Simulation Study of Equal Variance Assumption for Wilcoxon Rank Sum Test

George Daher

2023-12-19

INTRODUCTION

Parametric tests are almost always stronger than non-parametric provided the relevant parametric assumptions are met. These exact assumptions vary depending on the test being conducted but generally they require data to have a shape that is equal to a certain distribution with some parameter (or possibly multiple parameters). However, if the parametric assumptions for a method are not met then statisticians must use non-parametric alternatives.

If the goal is to test equality of center between two datasets of unpaired data, then there are a number of different methods that could be used, depending on the context. However, there are methods for testing center that are generally better (and others that are generally worse) in terms of power and required assumptions. One of the most common parametric tests of center is Welch's t-test and Mann Whitney Wilcoxon's rank sum (abbreviated MWW) test is generally considered its non-parametric alternative. The nature of Welch's t-test being a parametric test is that it has stricter assumptions for the data involved. Both tests require data to be identically and independently distributed with equal variance, but Welch's t-test also assumes that the data is approximately normal. In cases where the data is approximately normal relative efficiency calculations show that the MWW test is 95% as efficient as Welch's t-test. However, if the normality assumption is violated then the MWW test is more robust than the t-test.

The ability of the MWW test to ignore whether data is approximately normal is important. Especially with smaller sample size, the normality assumption is can be hard to satisfy and could require a separate test of normality to check. But the MWW test still requires that equal variance between the data sets being tested. Similarly to the normality assumption, the equal variance assumption can be difficult to satisfy for small sample sizes and could require a separate test of dispersion, such as the Ansari-Bradley test, to check. In both cases, performing multiple tests is undesirable because it lowers the significance of any conclusion and also affects the power of the entire testing procedure. In order to maintain consistency, individual significance levels for each test must be reduced or a combined test such as Lepage's Test must be considered.

In this paper we will examine the behavior of the MWW test and Welch t-test under departures from the unequal variance. This involves examining the distribution of test statistics and p-values across many simulations under slight to severe violations of the unequal variance assumption. Then we will compare overall procedure significance levels for performing a test of dispersion and then testing centers, testing centers without testing dispersion, and using Lepage's test to simultaneously test dispersion and center.

DECIDING SIMULATION PARAMETERS

The first step is to decide how the simulations should be run. This involves deciding both the distributions that should be simulated from and what conditions, in terms of true standard deviation ratio and shift in center the data should be sampled from.

In simulating parametric tests, the amount of choice in deciding which distributions to simulate from is considerably lower because the assumptions are stricter, unless the assumption violations are being tested via simulation. Due to the assumption of normality used for the t-test, one of the distributions we will use to simulate will be the normal distribution. Even though we are testing the performance of MWW, a non-parametric test, we will simulate from the normal distribution to establish a baseline to compare behavior between the t-test and Wilcoxon.

Considering that the goal is to analyze Wilcoxon test behavior under violations of the unequal variance, the focus of the simulations is to consider a number of non-parametric conditions. As a result, while the normal distribution is good enough the only distribution that will completely satisfy the parametric assumptions of the t-test, deciding which non-parametric distributions to simulate over is more difficult. Notably, there are a number of factors that can be considered. In order avoid over-complicating the simulation, because there are already a number

Deciding which distributions non-parametric distributions simulate other than the normal distribution is a more difficult decision. There are a number of factors that could be considered, but we decided to focus on choosing distributions to simulate from by considering skew and kurtosis. We first decided to consider distributions with varying kurtosis. The first reason behind this was due to the general fact that tail behavior of data is important to consider when evaluating parametric assumptions. Specifically, while the shape of the data could vary in terms of how many peaks, space between peaks and more, the kurtosis can be used as a general evaluation of tail width. Additionally, we can consider that the t-test and MWW test use two different measures of center. While the t-test considers the mean of the data, the MWW test uses the median of the data. Consequently, we would expect the MWW test to more resistant to datasets with high kurtosis. This means in addition to providing another setting in which to evaluate violations to the equal variance assumption in the Wilcoxon and t-tests, simulating distributions of varying kurtosis also allows more direct comparison of Welch t-test and MWW test performance.

Next, we considered how to test whether symmetry in the data affects results of the t-test and MWW test. It is important to note again that the t-test and MWW test use different measures over center. In symmetric distributions, this does not matter as much (except when considering kurtosis), because the expected mean and expected median are equal. Consequently, in symmetric cases we would generally expect the performance of the two-sample t-test and MWW t-test to be more similar. This means, similarly to with varying kurtosis, varying levels of skew (as a way to look asymmetry) can be used to examine MWW behavior on truly non-parametric with varying levels of center shift and variance ratio but also compare t-test behavior as the t-test assumptions are violated more and more. Specifically, we will focus on functions with a skew greater than 2 because that is generally seen to be a large skew.

As a result of this decision-making process, interspersed with some small preliminary simulations to examine basic behaviors, we decided on 6 distributions to simulate over, three symmetric distributions and 3 asymmetric distributions. First, as mentioned before, we are using the normal distribution, that is symmetric (meaning 0 skew) and has a kurtosis of 3. Second, we are using the unform distribution, because it also symmetric and has a lower kurtosis of $\frac{9}{5}$ than the normal distribution. Third, we are using the Laplace distribution, because it is symmetric but has a higher kurtosis, of 6. Fourth, as the first asymmetric distribution, we are using a chi-square distribution with 1 degree of freedom because it has a moderate skew of $2\sqrt{2}$ and a very large kurtosis of 15. Fifth, as another asymmetric distribution, we are using an exponential distribution with a rate equal to 1 because it has moderate skew of 2 and large kurtosis of 9. Sixth and finally, we are using the lognormal distribution with mean equal to 0 and variance equal to 1,

because it has a high skew and high kurtosis (approximately 6 and 100, respectively).

After deciding which distributions to use for the simulations, the next step was to decide under which conditions to run each of the simulations and which parameters to vary between simulations. The first two variables that we will vary between simulations will be the true standard deviation ratio and true shift in center. We decided to use standard deviation ratio rather than variance ratio out of convenience, but also because changes in standard ratio are less severe than equivalent changes in variance ratio. Specifically, we decided to use a true standard deviation ratio of 1, 1,25, 1.5 and 2 as levels of the parameter and we choose to have 4 levels for this parameter because the main goal of the simulation is to examine the Wilcoxon and t-test performance when the equal variance distribution is violated at different levels, which is what the standard deviation ratio (SDR) accounts for. Then, we decided that it would be best to account for shifts in center by number of standard deviations. After some preliminary analysis, we found that having setting the levels for true shift in center to be 0, 0.5 and 1 standard deviations was a sufficient. The shift in center could be done on an absolute numerical scale, but to the different types of distributions involved, we found that shifting the true center by some number (or fraction) of standard deviations was more consistent overall. Finally, we also decided to vary whether sample size between the two datasets simulated was equal were not, if . This way we could evaluate how the relationship between sample size ratio and standard deviation ratio affected test performance. Specifically, we decided to examine when the sample size ratio (of the smaller standard deviation to larger standard deviation data) was 1:2, 1:1 and 2:1 because these levels appeared to have some impact on simulation results.

Before running the simulations, we can note that the complexity of running a simulation where more parameters are varied grows exponentially. If we are only to vary one parameter, then we would only end 3-4 simulation, but as each parameter is varied then the number of simulations that need to be run increases multiplicatively by the number of levels the parameter is being run at. In this case, because we were using 6 distributions, 4 levels of SDR, 3 levels of true shift in center, and 3 levels of sample size ratio this meant that we needed to run 216 simulations. This is especially cumbersome because we wanted a very large number of samples runs for each simulation in order to mitigate the variability due to the random sample selection. Additionally, we are computing the Welch t-test, Wilcoxon Rank Sum Test, Ansari-Bradley Test of Dispersion and Lepage Test for each iteration, so there is another level of complexity. In total, while more simulations for each combination of parameters would have been better, we decided to run each simulation 30000 times. Running all the simulations ended up taking about 8 hours of compute time, split between 2 different computers (4 hours on each running simultaneously). The initial goal of 50000 would have taken about 12 hours of compute time.

SIMULATION METHOD

All the simulations were running using R where with each set of parameters is simulated from 30000 times. In each case an identical distribution is used to generate two different samples with the prespecified sizes (based on the sample size ratio). Then the y dataset is transformed by the specified SDR and shift in center. The reason that the standard deviation and center are adjusted after the sample is generated rather than adjusting the distribution from which the sample is drawn is due to implementation. With the method we ended up using, adjusting after the fact was easier and leads to the same results, provided the adjustment is the same regardless of the sample drawn. Once the two datasets are generated and appropriately adjusted (if needed), a t-test, MWW test, AB test and Lepage test are then each run. Due to the computational complexity challenges outlined above, these methods are all run using asymptotic behavior, which should be fairly accurate because the sample sizes used are 20 or larger. For each simulation iteration, the p-value and test statistics of each test are recorded.

After the simulations are done, conclusions are determined based on analysis and comparison on the distribution of the resulting p-values and test statistics for each test. Additionally, by comparing distributions where all parameters are equivalent except for the true shift in center (meaning the null hypothesis is true and true location shift is 0, or the true location is 0.5 or 1 standard deviation, as was decided as a parameter level above), conclusions can be made can be made on the power of tests.

In the output, we examine the distribution of p-values calculated using the MWW test under parameters where the null hypothesis is true. The expected behavior if the test is working correctly is that the p-values are uniformly distributed. This is because if they are uniformly distributed then the p-values accurately reflect the probability of getting a result as or more extreme under a true null hypothesis. Additionally, we can calculate the power of a test by the proportion of p-values that are significant (less than the set significance level) when there is some shift in center, because in the simulation this was the proportion of cases in which the false null hypothesis was correctly rejected.

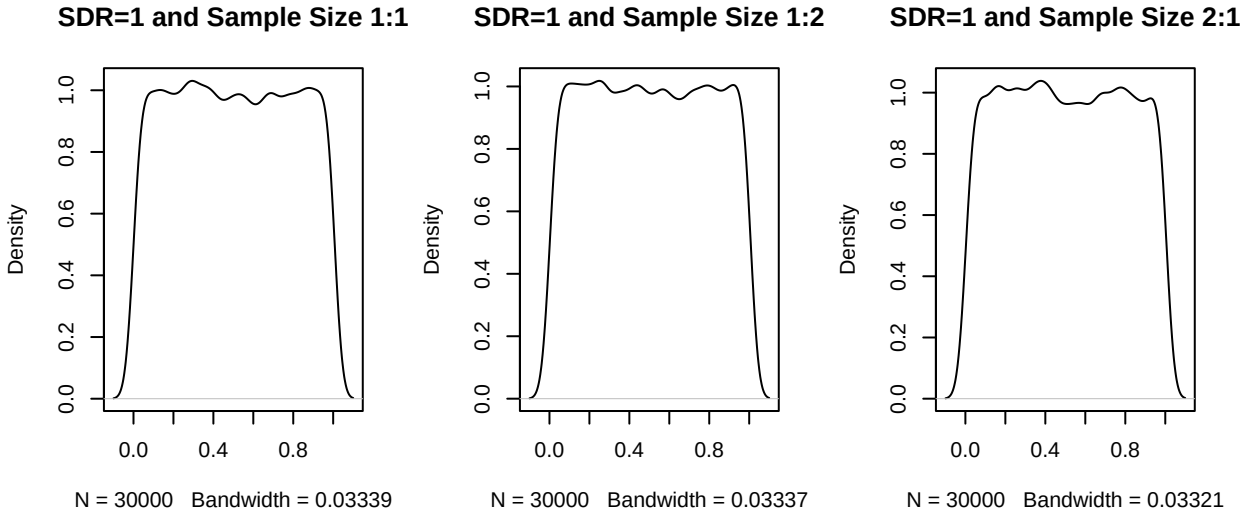
We also conduct very similar analysis over the p-value calculated Welch's t-test and the distribution of t-statistics under a true or false null hypothesis. The final step is to compare power and significance overall for both tests.

RESULTS

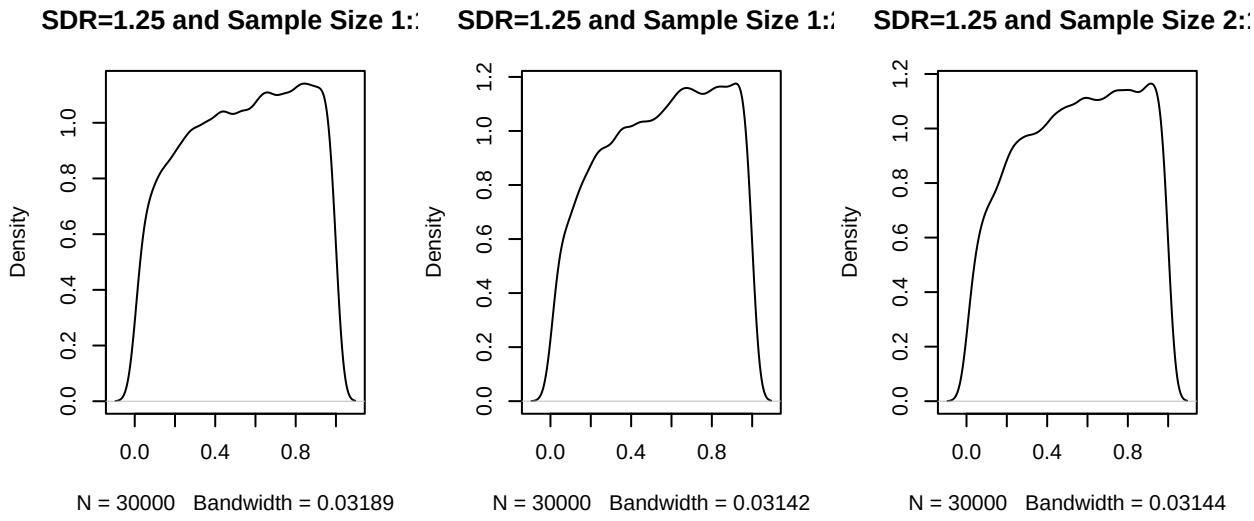
Part 1 - Wilcoxon Rank Sum Behavior under False Null Hypothesis

True Shift in Center = 0

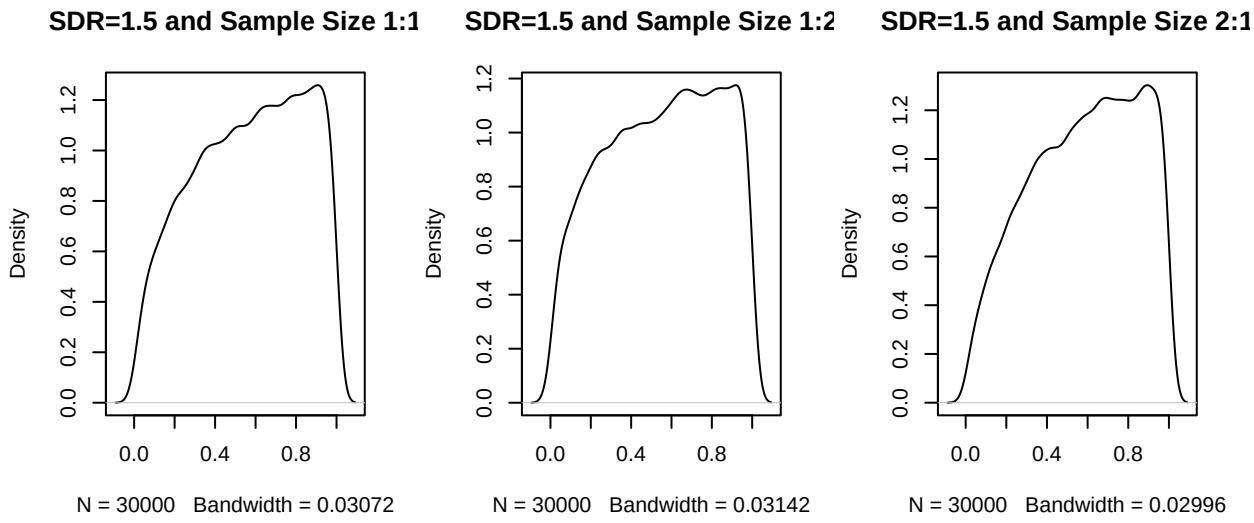
Normal Distribution with SDR=1



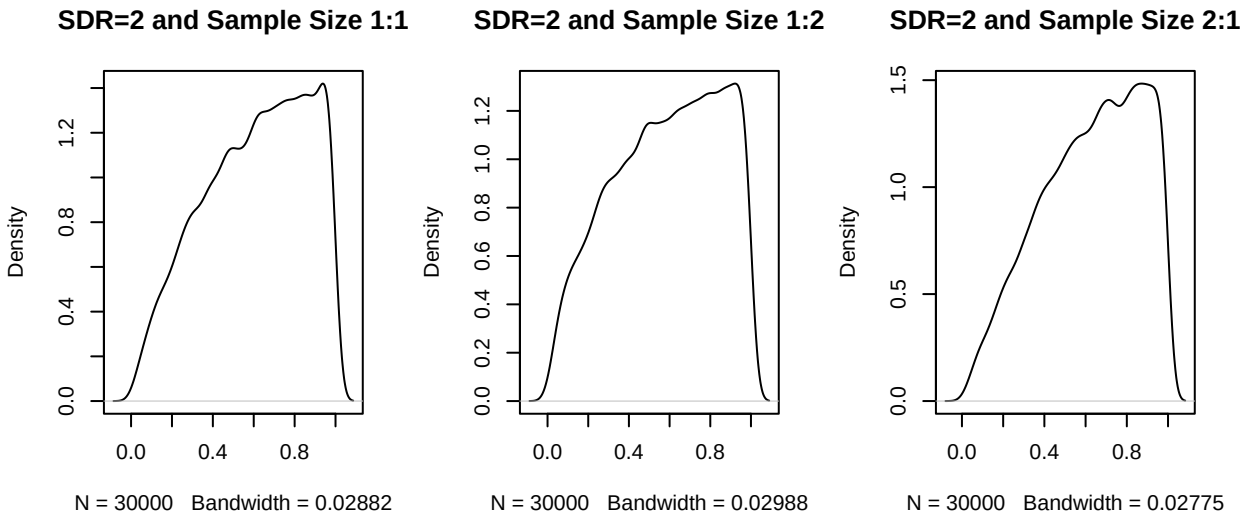
Normal Distribution with SDR=1.25



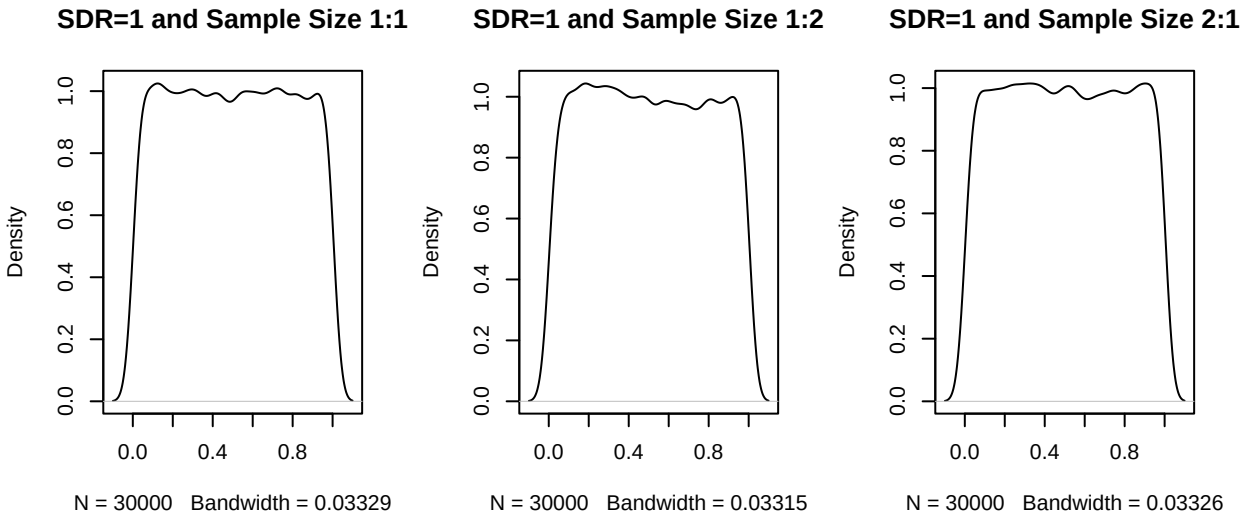
Normal Distribution with SDR=1.5



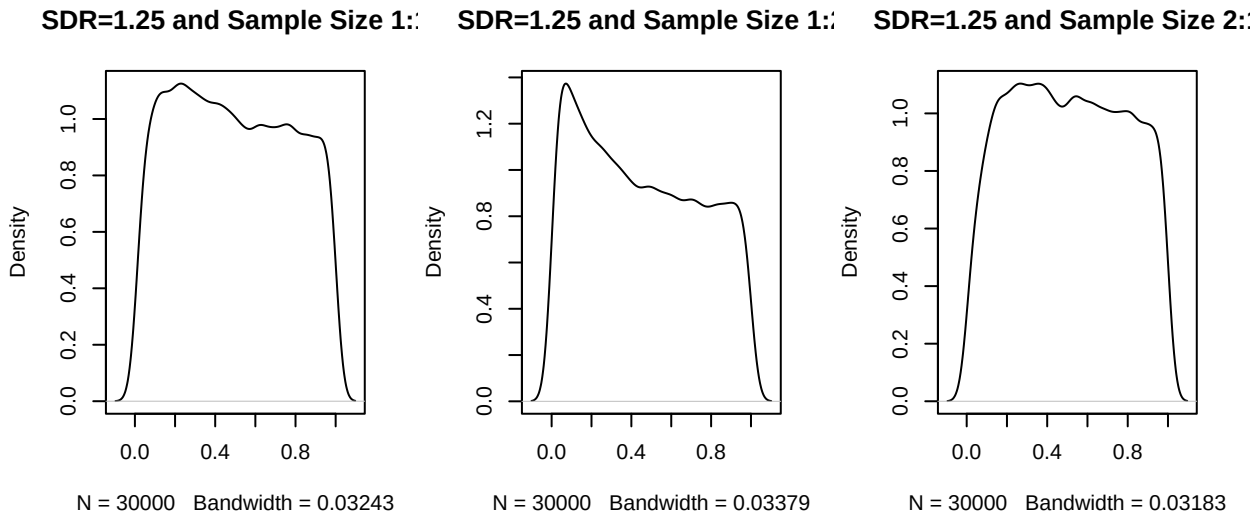
Normal Distribution with SDR=2



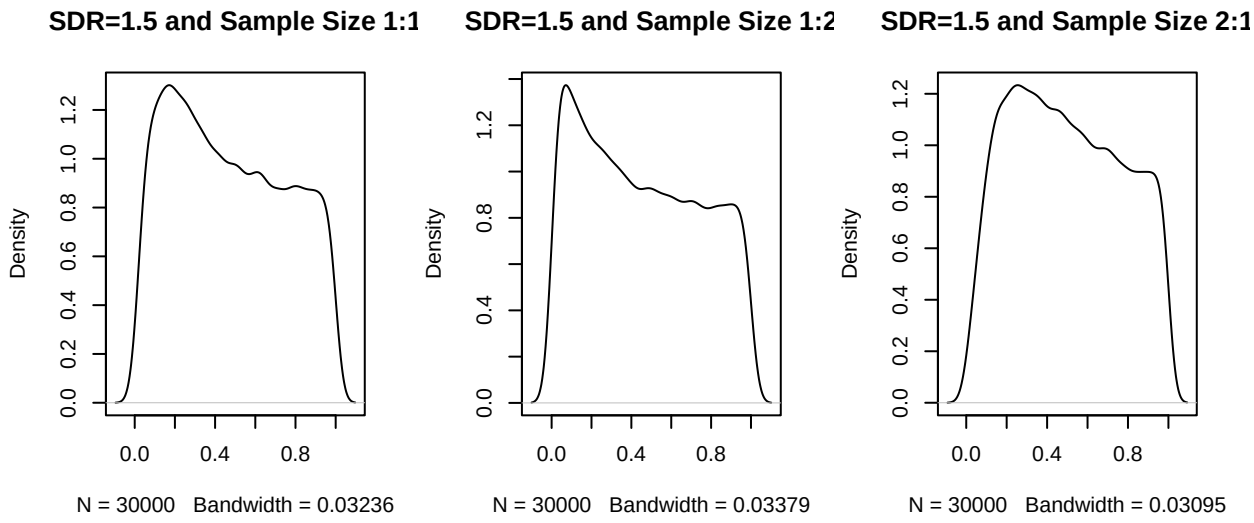
Chi-Square Distribution with SDR=1



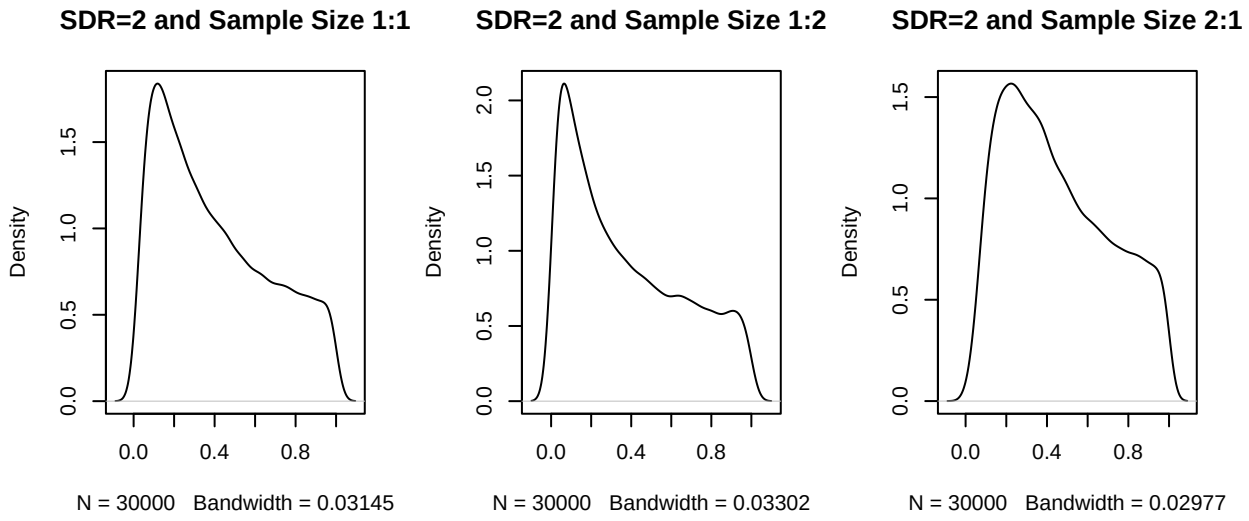
Chi-Square Distribution with SDR=1.25



Chi-Square Distribution with SDR=1.5



Chi-Square Distribution with SDR=2



Remarks

We can see for all the distributions used, under the null hypothesis and when the equal variance assumption is met (meaning SDR=1), the p-values are more or less uniformly distributed. This is given due to the performance of the test when the assumptions are satisfied but is worth showing for thoroughness.

However, as soon as we begin to violate the equal variance assumption, the p-values begin to lose their uniform shape and become skewed. In the case of the symmetric distributions, this shape is left-skewed and for asymmetric distribution, this shape is right-skewed. This means that the p-values calculated for the test are not accurate when the equal variance assumption is not met.

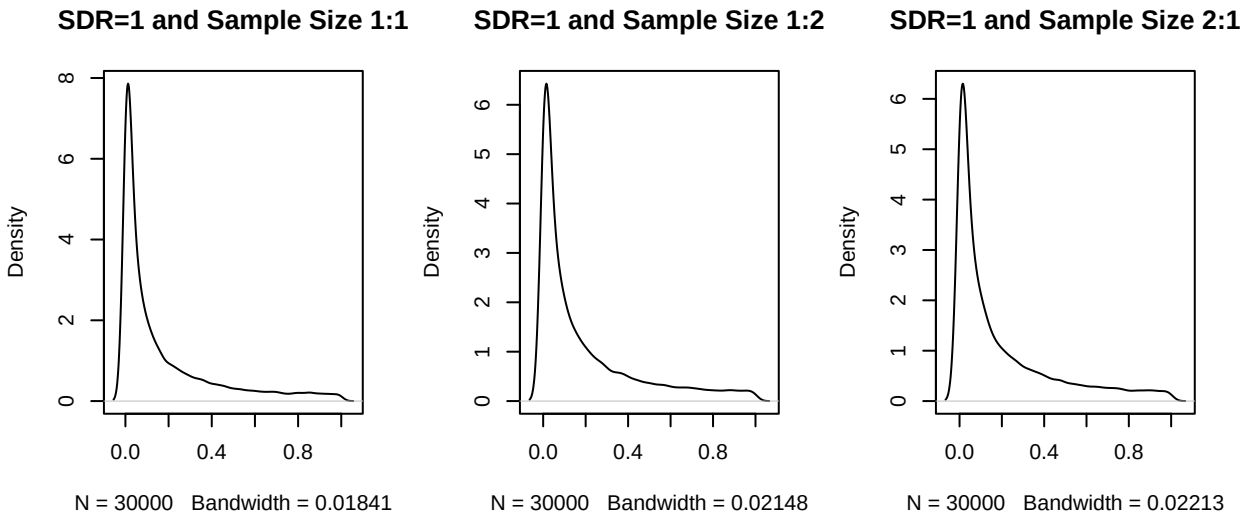
We can note that the p-values can still be used to see whether the test is significant if the data is symmetric. However, significance is a less frequent result meaning that the test is overall more conservative under these conditions. We can observe this because the distributions are left-skewed so the probability of receiving a p-value of p or less is less than p. In other words, according to the simulations, the actual p will be less than the observed p_value, so any observed significant results are still significant in this case. Conversely, the opposite case is true for the asymmetric data. The probability of receiving a p-value of p or less is greater than p. This means that the Type 1 error rate is reduced for symmetric data, and this usually indicates a decrease in test power, so we must examine these distributions under non-null conditions. For asymmetric data, power behavior is unclear.

A final note that we can make from this initial step is that the sample size ratio does have a minor effect on the distribution of the test statistic but it is a minimal effect compared the changes observable at different set SDR values.

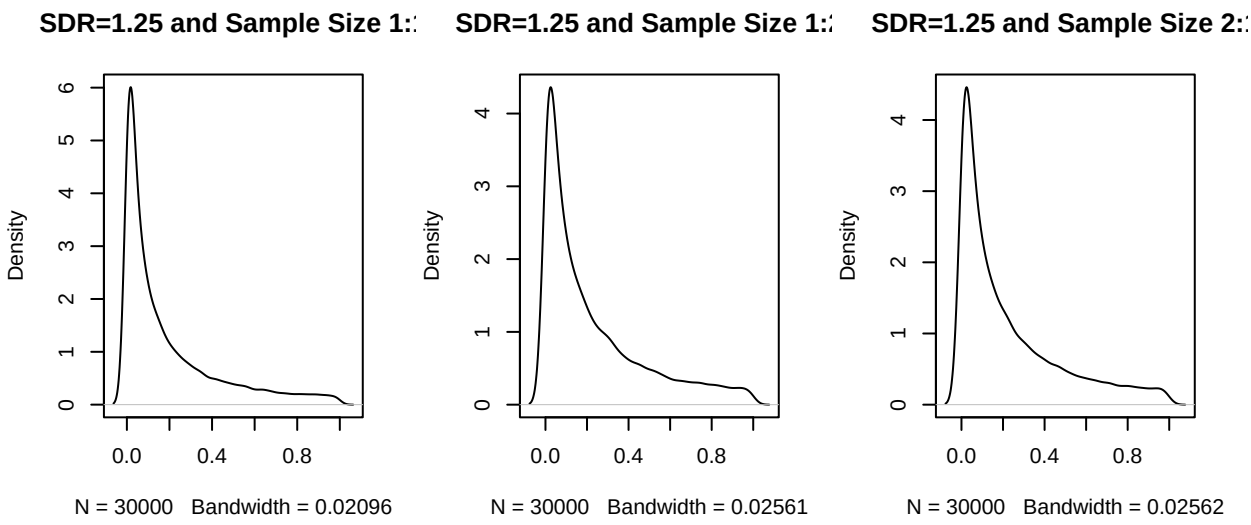
Part 2 - Wilcoxon Rank Sum Behavior under False Null Hypothesis

True Shift in Center = 0.5

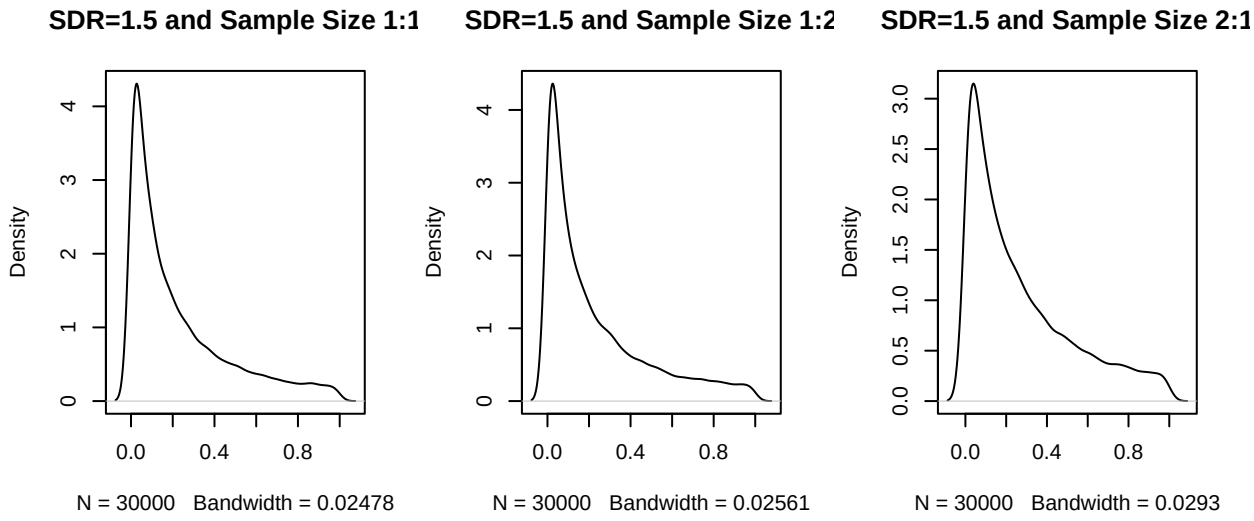
Normal Distribution with SDR=1



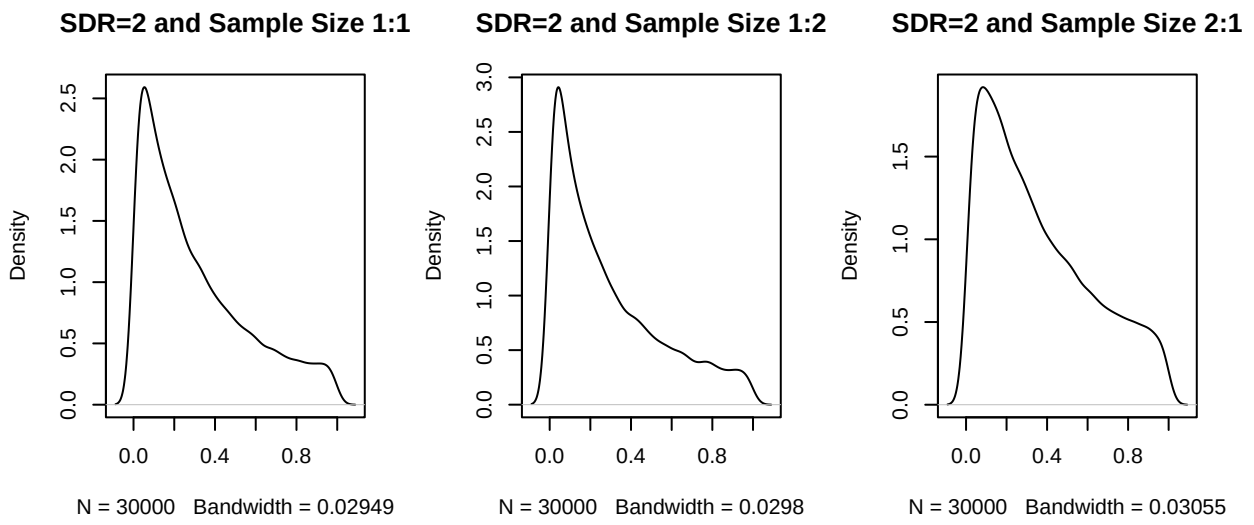
Normal Distribution with SDR=1.25



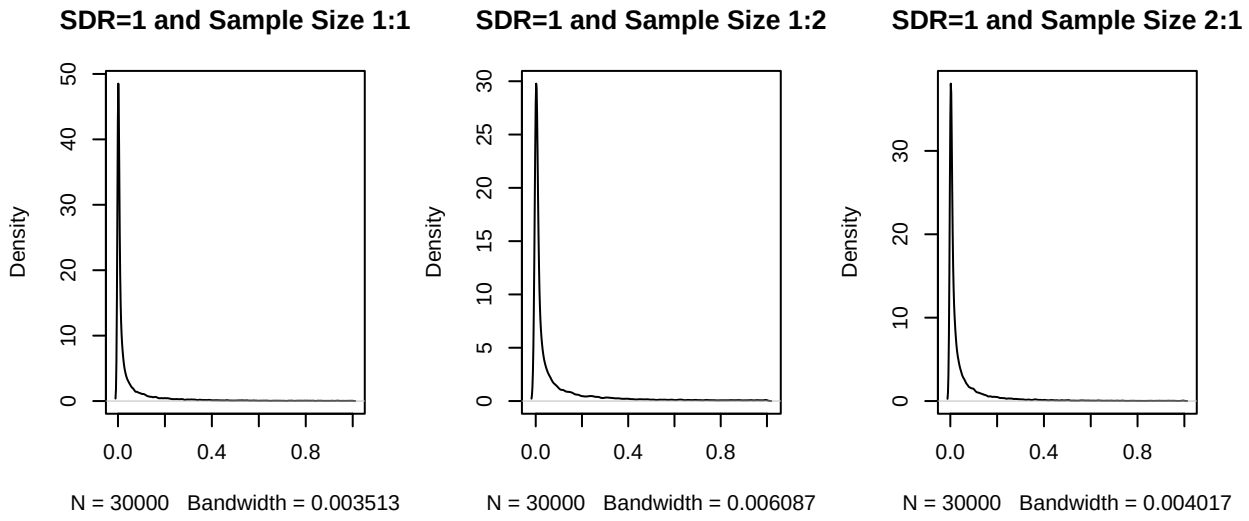
Normal Distribution with SDR=1.5



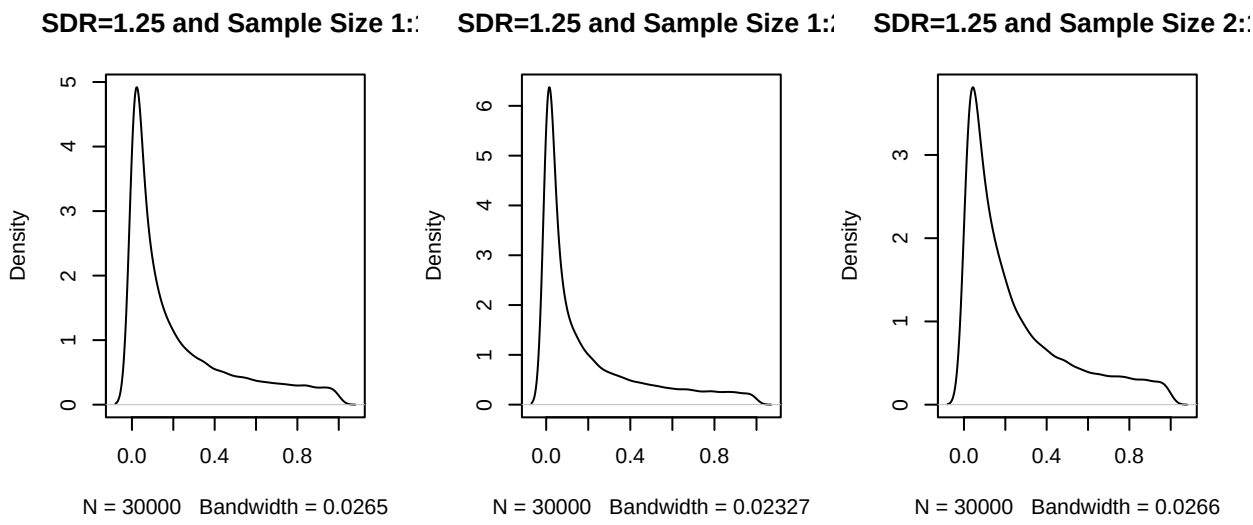
Normal Distribution with SDR=2



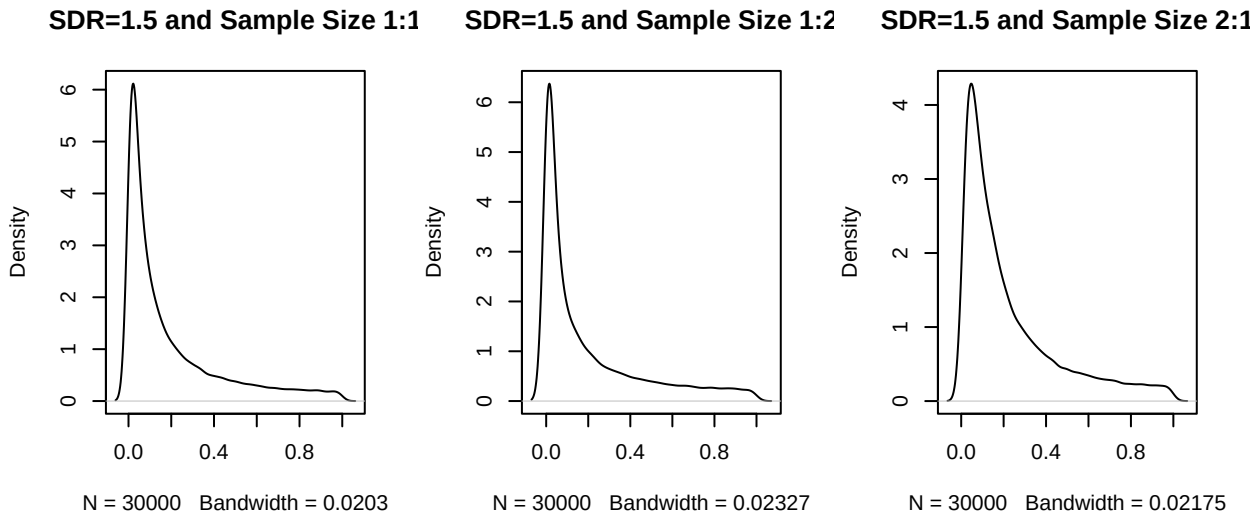
Chi-Square Distribution with SDR=1



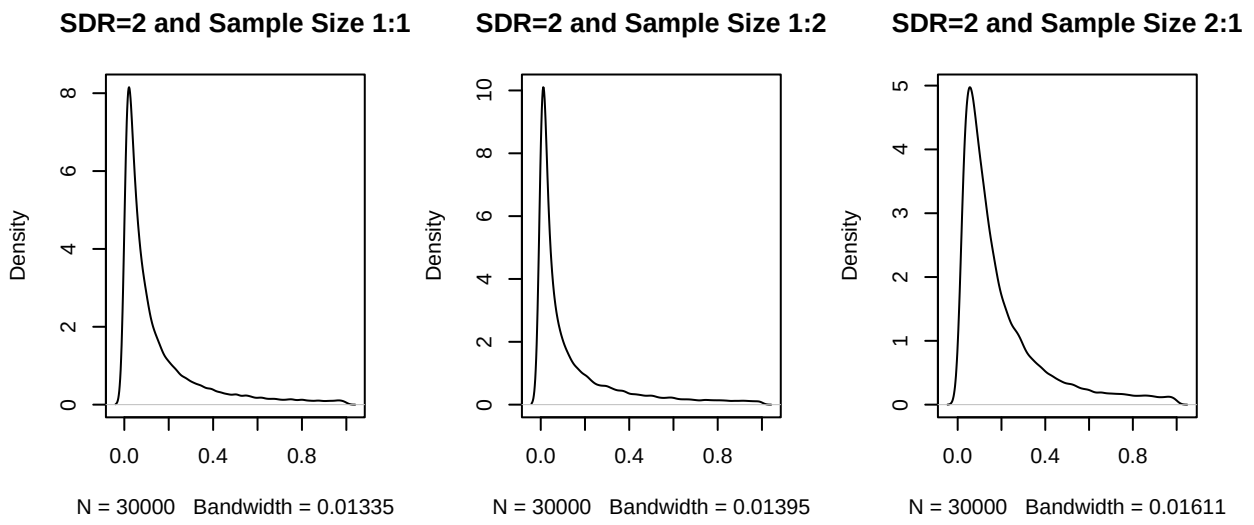
Chi-Square Distribution with SDR=1.25



Chi-Square Distribution with SDR=1.5

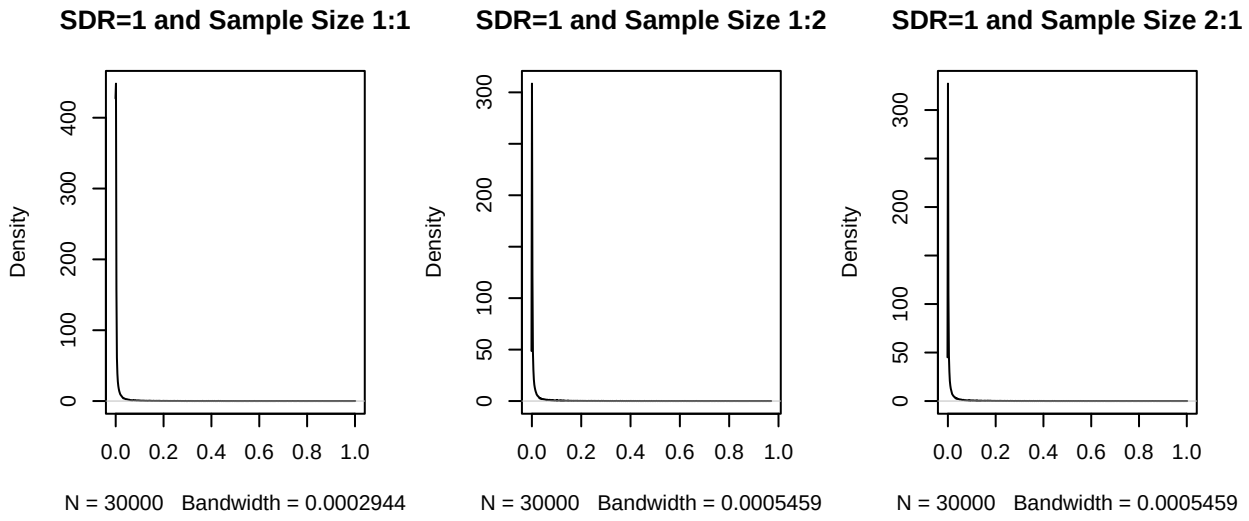


Chi-Square Distribution with SDR=2

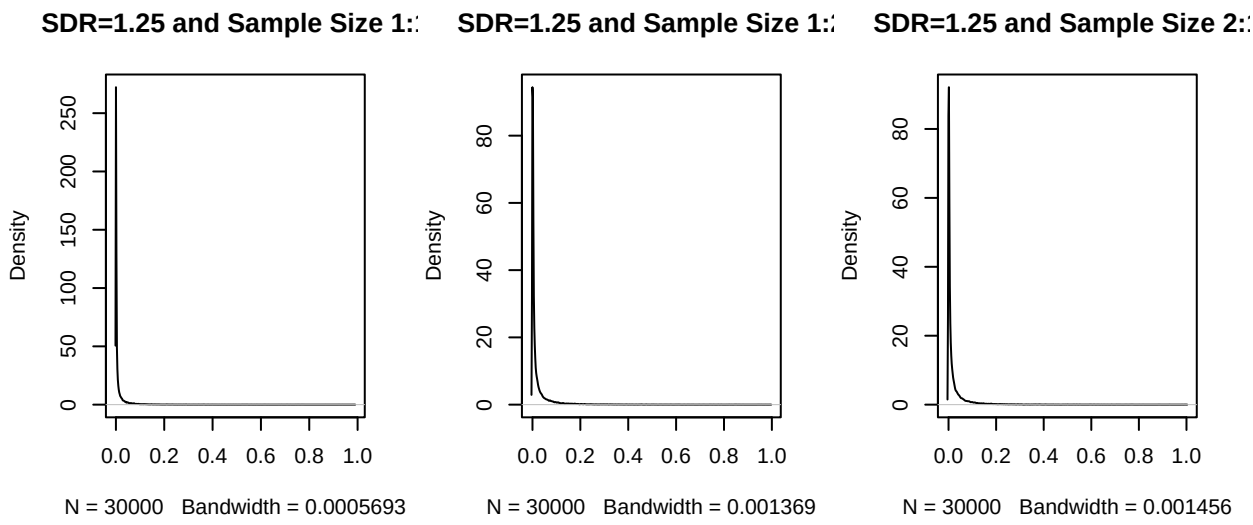


True Shift in Center = 1

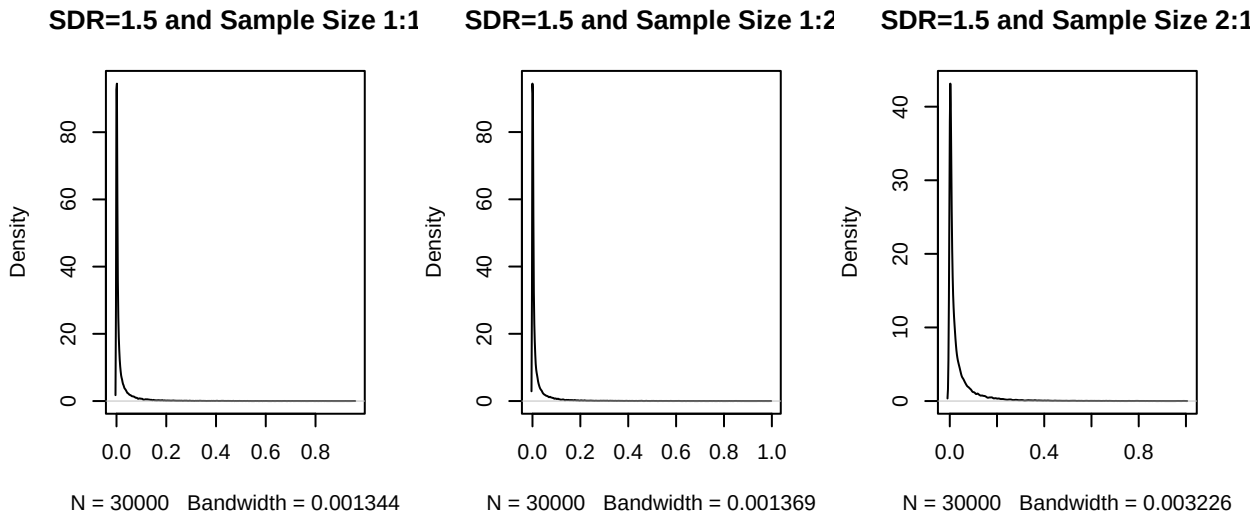
Normal Distribution with SDR=1



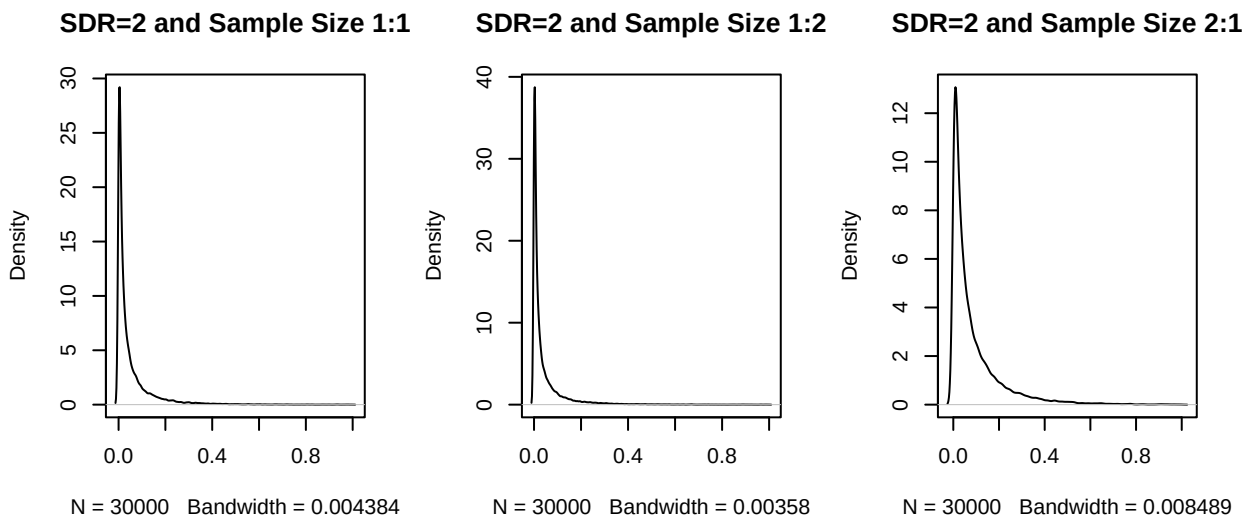
Normal Distribution with SDR=1.25



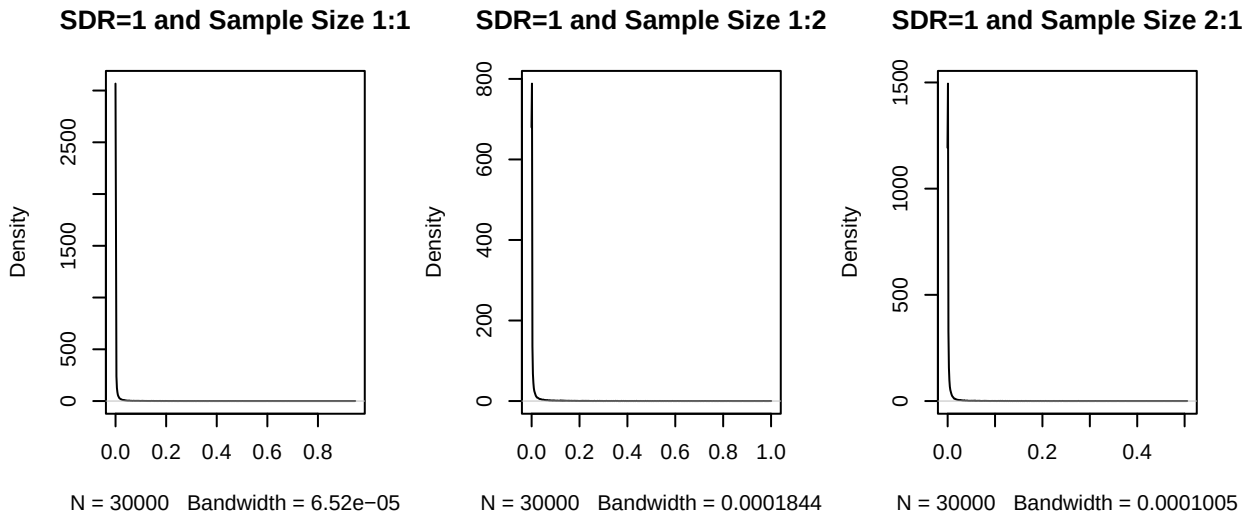
Normal Distribution with SDR=1.5



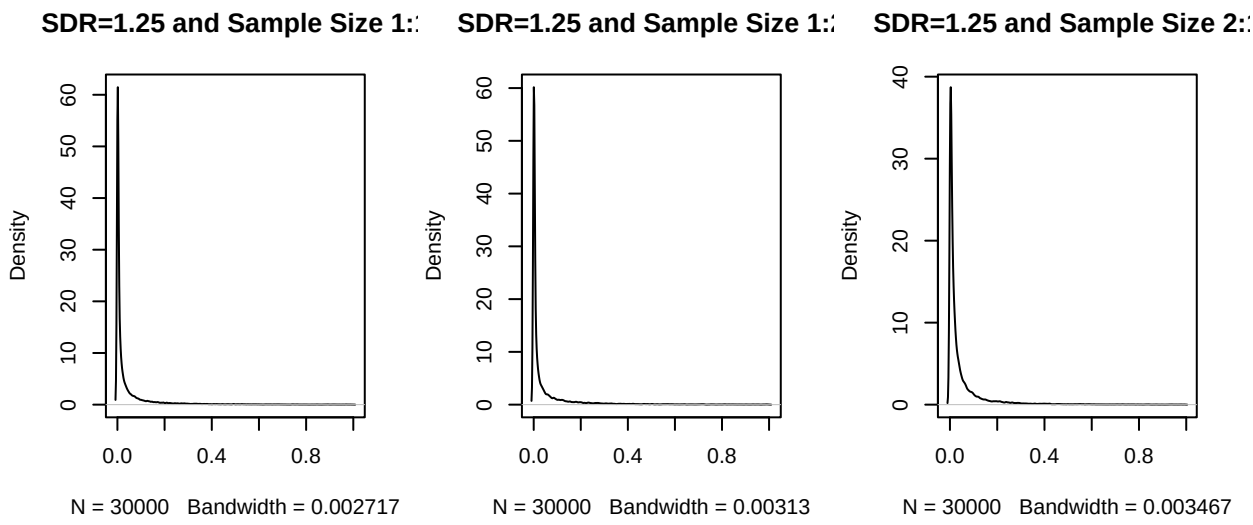
Normal Distribution with SDR=2



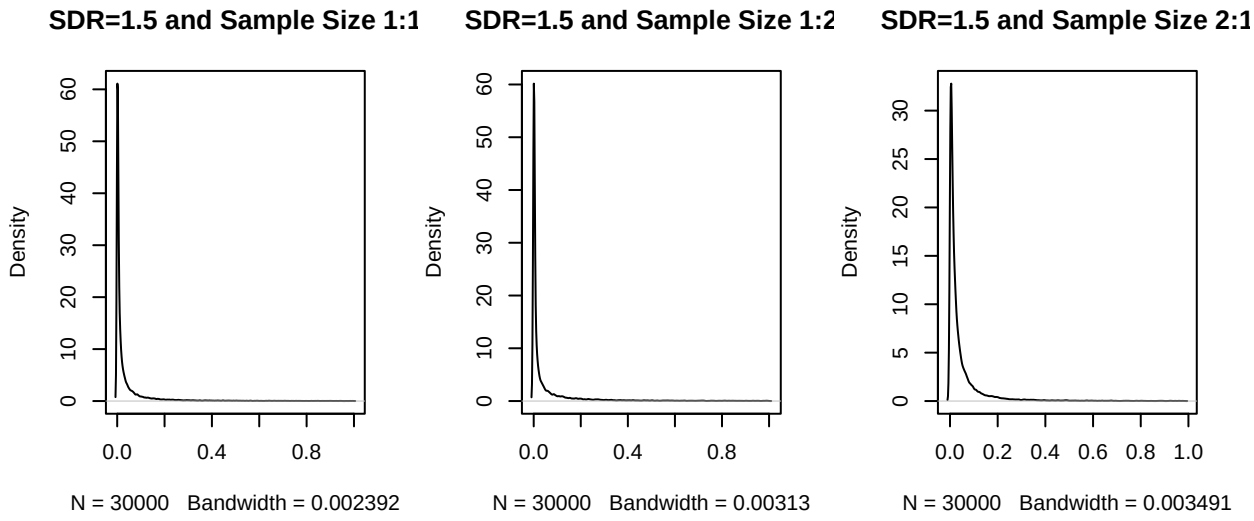
Chi-Square Distribution with SDR=1



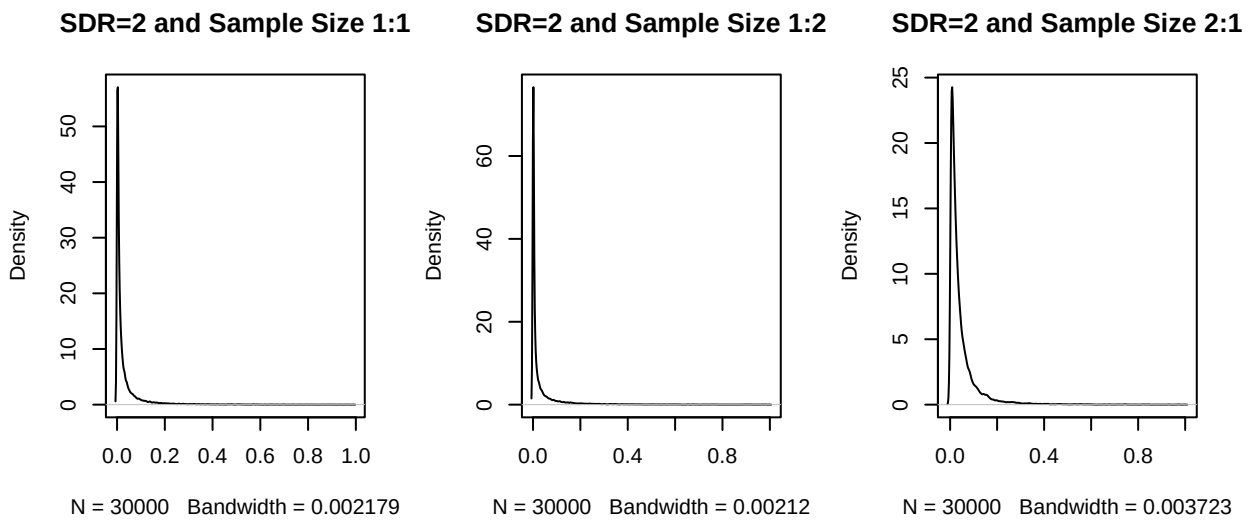
Chi-Square Distribution with SDR=1.25



Chi-Square Distribution with SDR=1.5



Chi-Square Distribution with SDR=2



Remarks

The following graphs show the distributions of p-values for a MWW test under various conditions where the null hypothesis of equal center is false. For the first set of graphs the true shift in center is 0.5 of a standard deviation and in the second set of graphs the true shift in center is 1 standard deviation.

We can see that in all the distributions used, there is a clear loss in power as the violation to the equal variance is more severe. Notably, as the standard deviation ratio increases, the tail weight of p-value distribution becomes noticeably larger. This means that the power of the test decreases as the SDR increases because a higher density of low p-values means that the test was more able to reject the null hypothesis of no difference in center, which we know to be false. As the true shift in center increases from 0.5 standard deviations to 1 standard deviation, this effect is still present but weaker because the power of the test increases as the true shift in center is further and further 0.

Part 3 - Power Calculations

Table of Power Values when True Shift in Center = 0.5

Distribution and SDR	Sample Size 1:1	Sample Size 1:2	Sample Size 2:1
norm SDR=1	0.4523000	0.4091333	0.4071000
norm SDR=1.25	0.3542000	0.2250667	0.3169667
norm SDR=1.5	0.2613667	0.2250667	0.2359667
norm SDR=2	0.1261333	0.1009333	0.1208333
unif SDR=1	0.4280667	0.3918667	0.3917333
unif SDR=1.25	0.3025333	0.1482667	0.2759333
unif SDR=1.5	0.1785667	0.1482667	0.1675333
unif SDR=2	0.0612000	0.0422667	0.0695333
lp SDR=1	0.3576667	0.3260333	0.3290000
lp SDR=1.25	0.2789000	0.1884000	0.2550000
lp SDR=1.5	0.2137000	0.1884000	0.1969333
lp SDR=2	0.1195333	0.0994333	0.1132333
chisq SDR=1	0.7703333	0.7007333	0.7517667
chisq SDR=1.25	0.7467000	0.6123333	0.7210333
chisq SDR=1.5	0.6792333	0.6123333	0.6269333
chisq SDR=2	0.3541667	0.3197333	0.3004000
exp SDR=1	0.7611667	0.6968667	0.7333000
exp SDR=1.25	0.6940667	0.5042000	0.6536333
exp SDR=1.5	0.5561667	0.5042000	0.4934000
exp SDR=2	0.2178000	0.1909000	0.1818333
lnorm SDR=1	0.1326000	0.1354667	0.1139667
lnorm SDR=1.25	0.0515333	0.0187333	0.0403000
lnorm SDR=1.5	0.0182667	0.0187333	0.0156000
lnorm SDR=2	0.0040000	0.0035333	0.0044000

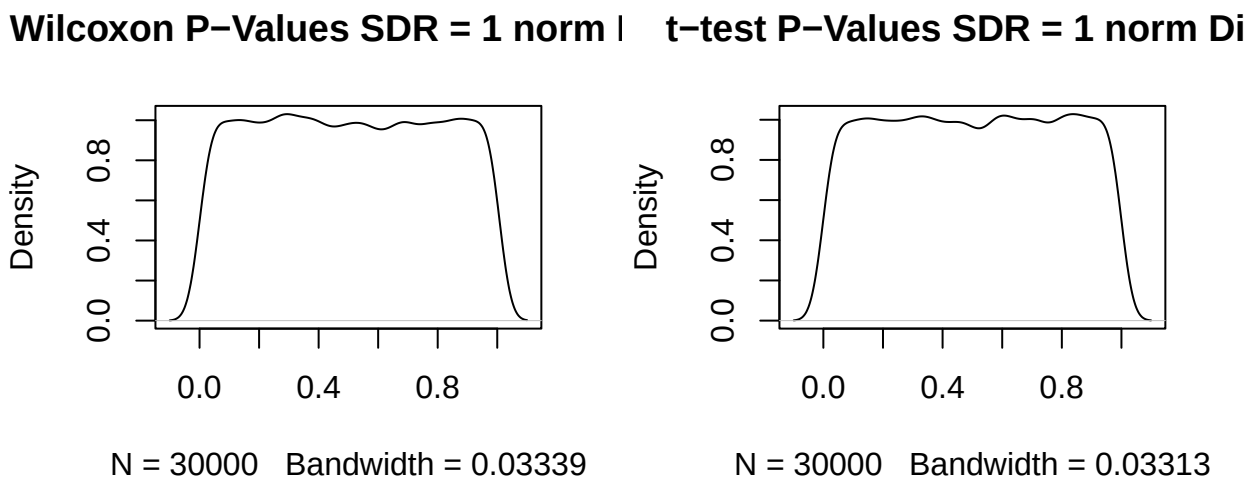
Table of Power Values when True Shift in Center = 0.5

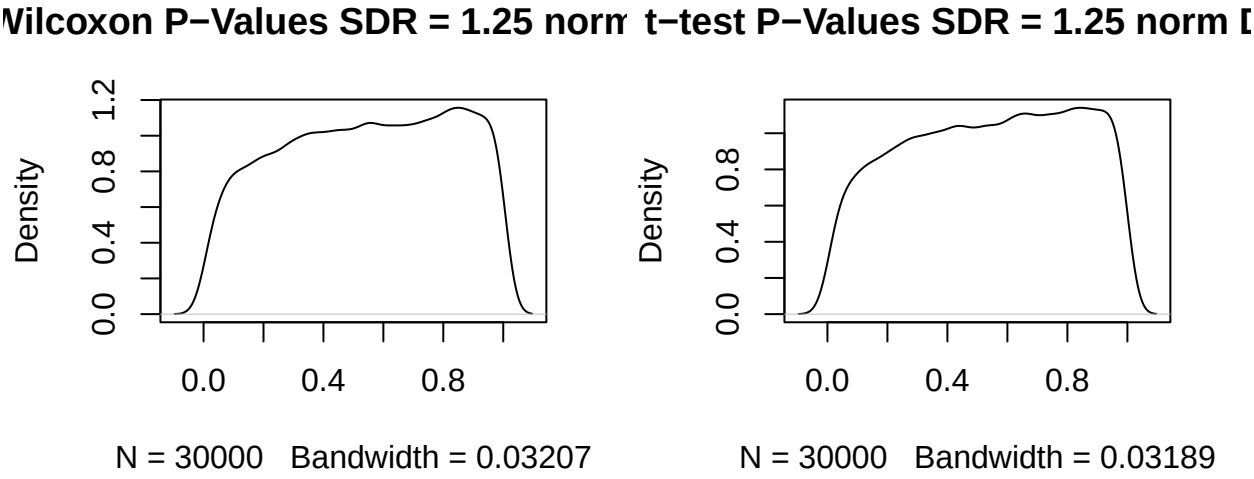
Distribution and SDR	Sample Size 1:1	Sample Size 1:2	Sample Size 2:1
norm SDR=1	0.9574667	0.9356333	0.9372333
norm SDR=1.25	0.9278333	0.8327000	0.8953000
norm SDR=1.5	0.8747333	0.8327000	0.8249333
norm SDR=2	0.7029667	0.6326667	0.6325333
unif SDR=1	0.9359333	0.9080333	0.9099333
unif SDR=1.25	0.8997000	0.7537000	0.8625667
unif SDR=1.5	0.8160333	0.7537000	0.7591000
unif SDR=2	0.4920667	0.4128000	0.4449333
lp SDR=1	0.8674667	0.8270333	0.8285000
lp SDR=1.25	0.8068333	0.6808000	0.7570000
lp SDR=1.5	0.7319667	0.6808000	0.6783333
lp SDR=2	0.5748667	0.5102333	0.5211000
chisq SDR=1	0.9845000	0.9562000	0.9880667
chisq SDR=1.25	0.9852667	0.9570667	0.9898333
chisq SDR=1.5	0.9855667	0.9570667	0.9899000
chisq SDR=2	0.9728000	0.9384667	0.9649333
exp SDR=1	0.9950000	0.9820667	0.9958667
exp SDR=1.25	0.9948667	0.9752667	0.9956667
exp SDR=1.5	0.9935000	0.9752667	0.9946667
exp SDR=2	0.9739667	0.9408333	0.9658000
lnorm SDR=1	0.3696000	0.3456667	0.3216333
lnorm SDR=1.25	0.2283667	0.1056667	0.1822000
lnorm SDR=1.5	0.1042667	0.1056667	0.0808333
lnorm SDR=2	0.0210667	0.0186667	0.0184333

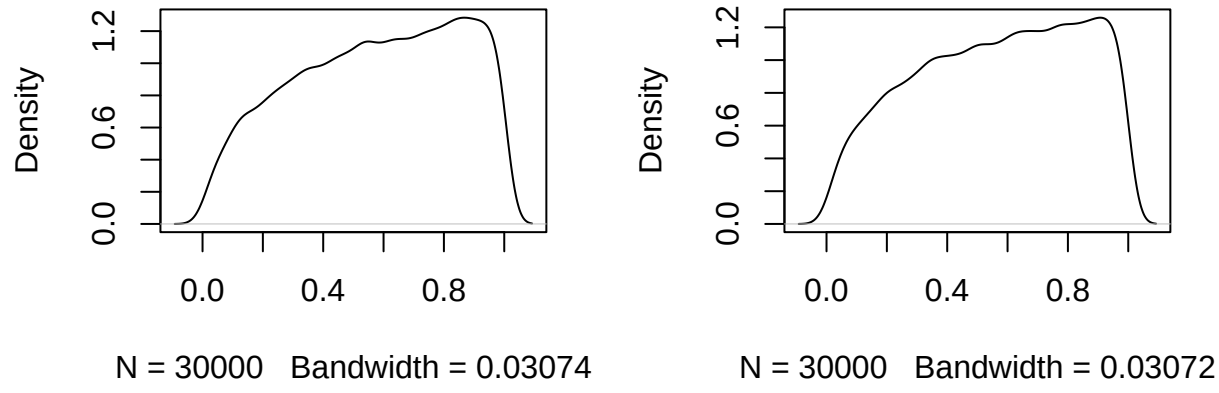
From these tables, we can observe numerically that as the SDR increases, the power of the test decreases in all cases. This makes sense when compared with our prior conclusions about the uneven distributions of p-values from the W test. It should also be noted that the rate at which power decreases for increases in SDR is not constant for all types distributions. Notably, the power of the test over an exponential or chi-square decreases less for equivalent changes in SDR than for other distributions. This behavior is likely due to some factor concerning the shape and skew of the underlying distributions do but more analysis would be needed for a better explanation of this behavior. The important result from these tables of calculated power values is that the power of the test decreases significantly as the SDR increases which means that as the equal variance assumption is more severely violated, the power of the Wilcoxon test decreases.

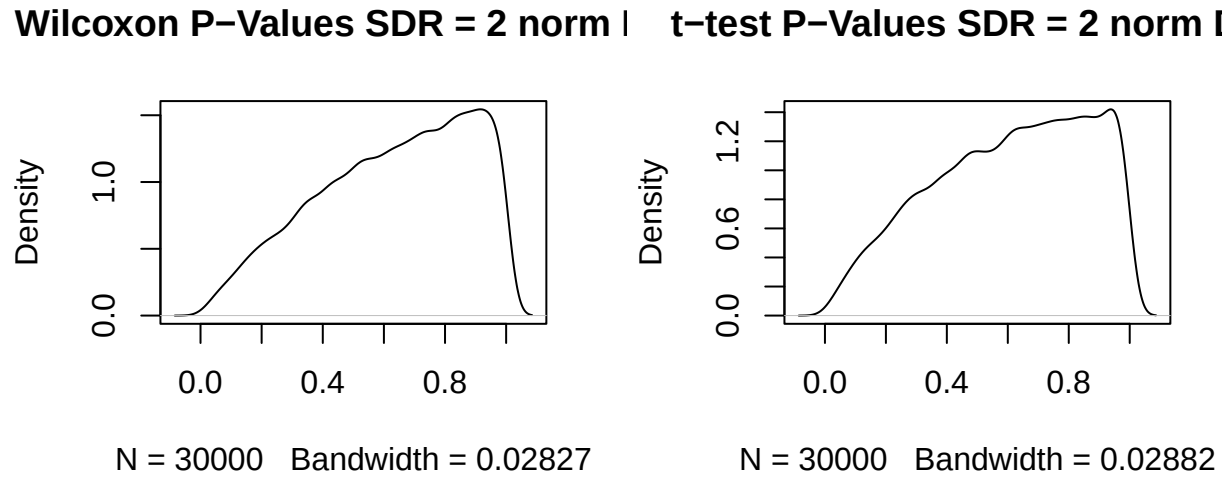
There is a caveat to this though because if we consider a larger shift such as one by 2 standard deviations, then the power of the test returns. Already, by comparing the power values for an true alternative center of 0.5 vs the power values for a true alternative center of 1 we can see that the decreases in power due to increased SDR are less dramatic when the gap between the null hypothesized center and actual center is larger.

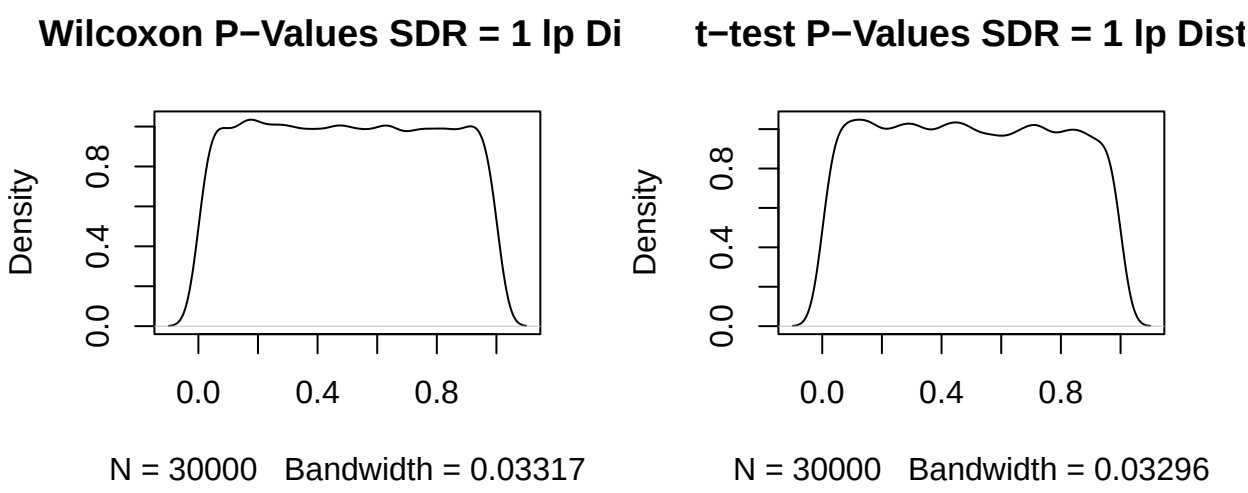
Part 4 - Comparing Wilcoxon Rank Sum vs. Welch's T-Test

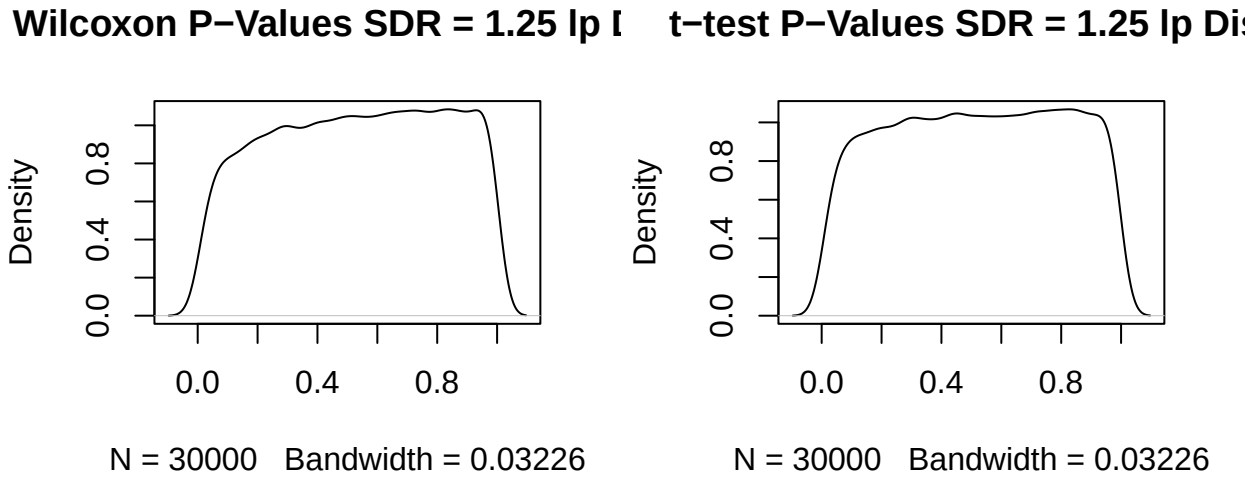


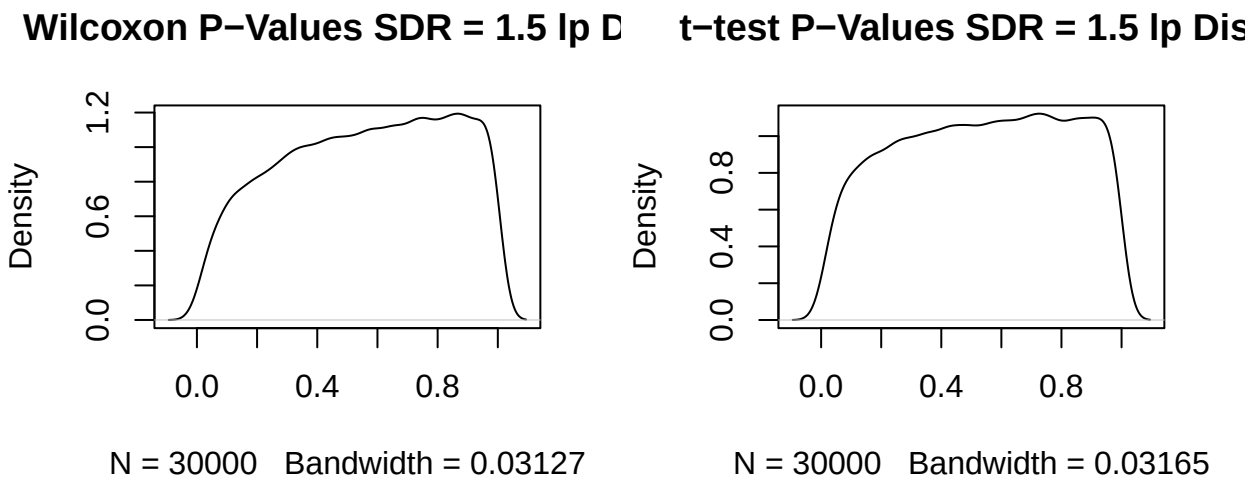


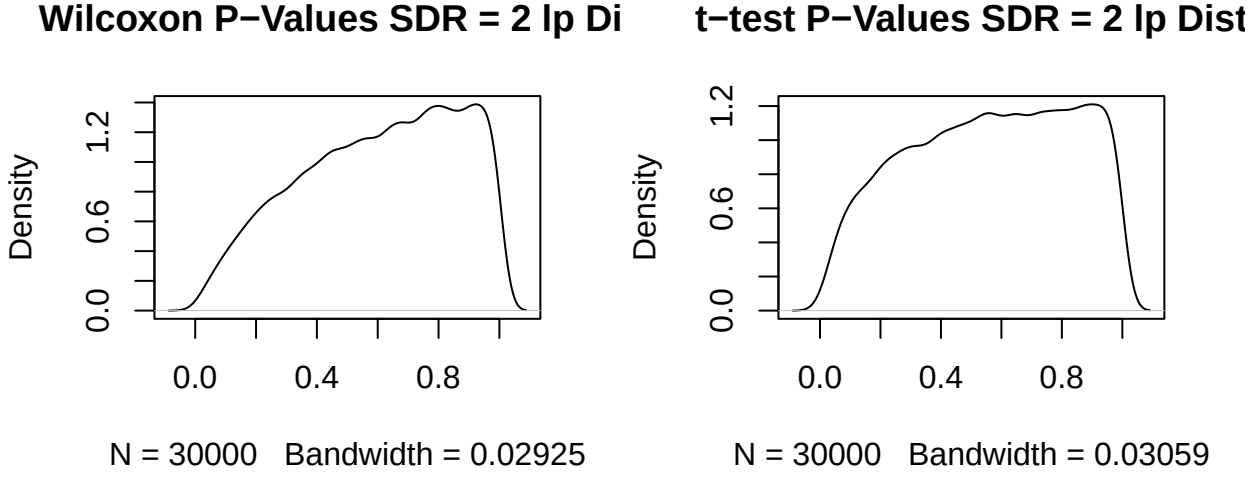

Wilcoxon P−Values SDR = 1.5 norm
t−test P−Values SDR = 1.5 norm D
Density
0.0
0.6
1.2
0.0
0.4
0.8
N = 30000 Bandwidth = 0.03074
N = 30000 Bandwidth = 0.03072

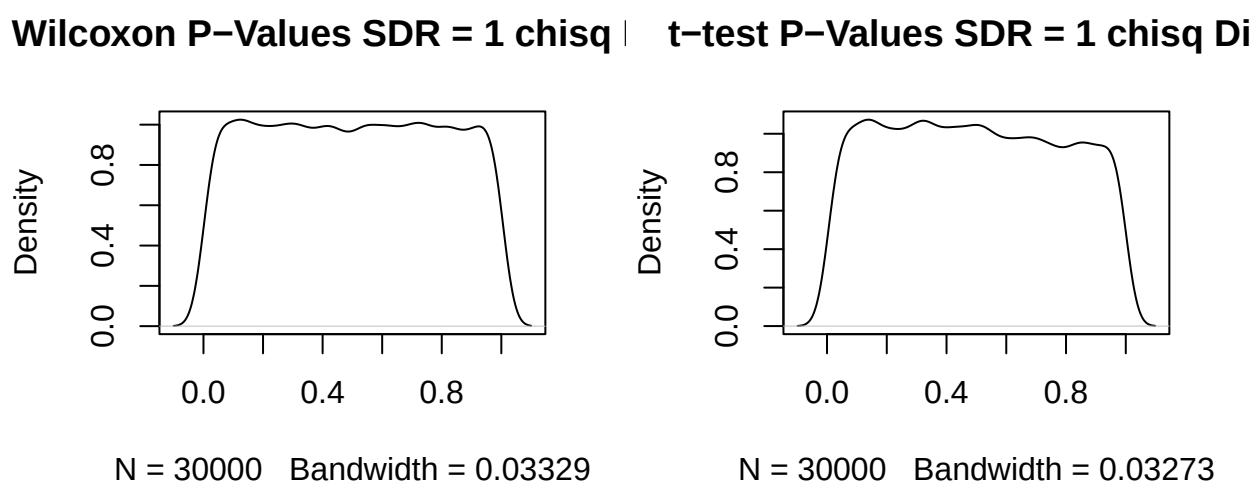

Wilcoxon P−Values SDR = 2 norm
t−test P−Values SDR = 2 norm Di
Density
0.0
1.0
0.6
1.2
0.0
0.4
0.8
N = 30000 Bandwidth = 0.02827
N = 30000 Bandwidth = 0.02882

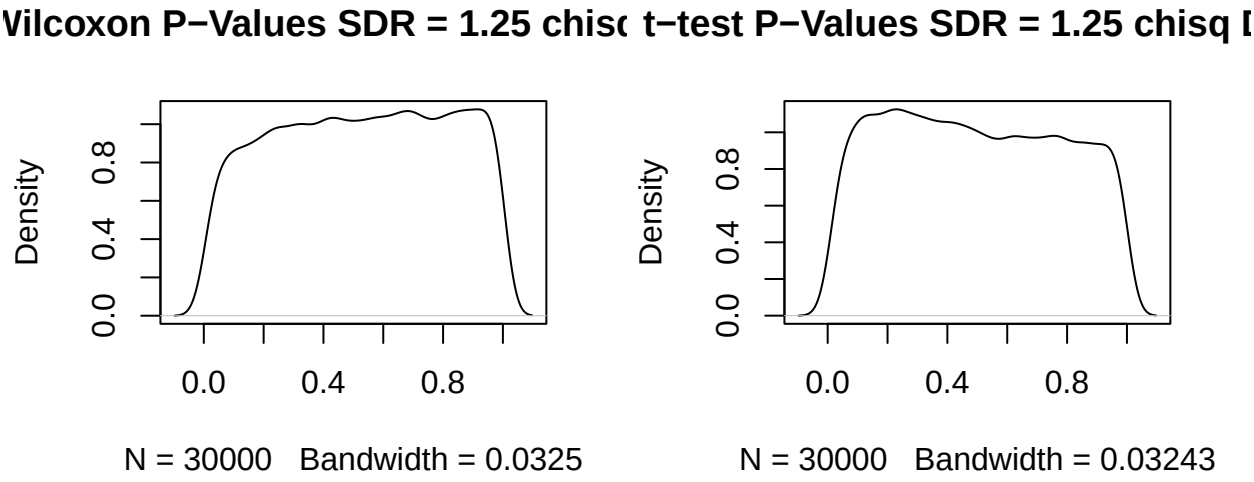

Wilcoxon P−Values SDR = 1 lp Di
t−test P−Values SDR = 1 lp Dist
Density
0.0
0.4
0.8
N = 30000 Bandwidth = 0.03317
N = 30000 Bandwidth = 0.03296

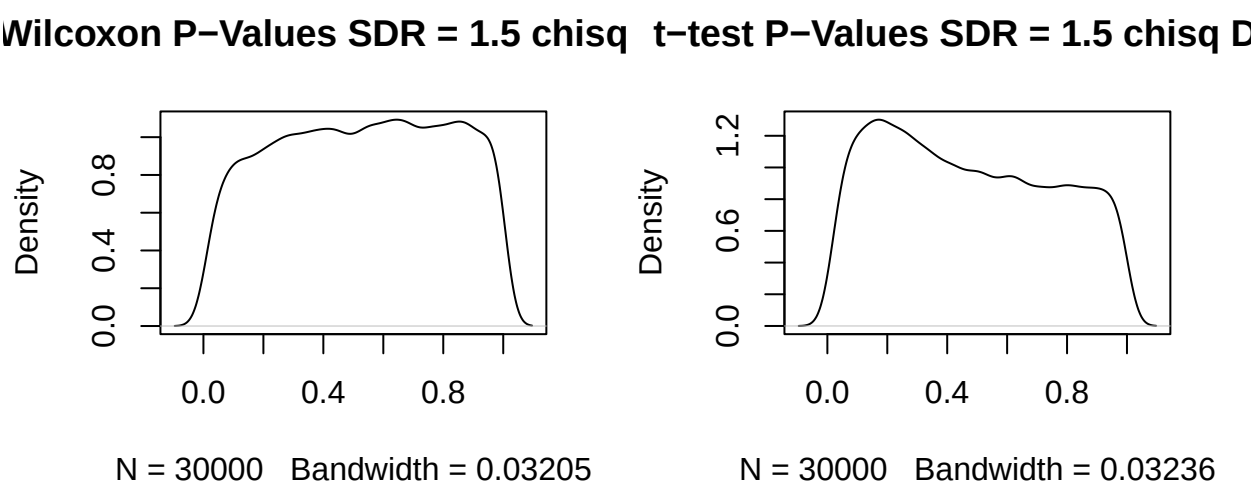

Wilcoxon P−Values SDR = 1.25 lp I
t−test P−Values SDR = 1.25 lp Di
Density
0.0
0.4
0.8
N = 30000 Bandwidth = 0.03226
N = 30000 Bandwidth = 0.03226

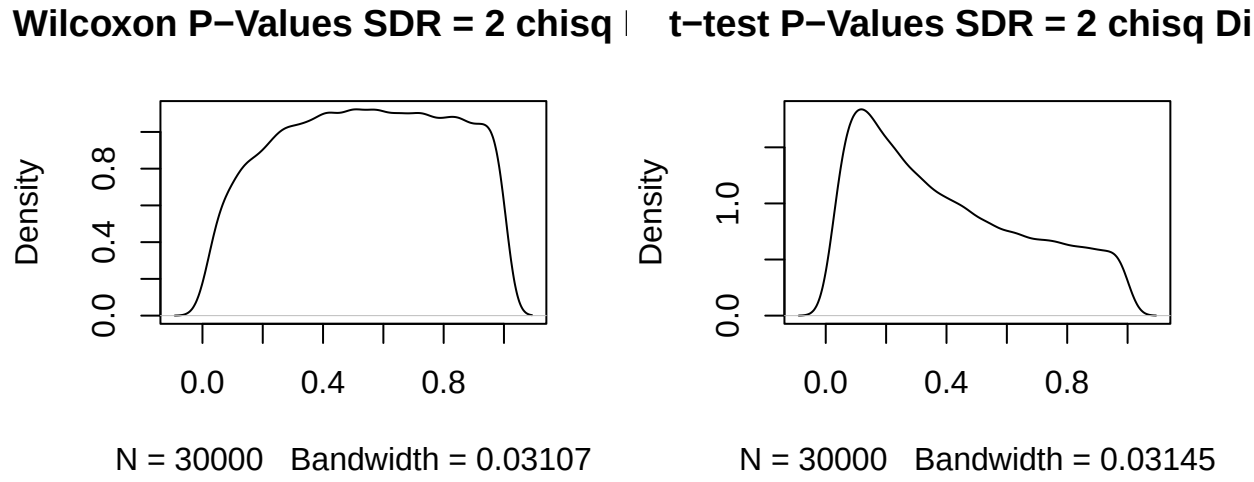

Wilcoxon P−Values SDR = 1.5 lp D
t−test P−Values SDR = 1.5 lp Dis
Density
0.0
0.6
1.2
0.4
0.8
0.0
0.4
0.8
N = 30000 Bandwidth = 0.03127
N = 30000 Bandwidth = 0.03165


Wilcoxon P−Values SDR = 2 lp Di
t−test P−Values SDR = 2 lp Dist
Density
0.0
0.6
1.2
0.4
0.8
0.0
0.4
0.8
N = 30000 Bandwidth = 0.02925
N = 30000 Bandwidth = 0.03059


Wilcoxon P−Values SDR = 1 chisq
t−test P−Values SDR = 1 chisq Di
Density
0.0
0.4
0.8
N = 30000 Bandwidth = 0.03329
N = 30000 Bandwidth = 0.03273


Vilcoxon P−Values SDR = 1.25 chisc
t−test P−Values SDR = 1.25 chisq I
Density
0.0
0.4
0.8
N = 30000 Bandwidth = 0.0325
N = 30000 Bandwidth = 0.03243





From these results, we can see that when the data is symmetric, the Wilcoxon Rank Sum test and Welch's t-test have comparable robustness, as the distribution of p-values is similarly shaped, even in cases where there is unequal variance. It is likely that this is at least partially due the fact that in symmetric data, the mean and median are equal. Additionally, it is conventional wisdom that for sample size greater than 30, if the data is not too irregularly shaped, than the normal assumption can be considered satisfied. In this case we used sample size equal to 30 for both data sets so it is likely that the normal assumption is at least partially true. If that is the case, it can be show that Wilcoxon's Ranked Sum is 95% has efficient as Welch's t-test.

However, when the data is asymmetric then t-test is considerably worse. Even when the variance is equal the distribution of p-values over the simulated tests is not uniform and is actually left-skewed. As a result in this case, actual alpha is higher than observed alpha on the probability of a type 1 error (false rejection of null hypothesis) is higher.

Due to this result, we will not consider the t test in the possible testing procedure, as Wilcoxon has comparable or significantly better performance in all cases.

Part 5 - Comparing Power of Testing Procedures

There are 3 possible options for testing procedures: Performing only a Wilcoxon Rank Sum Test; Performing an Ansari-Bradley Test of Dispersion then a Wilcoxon Rank Sum Test (if the AB test is passed); or performing a Lepage Test. Similar to in Part 3, we can calculate a table of simulated power values for of the procedures. In the case of performing 2 tests, the overall power of the procedure is the product of the power of each test performed in the procedure.

Table of Power Values when True Shift in Center = 0.5

Distribution and SDR	W Power	W & AB Power	Lepage Power
norm SDR=1	0.4523000	0.4746462	0.3523000
norm SDR=1.25	0.3542000	0.4370777	0.3440667
norm SDR=1.5	0.2613667	0.5095228	0.4534333
norm SDR=2	0.1261333	0.7774844	0.7465667
unif SDR=1	0.4280667	0.4498764	0.3276000
unif SDR=1.25	0.3025333	0.4287283	0.3494333
unif SDR=1.5	0.1785667	0.5974155	0.5513000
unif SDR=2	0.0612000	0.8983280	0.8683333
lp SDR=1	0.3576667	0.3903828	0.2841000
lp SDR=1.25	0.2789000	0.3596151	0.2648333
lp SDR=1.5	0.2137000	0.4103536	0.3337667
lp SDR=2	0.1195333	0.6303507	0.5663000
chisq SDR=1	0.7703333	0.9562179	0.9635667
chisq SDR=1.25	0.7467000	0.8879570	0.8830667
chisq SDR=1.5	0.6792333	0.7821032	0.7204000
chisq SDR=2	0.3541667	0.5268410	0.4612333
exp SDR=1	0.7611667	0.8767222	0.8816667
exp SDR=1.25	0.6940667	0.7497057	0.6836667
exp SDR=1.5	0.5561667	0.5871166	0.4533333
exp SDR=2	0.2178000	0.4905792	0.4375000
lnorm SDR=1	0.1326000	0.2488027	0.1736667
lnorm SDR=1.25	0.0515333	0.1167562	0.0647000
lnorm SDR=1.5	0.0182667	0.2446544	0.1526667
lnorm SDR=2	0.0040000	0.6451916	0.5017667

Table of Power Values when True Shift in Center = 0.5

Distribution and SDR	W Power	W & AB Power	Lepage Power
norm SDR=1	0.9574667	0.9582932	0.9168333
norm SDR=1.25	0.9278333	0.9340252	0.8906333
norm SDR=1.5	0.8747333	0.9080125	0.8992333
norm SDR=2	0.7029667	0.9058899	0.9519000
unif SDR=1	0.9359333	0.9371442	0.8839667
unif SDR=1.25	0.8997000	0.9087939	0.8561000
unif SDR=1.5	0.8160333	0.8782079	0.8720000
unif SDR=2	0.4920667	0.9091815	0.9515667
lp SDR=1	0.8674667	0.8743761	0.7971000
lp SDR=1.25	0.8068333	0.8276760	0.7508333
lp SDR=1.5	0.7319667	0.7954280	0.7546667
lp SDR=2	0.5748667	0.8125162	0.8360667
chisq SDR=1	0.9845000	0.9980723	0.9998333
chisq SDR=1.25	0.9852667	0.9967110	0.9992333
chisq SDR=1.5	0.9855667	0.9948059	0.9964667
chisq SDR=2	0.9728000	0.9824379	0.9695000
exp SDR=1	0.9950000	0.9980350	0.9996333
exp SDR=1.25	0.9948667	0.9969197	0.9983000
exp SDR=1.5	0.9935000	0.9949025	0.9924333
exp SDR=2	0.9739667	0.9757803	0.9482667
lnorm SDR=1	0.3696000	0.5776950	0.5252667
lnorm SDR=1.25	0.2283667	0.2943928	0.2015333
lnorm SDR=1.5	0.1042667	0.1999608	0.1329333
lnorm SDR=2	0.0210667	0.4916399	0.3756000

From this table we can observe a few things. First, as expected, in the case were the equal variance assumption is satisfied and SDR=1, then performing only the Wilcoxon test has the most power. Testing for variance when it is unnecessary reduces the power of the procedure and its overall significance. Second, we can observe that the power of Lepage's test is the most consistent when the variance is not truly equal. This also is a a result that makes sense because Lepage's test simultaneously considers differences in center and differences in variance.This is also an explanation for the robustness of Lepage's test throughout all scenario simulated under. Additionally, note that that the power of the Wilcoxon test becomes very small when the SDR is high, as explained in Part 3, so the overall power of a both procedures that use Wilcoxon's test is lower when SDR does not equal 1.

Another important thing to note is that the Wilcoxon test generally works as well as the other two methods if the equal variance assumption is mostly satisfied. However, if the equal variance assumption is violated, then in terms of test power it is better to perform the Ansari-Bradley and Wilcoxon tests or just Lepage's test. The method that performs better out of those two varies depending on the underlying shapes of the distributions, and this variation is very slight. This is where it is important to consider the goal of testing. If the only important results is whether two distributions are equal, then Lepage's test is sufficient. However, if we want to determine equal variance and then equal center independently, then performing Ansari-Bradley then Wilcoxon is required in this setting.
If we examine patterns for each level of SDR, separate from other observations about the data, then another pattern becomes clear. When SDR=1.25, the power of only performing the Wilcoxon test is very close to the power of performing Lepage's test. Hence we can reasonably conclude that if we have slight deviations from the equal variance assumption, where the ratio of standard deviations between 2 data sets is 1.25 or less, then the best procedure is to perform Wilcoxon alone.

Conclusion

There are number of results that can be taken from this simulation study, although none of them are particularly surprising. First, we can note that depending on the symmetry of the underlying distribuutions, the behavior of the Wilcoxon rank sum test changes. Specifically, if the underlying distribution of the data sets is symmetric, then the test yields p-values that are left-skewed, and if the underlying distribution of the data sets is asymmetric, then the test yields p-values that are right-skewed. If p-values are left-skewed it means that the type 1 error rate has decreased (actual alpha is less than observed value) whereas if p-values are right-skewed the type 1 error rate has increased (actual alpha is greater than observed value)l. Second, we can note the power of the Wilcoxon test (and t-test) decreases as the true standard deviation ratio SDR deviates from 1, indicating a violation of the equal variance assumption. Third, the Wilcoxon test is slightly robust to deviations from the equal variance asssumption, because the distribution of p-values holds its shape for the most part when SDR=1.25, and it is more robust to deviations from equal variance than Welch's t-test. Despite this however, as an overall procedure, Lepage's test is better than Wilcoxon in almost all cases, except when equal variance assumption can be met. This also true because Lepage's test offers a way to compare the overall equality of 2 datasets by looking at center and variance. Conversely, Wilcoxon compares 2 datasets only by their measure of center. Further simulation anad analysis could refine these results.

If a simulation study of this nature were to be repeated then it would be important to peform a larger number of simulations if possible. Also, it would be better to use a more methodical approach to deciding which distributions to sample from in the simulation step, that way a wider variety of distributions could be used beyond the named distribution that mostly fall into a couple categories (exponential, bell-shaped, etc.). Finally, testing a larger number of possible shift in centers and SDRs would lead to a more precise understanding of the behavior the test under a violation of the equal variance. From this study however, there are still a decent number of observations, mostly confirming behaviors of the Wilcoxon and t-test that follow the intuition behind those tests.