A Comparison of Permutation and Bootstrap Tests

He Jiang, supervised by Prof. Ery Arias-Castro

May 30, 2017

Abstract

This paper compares the permutation and bootstrap tests in the setting of a two sample testing of the equality of sample *mean*. Since Romano[4], Bradley and Tibshirani[2] already showed that the tests yield similar results in large sample hypothesis testings, but few, if any, literature investigated in the small sample cases, we will compare the permutation and bootstrap tests in the situation of a testing of the two sample *mean* in relatively small samples. For this paper, these samples will have equal size and variance. Specifically, we will investigate the relative value of the *Type I Error* rates and *Type II Error* rates of the permutation and bootstrap tests, and will use simulation results to investigate the required sample size for the tests to show similar results. We will also provide simulation results to visually understand the results proven by Romano, Bradley and Tibshirani of the large sample testing.

1 Introduction

1.1 Introduction to Paper

Permutation tests and bootstrap tests are two major ways of computer based non-parametric statistical tests that could be used under much less restrictive conditions than traditional parametric tests. In this paper, we will use mainly simulations to compare six tests under two different conditions. These six tests include the permutation test, the concatenated bootstrap test, the separated bootstrap test, and their respective studentized versions. The two situations are Normal samples and Exponential samples of various sample sizes. We will compare them using both the standards of *Type I Error* and *Power*, in the context of comparing the *mean* of two samples. We hope this paper will serve as a supplement of Romano's paper in the focus on small samples, and hope it will provide reference for selecting between the various available permutation and bootstrap tests, and provide further understanding of the permutation and bootstrap tests.

1.2 Permutation Test

Permutation Tests are usually used to test whether the two samples came from the same distribution. They rely on the assumption that the two samples came from the identical distribution, so that when the data are permuted, the new samples are still assumed to have the same distribution under the null assumption. Consider the case where we have two samples with $x_1, x_2, ..., x_m \sim$ i.i.d. F_x and $y_1, y_2, ..., y_n \sim$ i.i.d. F_y, and we want to test whether F_x is distributed the same as F_y. In this case, we could implement a permutation test, with test statistics of either:

$$T_1 = \bar{x} - \bar{y} \ \ or \ \ T_2 = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{{S_x}^2}{m} + \frac{{S_y}^2}{n}}} \tag{1}$$

We will call the permutation method using test statistic T_1 the regular permutation and the method using test statistic T_2 the studentized permutation.

To carry out the permutation methods, first use equation (1) to compute the test statistic T_a^0 from the observed samples [1], where a=1,2. Then concatenate the two samples into one sample $z_1, z_2, ...z_(m+n)$. Next use Monte

Carlo sampling to randomly sample B permutations from all permutations of $\{1, ..., m+n\}$. For each permutation b of the B permutations, let

$$X_i^b = Z_b(i) \;\; where \;\; i = 1, ..., m \tag{2}$$

$$Y_j^b = Z_b(j+m) \;\; where \;\; j = 1, ..., n \tag{3}$$

and compute T_a^b based on equation (1).

The p-value is then given by

$$p - value = \frac{\#\{b : T_a^b \geq T_a^0\} + 1}{B+1} \tag{4}$$

1.3 Concatenated Sampling Bootstrap Test

The bootstrap test has the advantage of not requiring a special symmetry that is needed for a permutation test, which means that it could be carried out more generally [2]. Similar to section 1.2, we will call the test using test statistic T_1 the regular concatenated sampling bootstrap and the test using test statistic T_2 the studentized concatenated sampling bootstrap.

To carry out the concatenated sampling bootstrap test, we also require that the two samples come from the same distributions, since the sampling will be similar to the permutation test, just with replacement. First we compute the test statistic T_a^0 from the observed samples, where a=1,2, and the test statistics are calculated from equation (1). Next we concatenate the two samples into one sample $z_1, z_2, ...z_(m+n)$, and let this sample's distribution be F_z. Then we randomly sample B samples with replacement from F_z. For each b of the B samples, let

$$X_i^b = Z_b(i) \;\; where \;\; i = 1, ..., m \tag{5}$$

$$Y_j^b = Z_b(j+m) \;\; where \;\; j = 1, ..., n \tag{6}$$

and compute T_a^b based on equation (1).

The p-value is then given by equation (4) as in section 1.2.

1.4 Seperated Sampling Bootstrap Test

The seperated sampling boostrap test is usually carried out when we want to test a hypothesis regarding the two sample means. It is preferred in this situation of testing the equality of two sample means because it does not interchange samples and thus each generated sample would be representative only of the observed sample that it came from. Thus using the seperated sampling bootstrap test, we can avoid interchanging two samples which came from different distributions. Again here we will call the method using test statistic T_1 as the regular seperated sampling bootstrap, and the method using test statistic T_2 as the studentized seperated sampling bootstrap.

In order to carry the seperated bootstrap test out, first we would need to compute the test statistic T_a^0 from the observed samples based on equation (1), where a=1,2, and then we would subtract $\bar{x}$ from every x and $\bar{y}$ from every y. After removing the sample mean from both samples, we would acquire two new samples, and two distribution functions $\hat{F}_x$ and $\hat{F}_y$ based on the new sample. Then for every $b = 1, 2, ..., B$, we generate $x_1^b, x_2^b, ..., x_m^b \sim$ i.i.d. $\hat{F}_x$ and $y_1^b, y_2^b, ..., y_n^b \sim$ i.i.d. $\hat{F}_y$. Finally, for each b, we will compute the test statistic according to equation (1) in section 1.2, and the p-value would again be given by equation (4) in section 1.2.

1.5 Description of Paper

Although the above mentioned tests would function well with different conditions, in many situations, however, it is very difficult to know or investigate what conditions the data actually satisfies prior to implementing the tests. Thus in this paper, we will compare the six tests together in the situation of a one-sided testing of the equality of the two sample *means*. We will be carrying out simulations in R using the methods described in sections 1.2 through 1.4 and in the context of specific cases as described in the following sections. Chapters 2-5 of this paper will be providing details about comparing the six tests when the two samples are of equal size and has same variance, in the context of both a Normal and a non-Normal situation. At the end, Chapter 6 will offer explanations of the results as well as conclusions.

2 Comparing Type I Error in Testing the Equality of Means of Two Normal Samples

2.1 Testing $H_0 : \mu_x = \mu_y \ \ vs \ \ H_1 : \mu_x > \mu_y$ with two Normal Distributions

Consider the case where we have two samples, $x_1, x_2, ..., x_m \sim$ i.i.d. $Normal(\mu_x, {\sigma_x}^2)$ and $y_1, y_2, ..., y_n \sim$ i.i.d. $Normal(\mu_y, {\sigma_y}^2)$. This case could arise in many situations, for instance if we want to compare the mean of a group of students' test scores to that of another group, or if we want to compare the mean of the height of a certain group of people to that of another group. In any case, we are interested in testing the following hypothesis:

$$H_0 : \mu(x) = \mu(y) \ \ vs \ \ H_1 : \mu(x) > \mu(y) \tag{7}$$

The first criteria of our comparison is to see how much *Type I Error* each test makes. *Type I Error* is the probability that we falsely rejected our null hypothesis when it is actually true. Thus a lower *Type I Error* would be desirable.

In order to investigate the *Type I Error* of each of the tests, we would generate two samples $x_1, x_2, ..., x_m \sim$i.i.d. Normal$(\mu_x, {\sigma_x}^2)$, and $y_1, y_2, ..., y_n \sim$i.i.d. Normal$(\mu_y, {\sigma_y}^2)$, with $\mu_x = \mu_y$ so that the null hypothesis is true. We will record the density plots of the p-values of the various tests, and then record the actual *Type I Error* rates with the various tests at the level of 5%.

Furthermore, since both samples came from the Normal Distribution, we could implement the z-test, which would serve as a great reference for our six tests.

In the circumstance described above where $\mu_x = \mu_y$, and the variance of the two samples are known to be ${\sigma_x}^2 \ \ and \ \ {\sigma_y}^2$, $\bar{x} - \bar{y} \sim Normal(0, \frac{{\sigma_x}^2}{m} + \frac{{\sigma_y}^2}{n})$, so the test statistic $z = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{{\sigma_x}^2}{m} + \frac{{\sigma_y}^2}{n}}} \sim Normal(0, 1)$.

From the above result, we know that a z-test with the Z statistic provided above could be implemented in this case with p-value equals to the area to the right of the Z value under the Standard Normal density curve. And specifically, in the case of a 5% test, we would reject H_0 if z is ≥ 1.645, i.e. the 95-th percentile of the Standard Normal Distribution.

Thus, we will use the six tests, as we discussed in Chapter 1, and the z-test, to test the above hypothesis.

2.2 Normal Sample Type I Error

In this section we will consider the case where ${\sigma_x}^2 = {\sigma_y}^2 = 1$ and $B = 10,000$. We will vary the sample sizes from $m = n = 10$, to $m = n = 50$, to $m = n = 100$, and then to $m = n = 1000$. The whole process is again repeated J=5,000 times. Note that B and J are large because we want more reliable simulation results, as few very unusual samples will not effect the result as much. We will show the p-value density plots of the four cases in Figure 1, and provide the specific detailed values of the *Type I Error* rates in Tables 1-4.

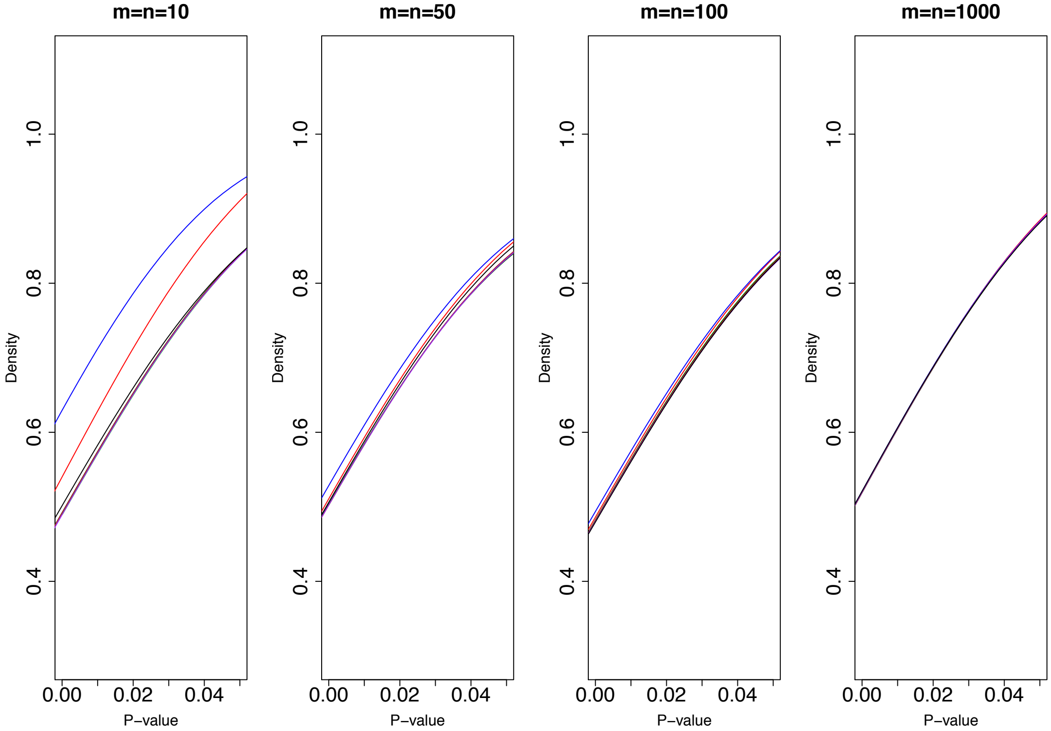

Figure 1: Normal P-Value Density from p=0 to p=0.05 (from left to right respectively: $m = n = 10, m = n = 50, m = n = 100, m = n = 1000$), with permutation as green, studentized permutation as darkgreen, concatenated bootstrap as red, studentized concatenated bootstrap as brown, seperated boostrap as blue, studentized seperated bootstrap as purple, and ztest as black

Table 1: Type I Errors of Normal Samples, with m=n=10 and ${\sigma_x}^2 = {\sigma_y}^2 = 1$

Test	**Total Trials**	**Number of Rejections**	**Type I Error**	**Margin of Error**
Perm.reg	5000	241	0.0482	0.005936896
Perm.stud	5000	242	0.0484	0.005948576
Boot.con.reg	5000	265	0.0530	0.006209777
Boot.con.stud	5000	241	0.0482	0.005936896
Boot.sep.reg	5000	339	0.0678	0.006968391
Boot.sep.stud	5000	234	0.0468	0.005854341
Z.test	5000	249	0.0498	0.006029555

Table 2: Type I Errors of Normal Samples, with m=n=50 and ${\sigma_x}^2 = {\sigma_y}^2 = 1$

Test	**Total Trials**	**Number of Rejections**	**Type I Error**	**Margin of Error**
Perm.reg	5000	261	0.0522	0.006165335
Perm.stud	5000	265	0.053	0.006209777
Boot.con.reg	5000	264	0.0528	0.006198704
Boot.con.stud	5000	260	0.052	0.006154162
Boot.sep.reg	5000	275	0.055	0.006319174
Boot.sep.stud	5000	257	0.0514	0.00612049
Z.test	5000	252	0.0504	0.006063854

Table 3: Type I Errors of Normal Samples, with m=n=100 and ${\sigma_x}^2 = {\sigma_y}^2 = 1$

Test	**Total Trials**	**Number of Rejections**	**Type I Error**	**Margin of Error**
Perm.reg	5000	239	0.0478	0.005913453
Perm.stud	5000	240	0.048	0.005925189
Boot.con.reg	5000	245	0.049	0.005983446
Boot.con.stud	5000	239	0.0478	0.005913453
Boot.sep.reg	5000	246	0.0492	0.005995014
Boot.sep.stud	5000	241	0.0482	0.005936896
Z.test	5000	234	0.0468	0.005854341

Table 4: Type I Errors of Normal Samples, with m=n=1000 and ${\sigma_x}^2 = {\sigma_y}^2 = 1$

Test	**Total Trials**	**Number of Rejections**	**Type I Error**	**Margin of Error**
Perm.reg	5000	250	0.05	0.006041015
Perm.stud	5000	253	0.0506	0.006075233
Boot.con.reg	5000	246	0.0492	0.005995014
Boot.con.stud	5000	253	0.0506	0.006075233
Boot.sep.reg	5000	250	0.05	0.006041015
Boot.sep.stud	5000	249	0.0498	0.006029555
Z.test	5000	249	0.0498	0.006029555

As can be seen from Figure 1 and Tables 1-4, starting from $m = n = 100$, the test results start to become extremely similar. Thus the simulation results show that *100* is a good size to be considered large for the results of the test to be extremely similar, given that the samples are of equal variance of 1 and equal size. It could also be noted that when the sample sizes are small, for instance when $m = n = 10$, the three studentized tests have very similar Type I Error Rates, which roughly match the curve of the z-test. For the non-studentized versions, the seperated bootstrap test has the highest *Type I Error*, followed by the concatenated bootstrap test, and then the permutation test, although the difference of the later two is not as significant.

It should come at no surprise that at the large sample case where $m = n = 1000$, that all the *Type I Error* rates are extremely similar and stayed around 5%, as we are carrying out our tests at the 5% level, and that Romano pointed out that the tests will have similar results in large samples [4]. Figures 2 and 3 will further illustrate this idea.

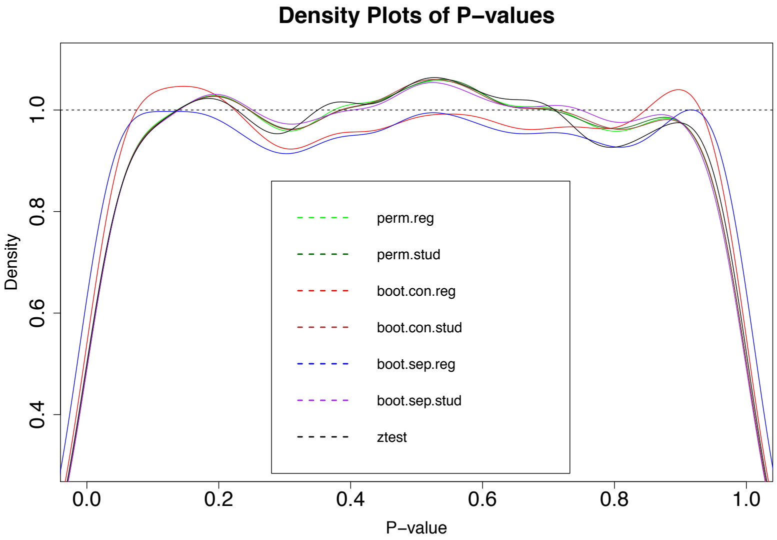

Figure 2: Normal Sample P-Value Density Plot with $m = n = 10, {\sigma_x}^2 = {\sigma_y}^2 = 1$

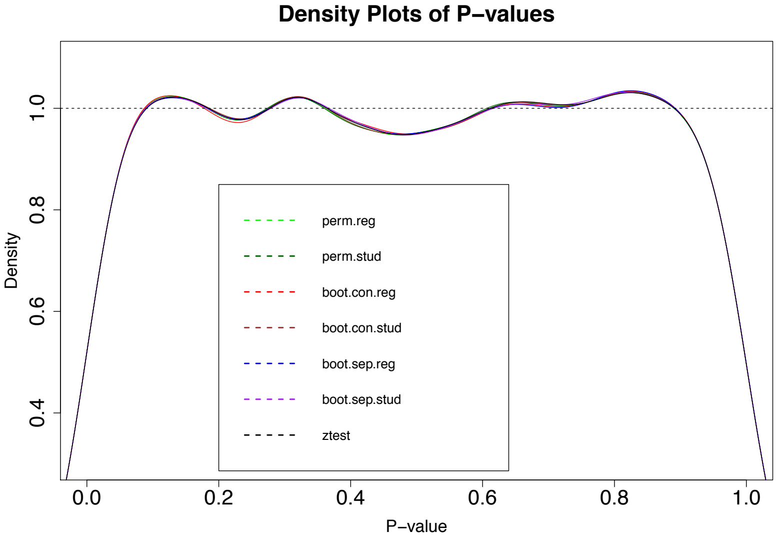

Figure 3: Normal Sample P-Value Density Plot with $m = n = 1000, {\sigma_x}^2 = {\sigma_y}^2 = 1$

3 Comparing Power in Testing the Equality of Means of Two Normal Samples

3.1 Using two Normal Distributions to test $H_0 : \mu_x = \mu_y \ vs \ H_1 : \mu_x > \mu_y$ and compare Power by increasing μ_x

In this chapter we will consider how to select from the six tests based on their *Power*, i.e. their ability to reject the Null Hypothesis when it is actually false.

Let $x_1, x_2, ..., x_m \sim$ i.i.d. $Normal(\mu_x, {\sigma_x}^2)$, and let $y_1, y_2, ..., y_n \sim$ i.i.d. $Normal(\mu_y, {\sigma_y}^2)$. To investigate the *Power*, i.e. to create a situtation where the null hypothesis is false, we need to generate two samples with different means. To achieve this goal, and w.l.o.g., assume that $\mu_y = 0$, and ${\sigma_x}^2 = {\sigma_y}^2 = 1$. Then for each $\mu_x = \mu$ from 0 to 2.5, with increments of 0.1 in the $m = n = 10$ case (or each μ from 0 to 0.25, with increments of 0.0125 in the $m = n = 1000$ case), we will carry out all six tests to test our hypothesis. We set B=10,000, and for each μ_x we will do each test 1,000 times. Note that here the Z statistic need to be adjusted accordingly to $z = \frac{(\bar{x}-\bar{y})-\mu}{\sqrt{\frac{{\sigma_x}^2}{m}+\frac{{\sigma_y}^2}{n}}}$ as the difference of $\mu_x - \mu_y = \mu$. For each μ we will record four data points, including the 90-th, 50-th and 10-th percentiles of the p-values of our test, as well as a power of each test in the context of a one-sided testing at the level of 5%.

Our goal is to observe both the *p-value plots* and the *power plots* to see if we can find any trends in them through our simulations.

3.2 P-value plots and Power plots with $m = n = 10$

As can be seen from Figure 5, in the case of small samples, we have very similar power curves at the very right(i.e. when μ is larger than 2.0 in this case). However, in the middle part of the graph, the curves differ from each other. The line representing the *Power* of the seperated sampling bootstrap(the blue line) runs on the top. The other lines run under it in a cluster in the middle of the graph. This result could show that the seperated bootstrap test has very high *Power* compared to the other tests in small sample testing.

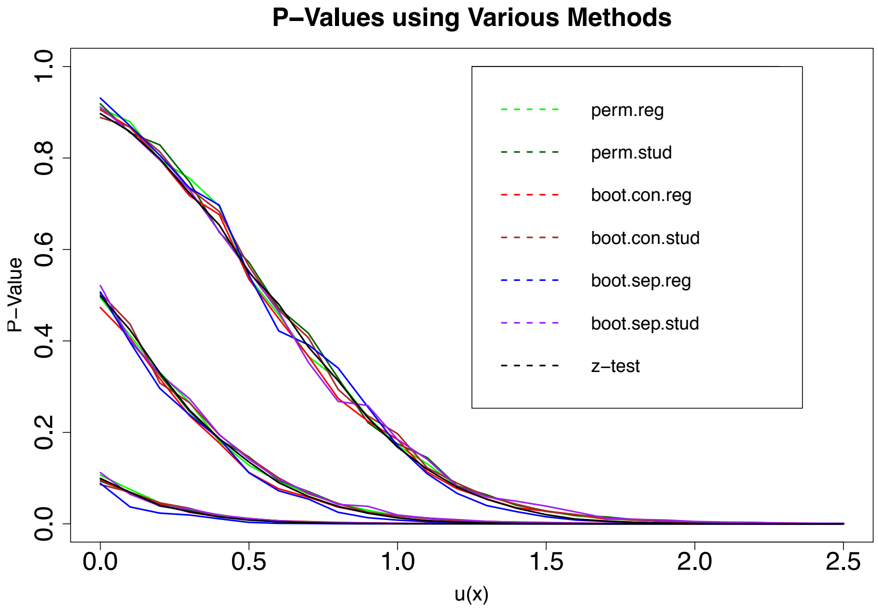

Figure 4: Normal Sample P-value Bands, with m=n=10 and ${\sigma_x}^2 = {\sigma_y}^2 = 1$

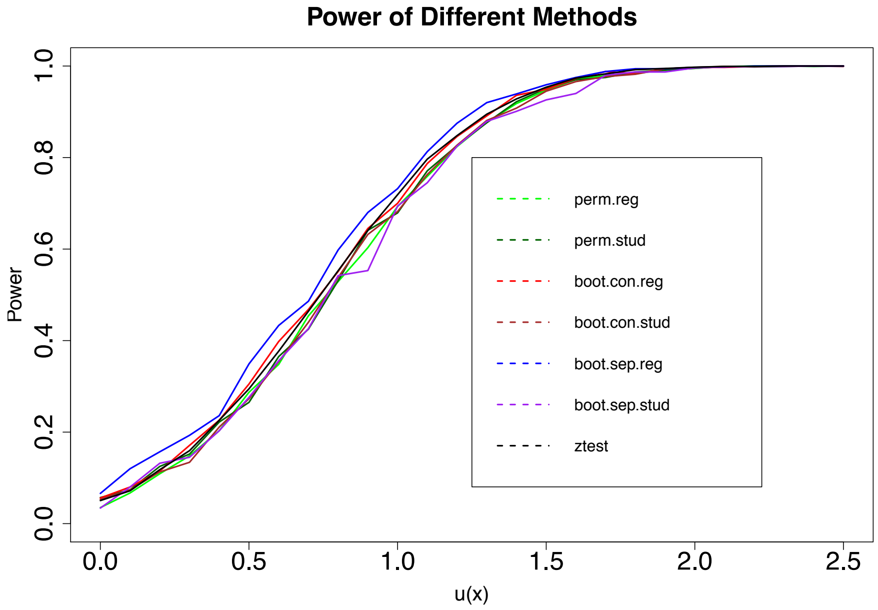

Figure 5: Normal Sample Power Plot, with m=n=10 and ${\sigma_x}^2 = {\sigma_y}^2 = 1$

3.3 P-value plots and Power plots with $m = n = 1000$

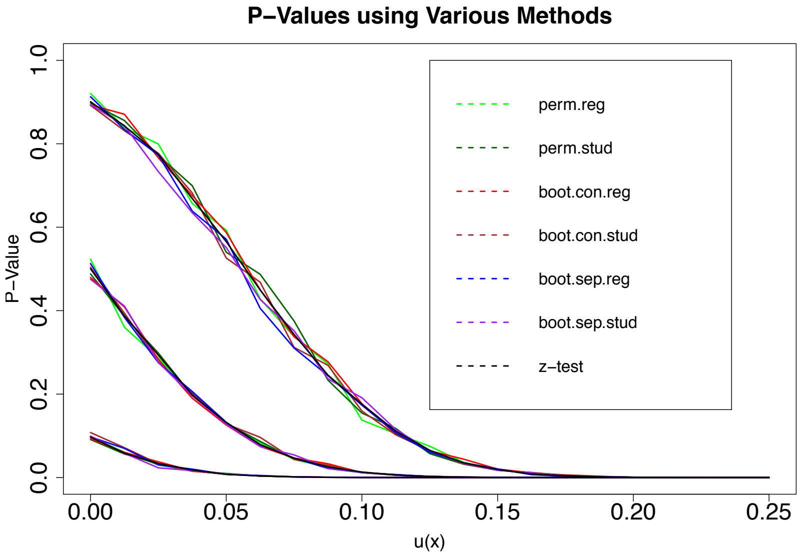

Figure 6: Normal Sample P-value Bands, with m=n=1000 and ${\sigma_x}^2 = {\sigma_y}^2 = 1$

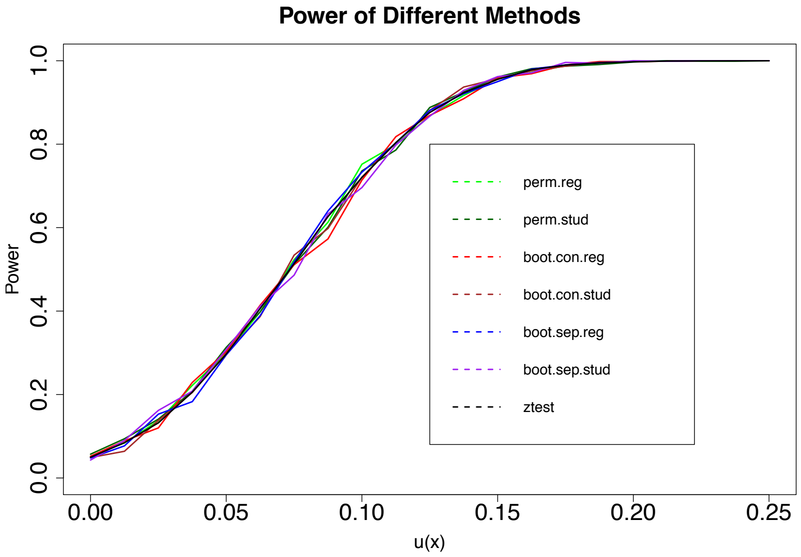

Figure 7: Normal Sample Power Plots, with m=n=1000 and ${\sigma_x}^2 = {\sigma_y}^2 = 1$

When the sample size is large, as can be seen from Figures 6 and 7, the six tests, together with the z-test, all yield similar p-value plots and power curves. Thus Figures 6 and 7 clearly showed the result proven by Romano. Thus we conclude that all the choices are similar in the aspect of *Power* when sample sizes are large.

4 Comparing Type I Error in Testing the Equality of Means of Two Exponential Samples

4.1 Using two Exponential Distributions to test $H_0 : \mu_x = \mu_y \ \ vs \ \ H_1 : \mu_x > \mu_y$

Not only can the permutation and bootstrap tests be used on Normal Samples, they could also be used on samples that are skewed and have long tails. We will be considering the Exponential distribution here. We will again use the six tests which we discussed in Chapter 1 to test the above hypothesis. First, we will generate two samples, $x_1, x_2, ..., x_m \sim$i.i.d. Exponential$(\lambda) + \mu$, i.e. an exponential distribution with parameter λ that is shifted to the right by μ, and $y_1, y_2, ..., y_n \sim$i.i.d. Exponential(λ). Then we will use the six test statistics to test the goal hypothesis. This case could arise in many situations, for instance if we want to compare the time it takes to cure a certain disease for a group of patients who used a certian medicine, to the time of another group of patients who have not used that medicine.

In order to test the Type I Error, we will set $\mu = 0$, so that we could gurantee that H_0 is true.

4.2 Type I Error

In this section we will consider the case where $\lambda = 1$ and $B = 10,000$. We will vary the sample sizes from $m = n = 10$, to $m = n = 100$, and then to $m = n = 1000$ ($m = n = 50$ is skipped because we already found out that 50 is not quite large enough for the tests to show very similar results in section 2.2). The whole process is again repeated J=5,000 times. Note that B and J are large because we want more reliable simulation results, as few very unusual samples will not effect the result as much. We will show the p-value density plots of the four cases in Figure 8, and provide the specific detailed values of the *Type I Error* rates in Tables 5-7.

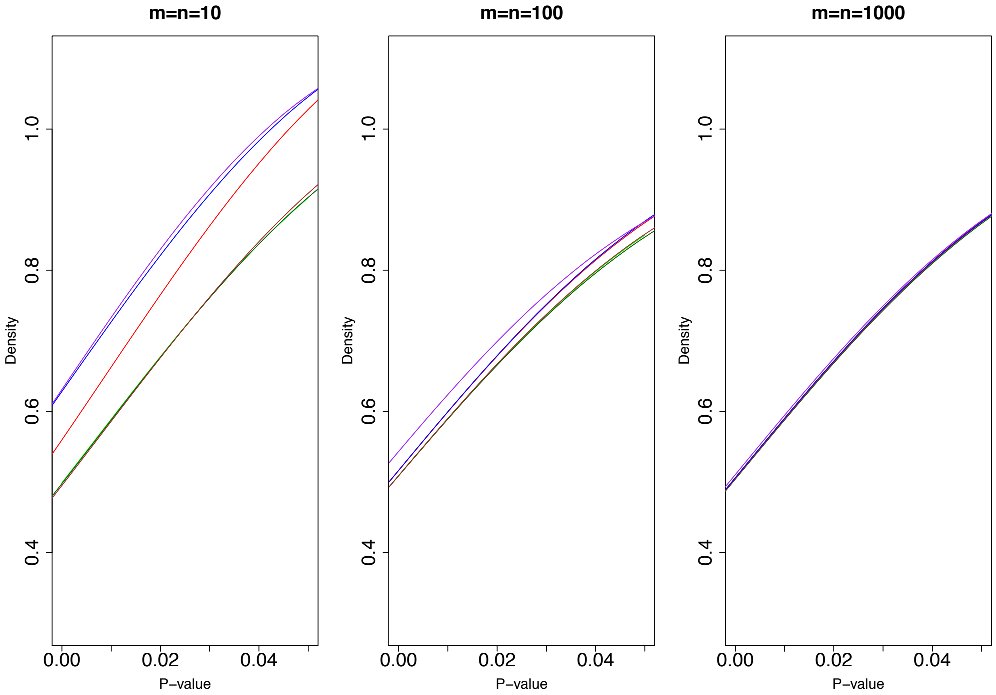

Figure 8: Exponential P-Value Density from p=0 to p=0.05 (from left to right respectively, $m = n = 10, m = n = 100, m = n = 1000$, with permutation as green, studentized permutation as darkgreen, concatenated bootstrap as red, studentized concatenated bootstrap as brown, seperated boostrap as blue, and studentized seperated bootstrap as purple)

Table 5: Type I Errors of Exponential Samples, with m=n=10 and ${\sigma_x}^2 = {\sigma_y}^2 = 1$

Test	**Total Trials**	**Number of Rejections**	**Type I Error**	**Margin of Error**
Perm.reg	5000	239	0.0478	0.005913453
Perm.stud	5000	233	0.0466	0.005842431
Boot.con.reg	5000	260	0.0520	0.006154162
Boot.con.stud	5000	232	0.0464	0.005830492
Boot.sep.reg	5000	314	0.0628	0.006724485
Boot.sep.stud	5000	327	0.0654	0.006852749

Table 6: Type I Errors of Exponential Samples, with m=n=100 and ${\sigma_x}^2 = {\sigma_y}^2 = 1$

Test	**Total Trials**	**Number of Rejections**	**Type I Error**	**Margin of Error**
Perm.reg	5000	257	0.0514	0.00612049
Perm.stud	5000	255	0.051	0.006097914
Boot.con.reg	5000	258	0.0516	0.00613174
Boot.con.stud	5000	254	0.0508	0.006086587
Boot.sep.reg	5000	259	0.0518	0.006142964
Boot.sep.stud	5000	277	0.0554	0.006340769

Table 7: Type I Errors of Exponential Samples, with m=n=1000 and ${\sigma_x}^2 = {\sigma_y}^2 = 1$

Test	**Total Trials**	**Number of Rejections**	**Type I Error**	**Margin of Error**
Perm.reg	5000	249	0.0498	0.006029555
Perm.stud	5000	243	0.0486	0.005960227
Boot.con.reg	5000	243	0.0486	0.005960227
Boot.con.stud	5000	241	0.0482	0.005936896
Boot.sep.reg	5000	247	0.0494	0.006006555
Boot.sep.stud	5000	251	0.0502	0.006052447

Here, although Tables 5-7 show that the numerical values of the *Type I Error* rates tend similar when $m = n = 100$, Figure 8 still showed some minor difference in the p-value density curves of the tests at this sample size. Therefore, we should conclude that the sample size for the tests to behave similar results in this case is around or a little above $m = n = 100$, given that the samples are of equal variance of 1 and equal size. It could also be noted from Figure 8 that the studentized and non-studentized versions of the bootstrap test have the highest *Type I Error* when sample sizes are small, for instance $m = n = 10$.

Similar to the Normal case, when sample sizes are large, all *Type I Error* rates are extremely similar and stayed around 5%, as we are carrying out our tests at the 5% level, and that Romano pointed out that the tests will have similar results in large samples [4]. Figures 9 and 10 will further illustrate this idea.

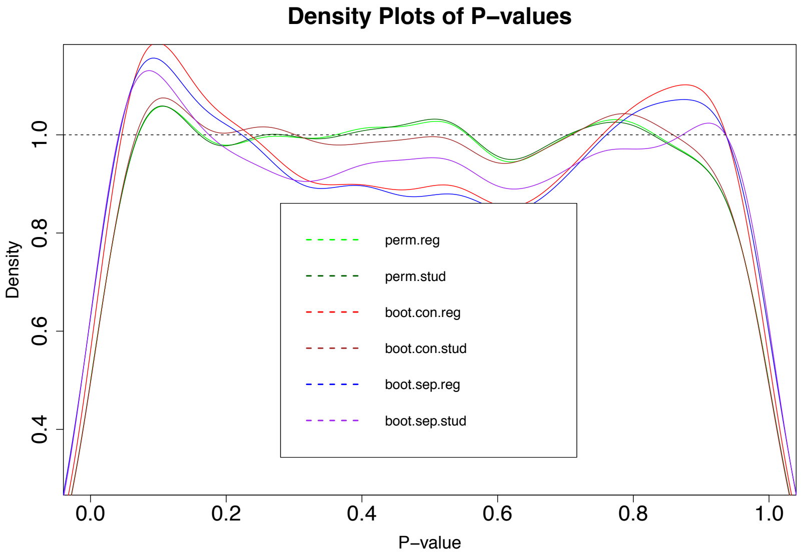

Figure 9: Exponential Sample P-Value Density Plot with $m = n = 10, {\sigma_x}^2 = {\sigma_y}^2 = 1$

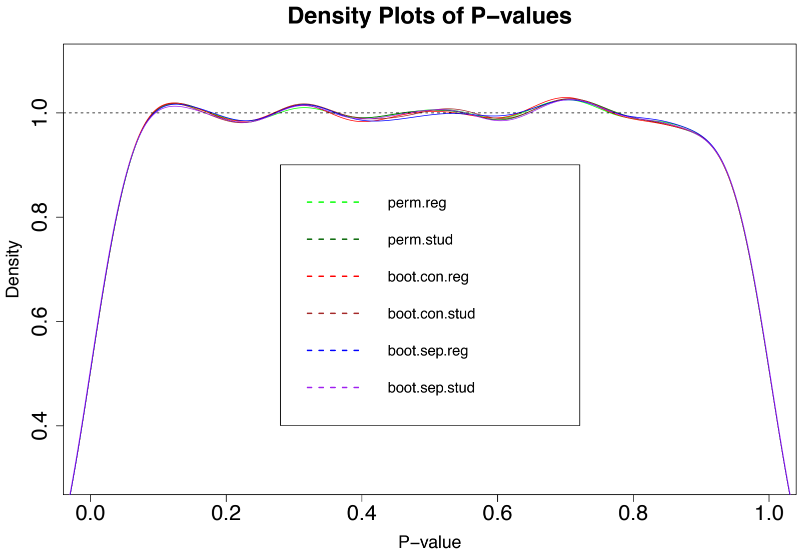

Figure 10: Exponential Sample P-Value Density Plot with $m = n = 1000, {\sigma_x}^2 = {\sigma_y}^2 = 1$

5 Comparing Power in Testing the Equality of Means of Two Exponential Samples

5.1 Using two Exponential Distributions to test $H_0 : \mu_x = \mu_y \ vs \ H_1 : \mu_x > \mu_y$ and compare power by increasing the shift μ

In this chapter we will consider how to select from the six tests based on their *Power*, i.e. their ability to reject the null hypothesis when it is actually false.

Let $x_1, x_2, ..., x_m$ $\sim$i.i.d. Exponential(λ) + μ, and $y_1, y_2, ..., y_n$ $\sim$i.i.d. Exponential(λ). Similar to Chapter 3, for each μ from 0 to 3, with increments of 0.1 when $m = n = 10$ (or each μ from 0 to 0.25, with increments of 0.0125 with $m = n = 1000$), we will carry out all six tests to test our hypothesis. We set B=10,000, and for each μ we will do each test 1,000 times. For each μ we will record four data points, including the 90-th, 50-th and 10-th percentiles of the p-values of our test, as well as a power of each test in the context of a one-sided testing at the level of 5%.

Our goal is to observe both the *p-value plots* and the *power plots* to see if we can find any trends in them through our simulations.

5.2 P-value plots and Power plots with $m = n = 10$

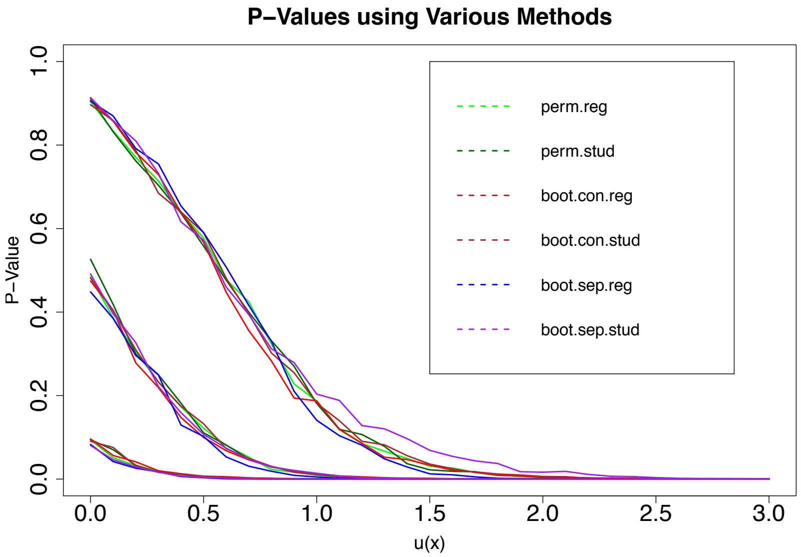

Figure 11: Exponential Sample P-value Bands with $m = n = 10$ with ${\sigma_x}^2 = {\sigma_y}^2 = 1$

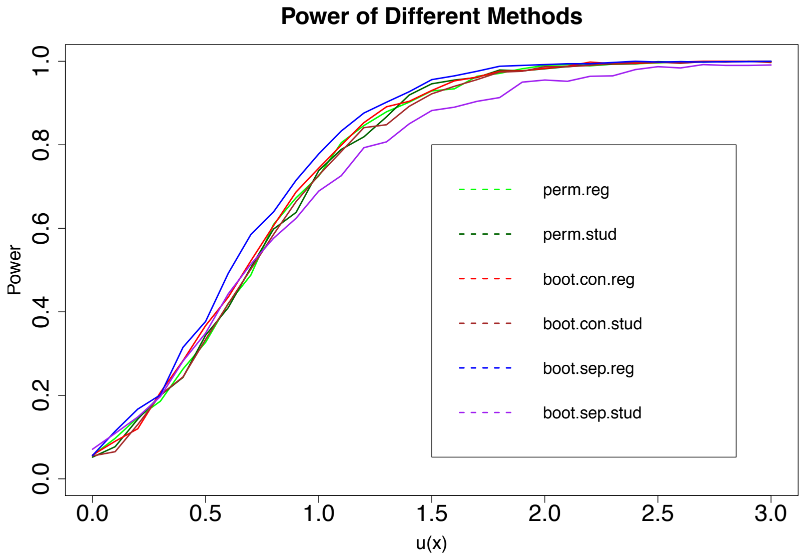

Figure 12: Exponential Sample Power Plots with $m = n = 10$ with ${\sigma_x}^2 = {\sigma_y}^2 = 1$

As can be seen from Figure 12, in the case of small samples, we have again very different curves in the middle. The line representing the power of the seperated sampling bootstrap(the blue line) runs on the top. Then we have another cluster of lines below the blue line, with the line representing the concatenated sampling bootstrap(the red line) relatively on the top of the cluster. The line representing the seperated studentized boostrap(the purple line) test runs on the bottom for μ from 1.0 to 2.5.

5.3 P-value plots and Power plots with $m = n = 1000$

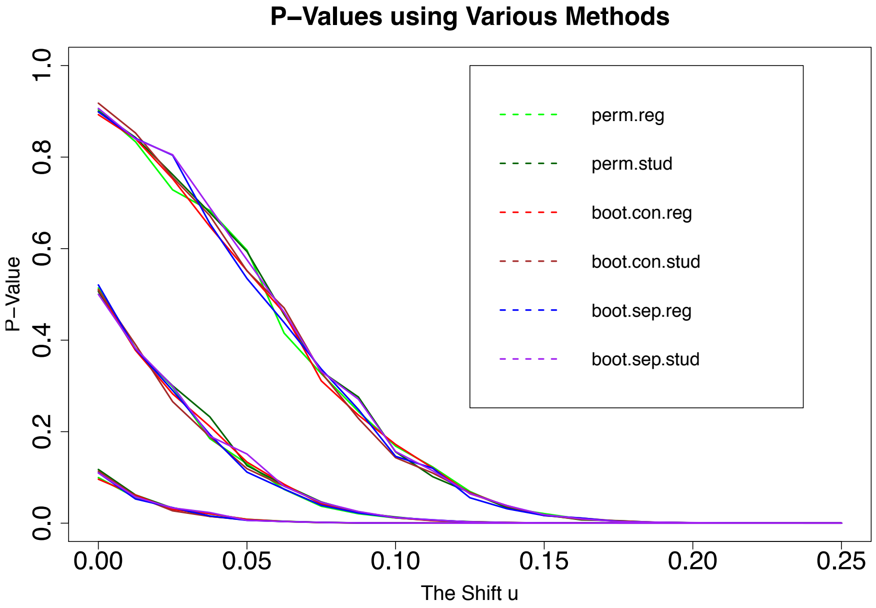

Figure 13: Exponential Sample P-value Bands with $m = n = 1000$ with ${\sigma_x}^2 = {\sigma_y}^2 = 1$

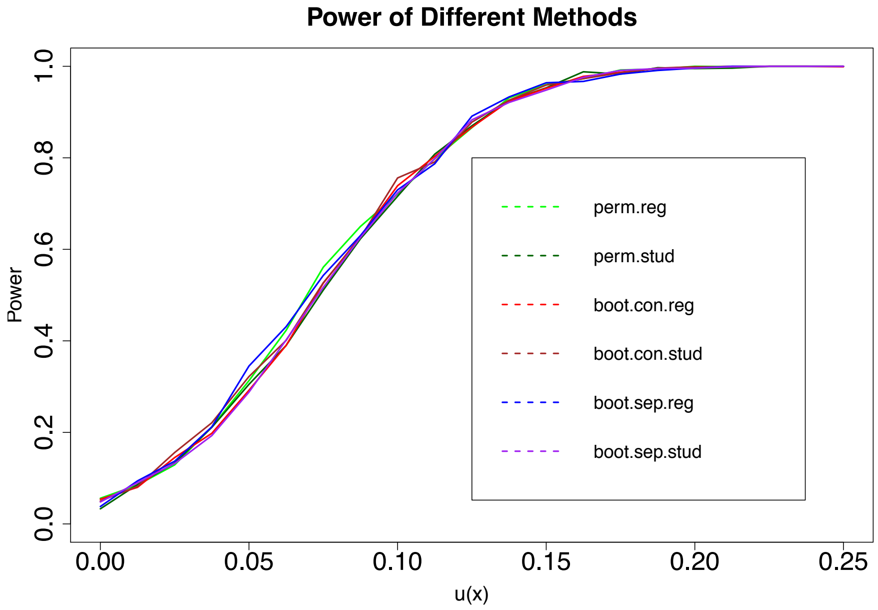

Figure 14: Exponential Sample Power Plots with $m = n = 1000$ with ${\sigma_x}^2 = {\sigma_y}^2 = 1$

Similar to Section 3.3, when the sample sizes are large, the six tests all yield similar p-value plots and power curves, as can be seen from Figure 14. Thus Figure 14 also illustrated the result proven by Romano. Therefore we conclude that all the choices are similar in the aspect of *Power* when sample sizes are large in this Exponential sample case.

6 Explanations and Conclusions

6.1 Explanations on the Observed Results

Consider Chapters 2 and 4, where we were interested in comparing the *Type I Error* of the six tests. As pointed out by Ernest[3], Bradley and Tibshirani[2], the permutation test is an exact test, meaning that if we set a test at α level, then the *Type I Error* of the permutation test is going to be less than or equal to α. Thus in sections 2.2 and 4.2, we can find p-values of permutation tests, which are carried out at the 5% level, smaller than(or approximately exactly) 0.05. Note that althought some values are larger than 5%, for instance the 0.0522 in Table 2, we have to also take into consideration the Margin of Error. So in the case of the 0.0522 in Table 2, once we subtract the Margin of Error, 0.006, we will get a lower bound of 0.0462, which does not contradict the idea that the permutation test is an exact test. The p-value density plots with small sample sizes also showed a lower density for the permutation test, which means that it rejected the least amount of times, i.e. it has the lowest *Type I Error* of the tests. The bootstrap tests, on the other hand, are only guranteed to be accurate as the sample size goes to infinity [2]. This is why we could see such a large *Type I Error* in the case of the bootstrap methods in the setting of small sample testings. When we consider the case of large samples, on the other hand, since the distribution functions of the permutation and bootstrap tests were uniformly close in the sense that the supremum of their difference tends to 0 as sample size gets large [4], it is not surprising to see that in both of the large sample cases we have very similar density curves and very similar *Type I Error* when the tests are carried out at the 5% level, which illustrated Romano's ideas.

Then we will consider Chapters 3 and 5, where we were comparing the *Power*. Since we are using the same test statistic as the one we used to test for *Type I Error*, analogous results from the concept of *Type I Error* will still hold here[4]. Thus we could see in the small sample case that the permutation test is still the test that rejects very conservatively, and the bootstrap tests reject the null hypothesis more aggresively, resulting in a higher power when the null hypothesis is actually false. Also, as before, we could observe a huge similarity of the power curves for large samples, which again confirmed Romano's idea using simulations.

6.2 Conclusions

We can conclude from the above simulation results that when we are comparing small samples of same size and equal variances, the seperated bootstrap test has the highest *Type I Error* rate and *Power*, the concatenated bootstrap test, although similar to the remaining tests, have a relatively high *Type I Error* rate and *Power*. The other tests behave similarly from the simulation results.

We could also conclude that starting at the sample size of 100 when the samples are of equal variance of 1 and equal sample size, the choice of the test would not make much difference.

Acknowledgement

I would like to show my deepest appreciation to Professor Ery Arias-Castro for his inspiring guidance and constant encouragement throughout the project. This project would be impossible without his countless hours of counseling and insightful comments on the key points. I would also like to thank my friend Ziyang Diao, for always giving me encouragement throughout the process of writing this paper.

References

[1] Ery Arias-Castro. *Two Sample Numerical Data*, 2016. UCSD Computational Statistics Lecture Notes.

[2] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall/CRC, 1994.

[3] Michael D. Ernest. Permutation methods: A basis for exact inference. *Statistical Science*, 19(4):676–685, 2004.

[4] Joseph P. Romano. Boostrap and randomization tests of some nonparametric hypothesis. *The Annals of Statistics*, 17(1):141–159, 1989.