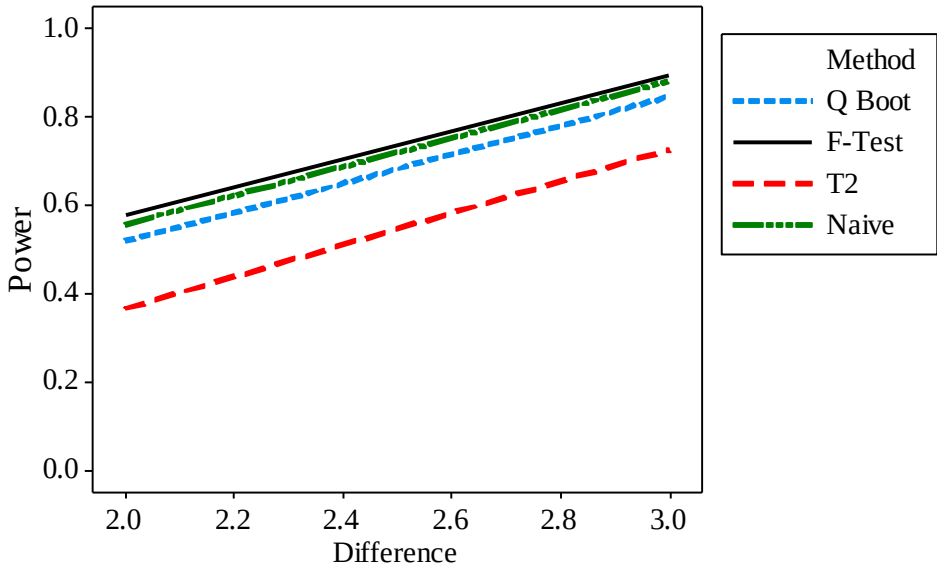

Figure 4-16. Power for r = 3, X1 ~ Chi-square, X2 and X3 ~ Normal; Homoscedastic, Balanced, N = 10

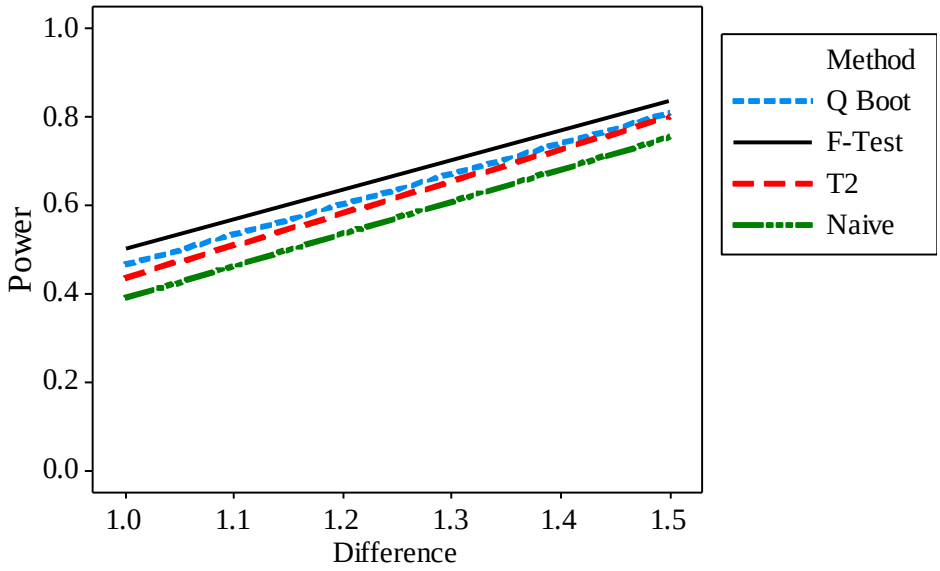

Figure 4-17. Power for r = 3, X_1 ~ Chi-square, X_2 and X_3 ~ Normal; Homoscedastic, Balanced, N = 30

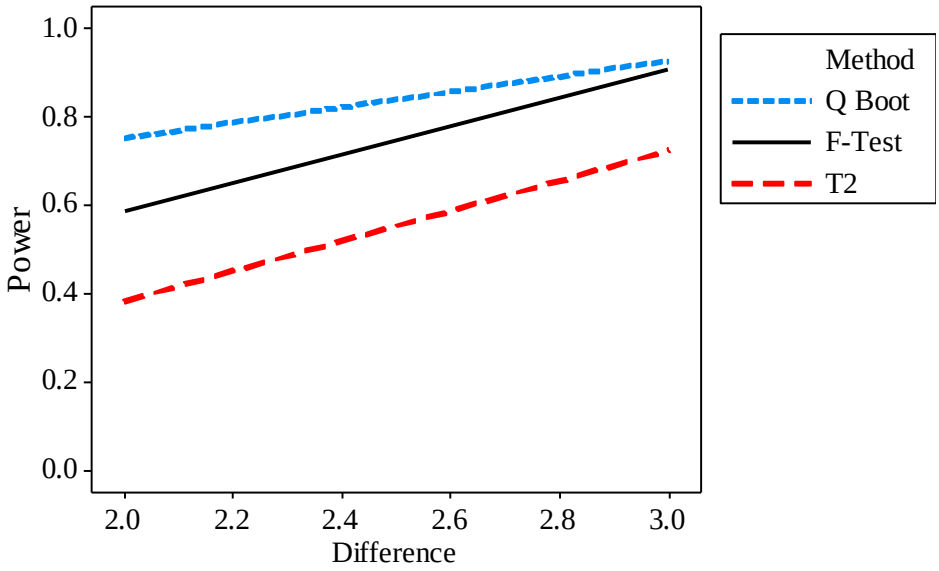

Figure 4-18. Power for r = 3, X_1 ~ Chi-square, X_2 and X_3 ~ Normal; Heteroscedastic, Unbalanced, Large/Large

Table 4-7. Details of r = 3 simulation cases, all groups drawn from Normal distribution

	Homoscedastic, Balanced		Heteroscedastic, Balanced		Heteroscedastic, Unbalanced, Large/Large		Heteroscedastic, Unbalanced, Small/Large	
i	n_i	σ_i^2	n_i	σ_i^2	n_i	σ_i^2	n_i	σ_i^2
1	10/30	1	10/30	1	5	1	5	9
2	10/30	1	10/30	4	10	4	10	4
3	10/30	1	10/30	9	15	9	15	1

Table 4-8. Level for r = 3, all groups drawn from Normal distribution

Description	Naïve	Q Boot	F-Test	T_2
Homoscedastic, Balanced, n = 10	5.40%	5.05%	5.05%	2.90%
Homoscedastic, Balanced, n = 30	2.55%	5.10%	5.10%	5.40%
Heteroscedastic, Balanced, n = 10	4.95%	4.85%	7.25%	2.90%
Heteroscedastic, Balanced, n = 30	2.45%	4.60%	6.10%	3.70%
Heteroscedastic, Unbalanced, Large/Large	4.95%	5.45%	2.40%	0.85%
Heteroscedastic, Unbalanced, Small/Large	10.40%	4.95%	18.00%	2.70%

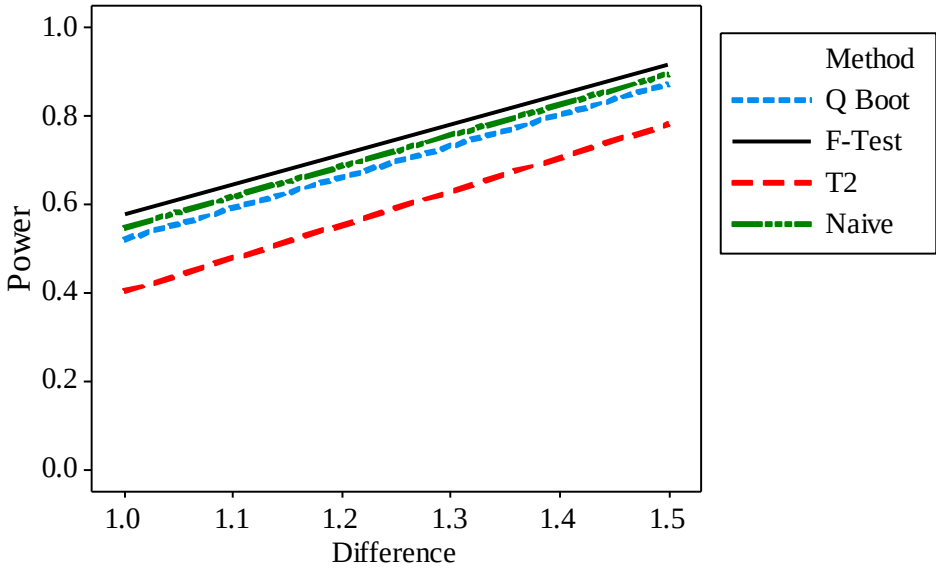

Figure 4-19. Power for r = 3, all groups Normal; Homoscedastic, Balanced, n = 10

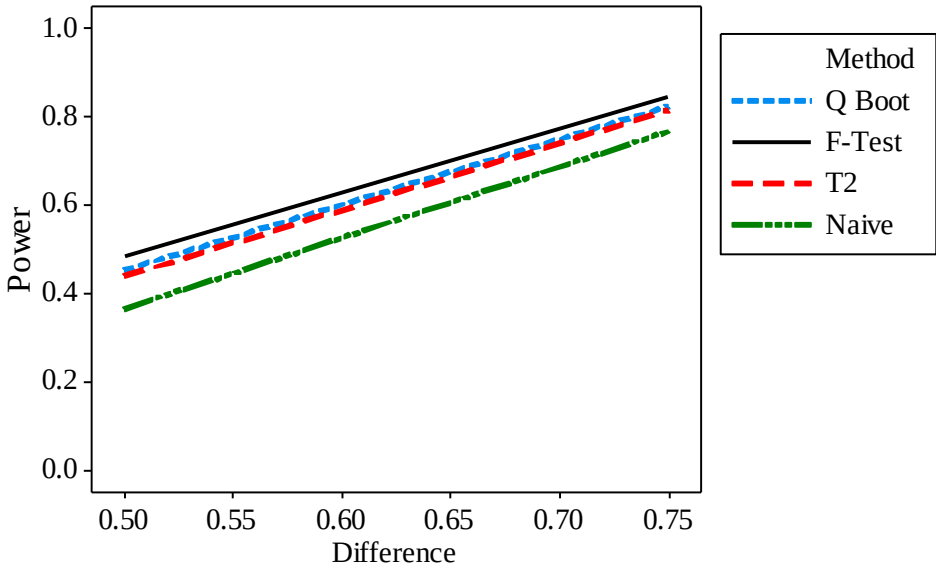

Figure 4-20. Power for r = 3, all groups Normal; Homoscedastic, Balanced, n = 30

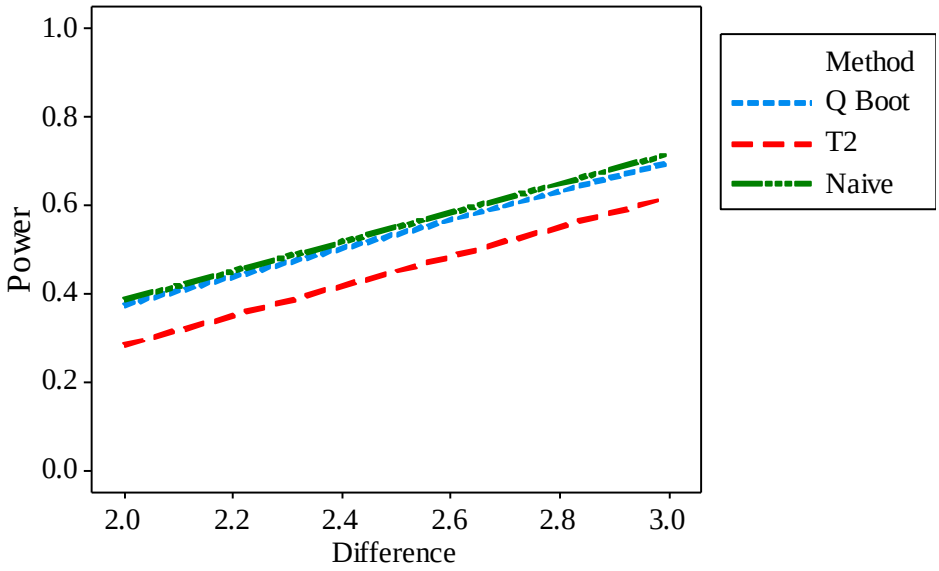

Figure 4-21. Power for r = 3, all groups Normal; Heteroscedastic, Balanced, n = 10

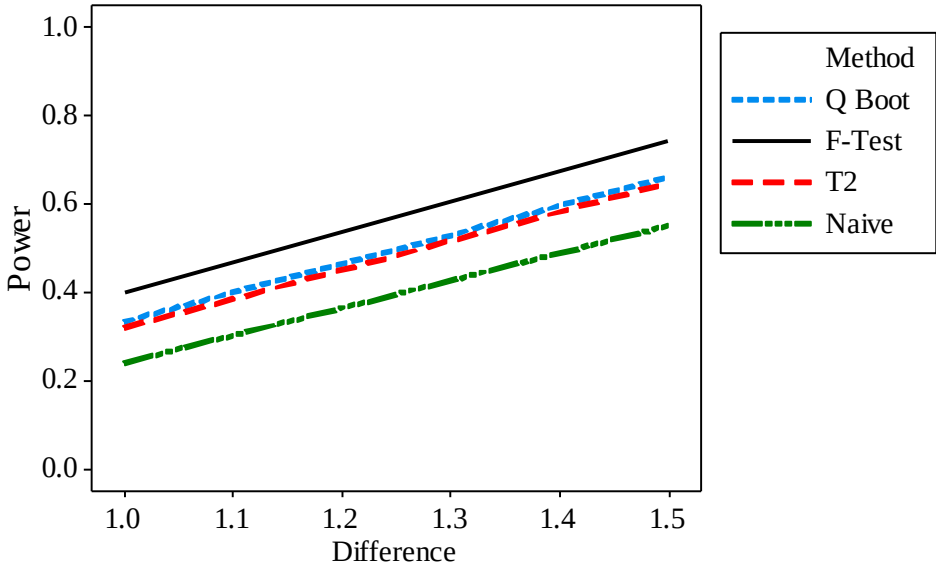

Figure 4-22. Power for r = 3, all groups Normal; Heteroscedastic, Balanced, n = 30

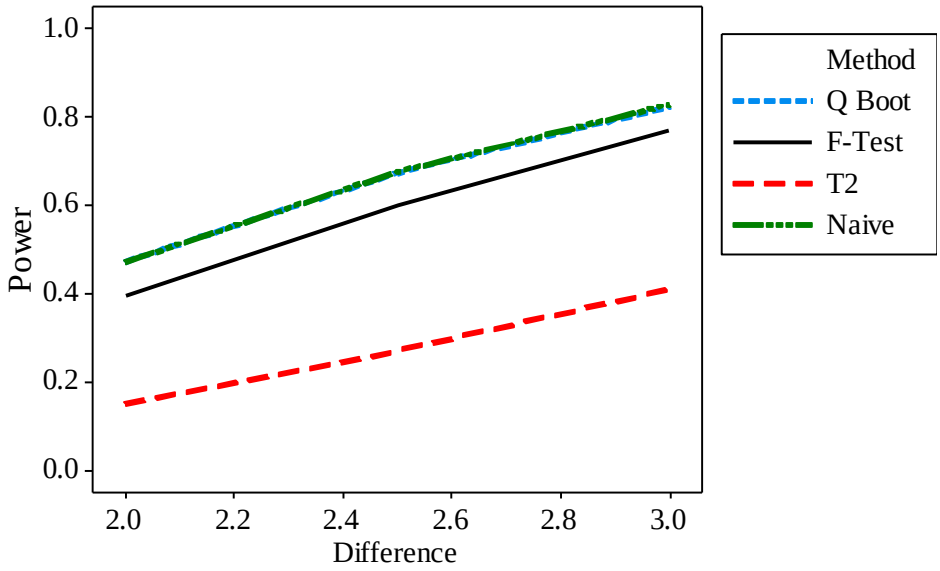

Figure 4-23. Power for r = 3, all groups Normal; Heteroscedastic, Unbalanced, Large/Large

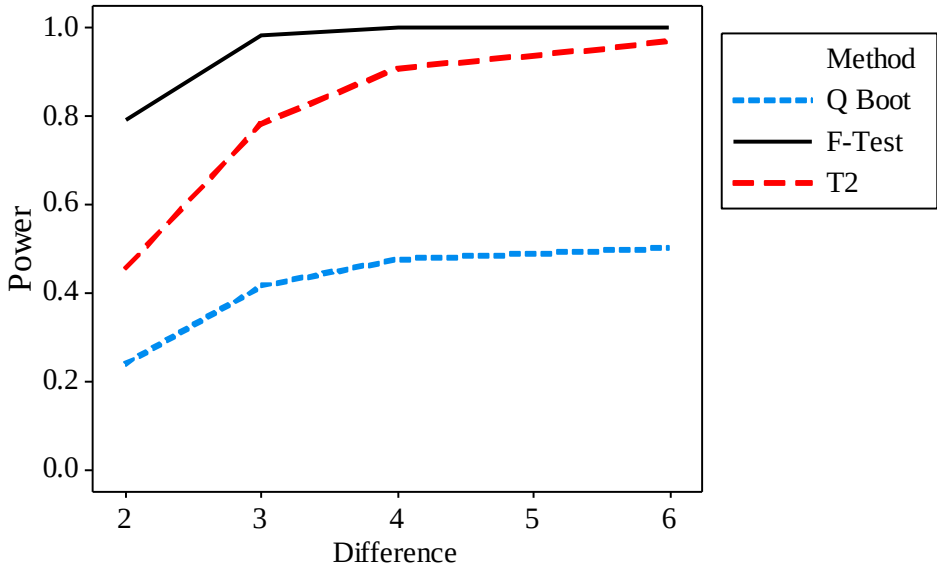

Figure 4-24. Power for r = 3, all groups Normal; Heteroscedastic, Unbalanced, Small/Large

Simulation Results for r = 5

Table 4-9. Details of r = 5 simulation cases, all groups drawn from Normal distribution

	Homoscedastic, Balanced		Heteroscedastic, Balanced		Heteroscedastic, Unbalanced, Large/Large		Heteroscedastic, Unbalanced, Small/Large	
i	n_i	σ_i^2	n_i	σ_i^2	n_i	σ_i^2	n_i	σ_i^2
1	10	1	10	1	5	1	5	9
2	10	1	10	1	10	1	10	9
3	10	1	10	4	10	4	10	4
4	10	1	10	9	15	9	15	1
5	10	1	10	9	20	9	20	1

Table 4-10. Level Error for r = 5, all groups drawn from Normal distribution

Description	Naïve	Q Boot	F-Test	T_2
Homoscedastic, Balanced	10.15%	4.35%	4.90%	1.75%
Heteroscedastic, Balanced	7.50%	3.70%	6.05%	1.40%
Heteroscedastic, Unbalanced, Large/Large	8.20%	5.30%	3.05%	0.90%
Heteroscedastic, Unbalanced, Small/Large	10.40%	4.90%	18.25%	3.35%

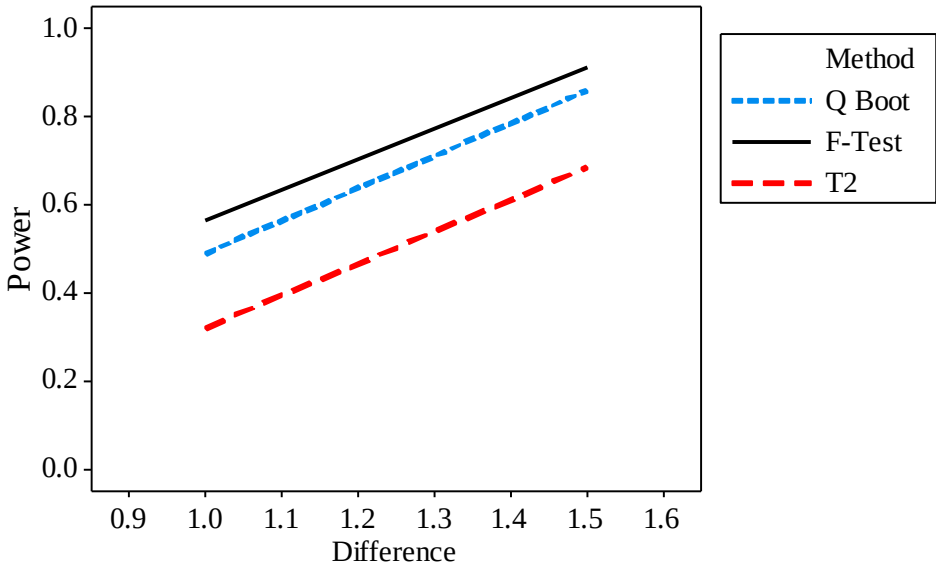

Figure 4-25. Power for r = 5, all groups Normal; Homoscedastic, Balanced

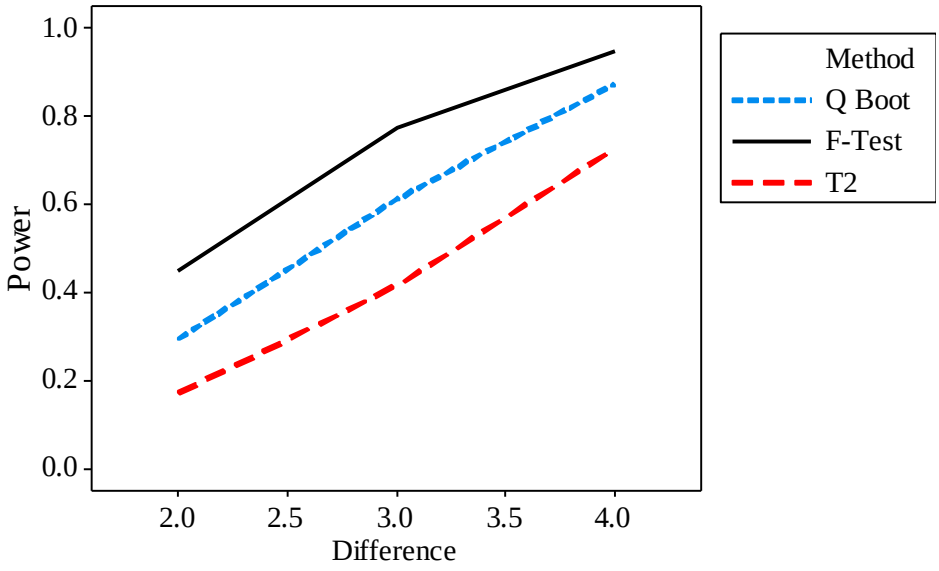

Figure 4-26. Power for r = 5, all groups Normal; Heteroscedastic, Balanced

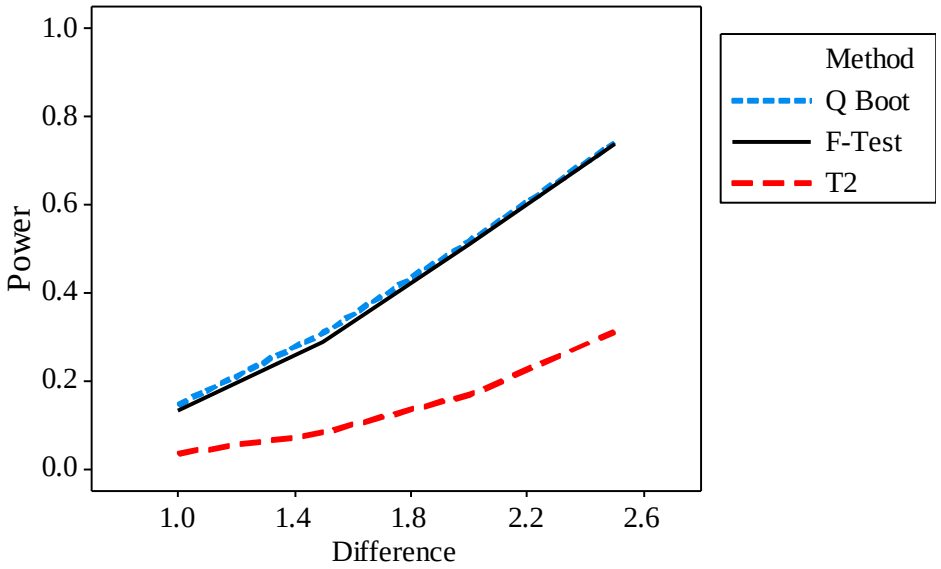

Figure 4-27. Power for r = 5, all groups Normal; Heteroscedastic, Unbalanced Large/Large

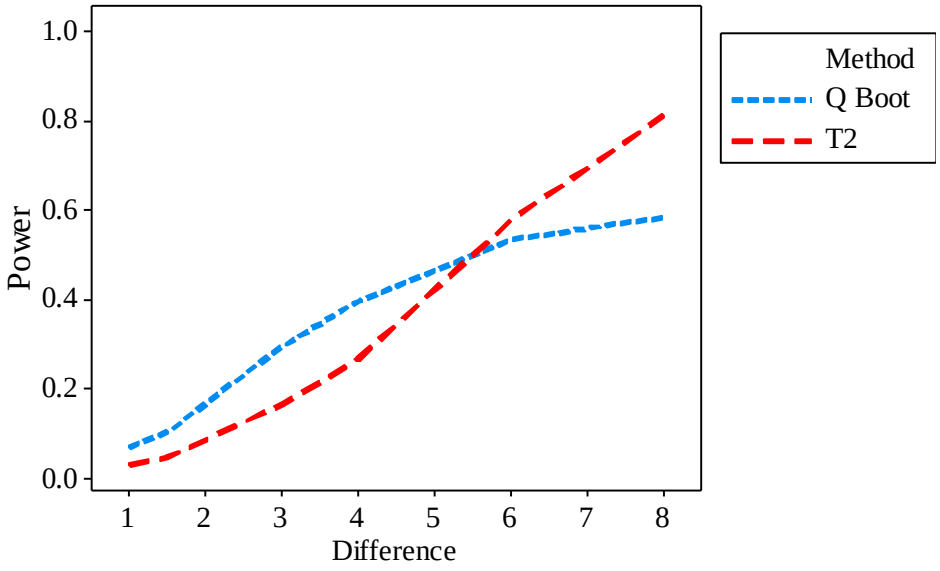

Figure 4-28. Power for r = 5, all groups Normal; Heteroscedastic, Unbalanced, Small/Large

Simulation Results for r = 10

Table 4-11. Details of r = 10 simulation cases, all groups drawn from Normal distribution

	Homoscedastic, Balanced		Heteroscedastic, Balanced		Heteroscedastic, Unbalanced, Large/Large		Heteroscedastic, Unbalanced, Small/Large	
i	n_i	σ_i^2	n_i	σ_i^2	n_i	σ_i^2	n_i	σ_i^2
1	10	1	10	1	5	1	5	25
2	10	1	10	1	5	1	5	25
3	10	1	10	4	10	4	10	16
4	10	1	10	4	10	4	10	16
5	10	1	10	9	15	9	15	9
6	10	1	10	9	15	9	15	9
7	10	1	10	16	20	16	20	4
8	10	1	10	16	20	16	20	4
9	10	1	10	25	25	25	25	1
10	10	1	10	25	25	25	25	1

Table 4-12. Level Error for r = 10

Description	Naïve	Q Boot	F-Test	T_2
Homoscedastic, Balanced	33.50%	3.55%	6.05%	1.90%
Heteroscedastic, Balanced	18.25%	4.05%	8.55%	2.10%
Heteroscedastic, Unbalanced, Large/Large	24.95%	3.90%	1.75%	0.30%
Heteroscedastic, Unbalanced, Small/Large	17.15%	1.50%	32.90%	1.15%

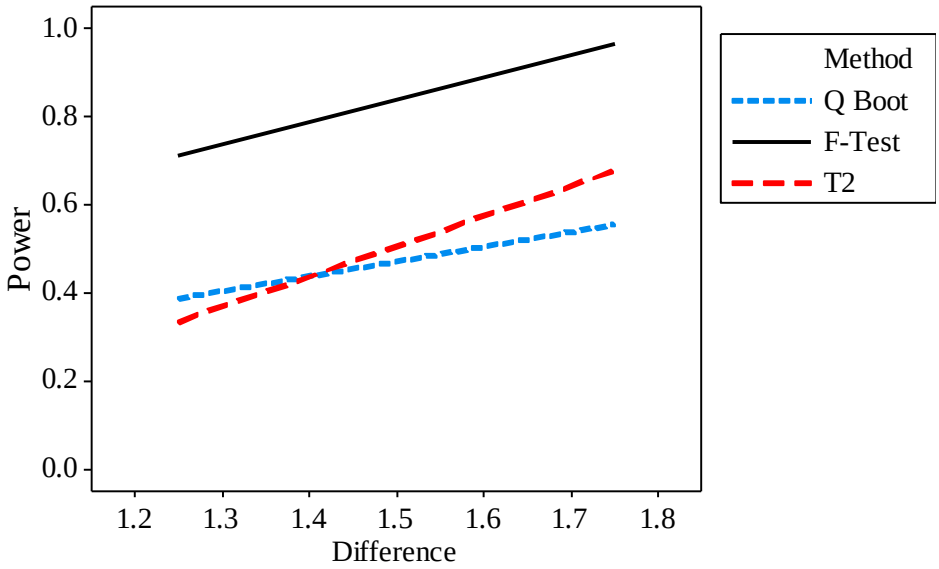

Figure 4-29. Power for r = 10, all groups Normal; Homoscedastic, Balanced

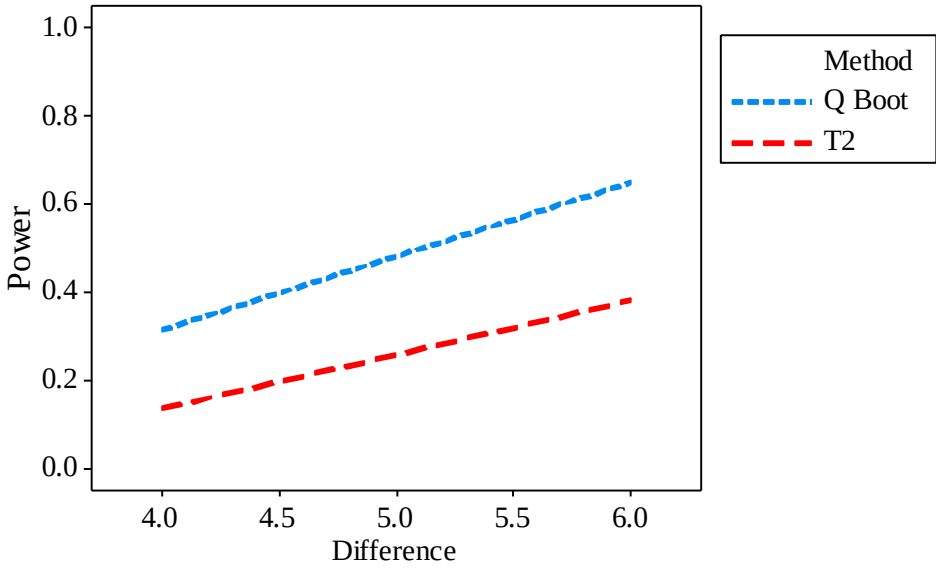

Figure 4-30. Power for r = 10, all groups Normal; Heteroscedastic, Balanced

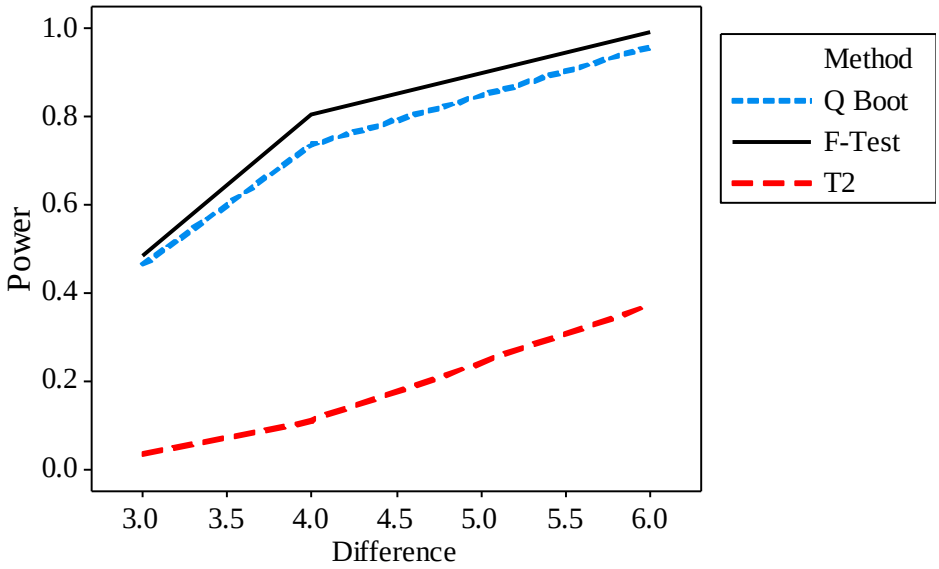

Figure 4-31. Power for r = 10, all groups Normal; Heteroscedastic, Unbalanced, Large/Large

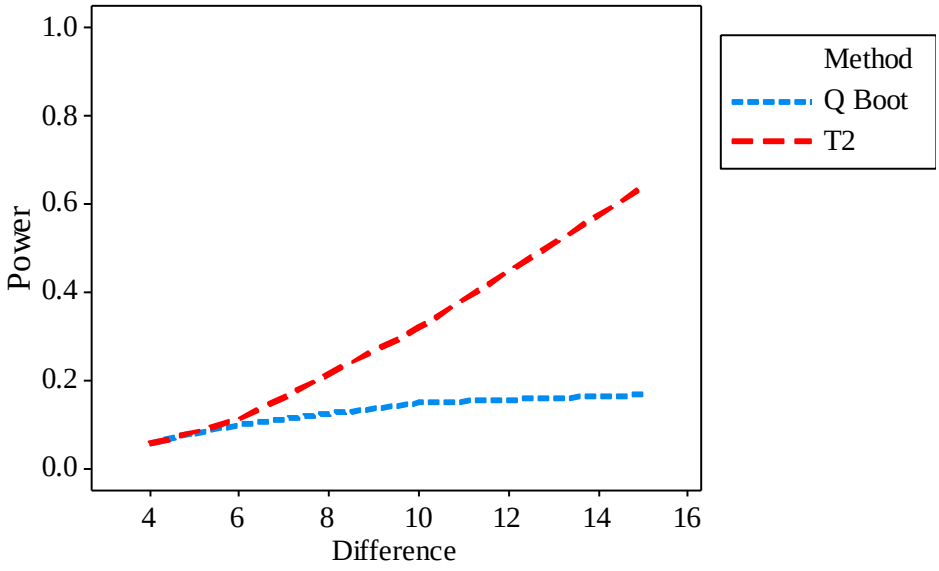

Figure 4-32. Power for r = 10, all groups Normal; Heteroscedastic, Unbalanced, Small/Large

Chapter 5 Conclusions

The quantile bootstrap method detailed here provides a means of testing the equality of group means (or of testing the equality of a group mean to a hypothesized value in the 1-sample case) under the alternative hypothesis, which accomplishes improved power while retaining accurate level error in most of the data scenarios that were simulated in this study. The notable and serious exception is in the case where one or more of the groups is strongly skewed. In that situation, level error is not accurate and the test cannot be trusted. As well, the quantile bootstrap method did not demonstrate improved power over Fisher and Hall's T_2 method/statistic in the heteroscedastic unbalanced case where groups with larger variance are also smaller in size. It is yet unclear why these scenarios present problems to the quantile bootstrap method, and further investigations into these scenarios explicitly would be warranted. While beyond the scope of this study, it would be interesting to see how the quantile bootstrap method performs under other measures of central tendency, such as the median, which would likely not be as sensitive as the mean to outlying data values common in datasets drawn from strongly skewed distributions.

As well, it cannot go without saying that the resampling-within-resampling component of the quantile bootstrap method adds considerable computational expense to the traditional bootstrap methods it is based on. In this study, we compared the performance of Fisher and Hall's T_2 method/statistic. Fisher and Hall's method required B iterations, whereas the quantile bootstrap method required B x M iterations. Computing power today is such that the multiplicative computational expense of the quantile bootstrap is not likely a significant detriment to the method, but it is noteworthy.

Bibliography

Nicholas I. Fisher and Peter Hall. On bootstrap hypothesis testing. *Australian Journal of Statistics*, 32(2):177-190, 1990.

G. V. Glass, P. D. Peckman, and J. R. Sanders. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42(3):237-288, 1972.

D. W. Zimmerman. Inflation of type I error rates by unequal variances associated with parametric, nonparametric, and rank-transformation tests. *Psicológica*, 25(1):103-133, 2004.

Appendix A

R Code

```
#########################################################################
# function takes indat, a vector of data to test mean(indat) = 0
#
QBoot1 <- function(indat,M,B){
#
#########################################################################

       resamp <- function(z,N){
              zlist  <- rep(list(z),N)
              zsamp  <- lapply(zlist, sample, size=length(z), replace=TRUE)
              return(zsamp)
              }

## Stage 1 #########

       datsamp      <- resamp(indat,B)
       datbar       <- sapply(datsamp, mean)
       nvpval       <- sum(datbar < 0)/B

## Stage 2 #########

       datcenter    <- indat - mean(indat)
       datcsamp     <- resamp(datcenter,M)
       datcstar     <- lapply(datcsamp, resamp, N=B)
       datcstarbar  <- lapply(datcstar, function(z) sapply(z,mean))
       qbar         <- sapply(datcstarbar, quantile, probs=nvpval)

       pval         <- sum(qbar > 0)/M

       return(pval)
       }


#########################################################################
# function takes indat, a 2-item list where each item is a group's data.
#      tests whether group 2's mean is greater than group 1's mean
#
QBoot2 <- function(indat,M,B){
#
#########################################################################

resamplist <- function(z,N){
       zll          <- rep(list(z),N)
       zsamp        <- lapply(zll, function(zl) lapply(zl,function(z)
                    sample(z,size=length(z),replace=TRUE)))
       return(zsamp)
       }

getbar <- function(zlist) return(mean(zlist[[2]]) - mean(zlist[[1]]))

Center <- function(zlist){
       return(lapply(zlist,function(z) z - mean(z)))
       }

## Stage 1 #########
```

```
       datsamp      <- resamplist(indat,B)  # list with B items, each a list of
                    resampled data

       nvMD         <- sapply(datsamp,getbar)
       nvMDpval     <- sum(nvMD < 0) / length(nvMD)


## Stage 2 #########

       datcenter    <- Center(indat)
       datcsamp     <- resamplist(datcenter,M)
       datcstar     <- lapply(datcsamp, resamplist, N=B)
       datcstarMD   <- lapply(datcstar, function(z) sapply(z,getbar))
       qMD          <- sapply(datcstarMD, quantile, probs=nvMDpval)
       MDpval       <- sum(qMD > 0) / length(qMD)

       return(MDpval)
       }

#########################################################################
# function takes indat, a list where each item is a group's data
#
QBoot <- function(indat,M,B){
#
#########################################################################

       ssq <- function(y){
              sum((y-mean(y))^2)
              }

       T3stat <- function(zlist){

              ni     <- sapply(zlist,length)
              r      <- length(zlist)
              mean.i       <- sapply(zlist,mean)
              mean.all<- mean(mean.i)

              T3num  <- sum((mean.i - mean.all)^2)
              T3den  <- ((r-1)/r)*sum((1/((ni-1)*ni))*sapply(zlist,ssq))

              return(T3num - T3den)
              }

       resamplist <- function(z,N){
              zll    <- rep(list(z),N)
              zsamp  <- lapply(zll, function(zl) lapply(zl,function(z)
              sample(z,size=length(z),replace=TRUE)))
              return(zsamp)
              }

       Center <- function(zlist){
              return(lapply(zlist,function(z) z - mean(z)))
              }

## Stage 1 #########

       datsamp <- resamplist(indat,B)

       nvT3   <- sapply(datsamp,T3stat)
       nvpval <- sum(nvT3 < 0) / length(nvT3)
```

```
## Stage 2 #########

       datcenter     <- Center(indat)
       datcsamp      <- resamplist(datcenter,M)
       datcstar      <- lapply(datcsamp, resamplist, N=B)
       datcstarT3    <- lapply(datcstar, function(z) sapply(z,T3stat))
       qT3           <- sapply(datcstarT3, quantile, probs=nvpval)
       T3pval        <- sum(qT3 > 0)/M

       return(T3pval)
       }
```