

Choosing Between Parametric or Non-parametric Tests

Abstract: A common question in comparing two sets of measurements is whether to use a parametric testing procedure or a non-parametric procedure. The question is even more important in dealing with smaller samples. Here, using simulation, several parametric and non-parametric tests, such as, t-test, Normal test, Wilcoxon Rank Sum test, van-der Waerden Score test, and Exponential Score test are compared.

Introduction

Let us consider two independent random samples x_1, x_2, \dots, x_m and y_1, y_2, \dots, y_n are taken from two populations. To compare the two samples, a common practice is to compare their means, in other words testing the statistical hypothesis:

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2$$

Where H_0 indicates the null hypothesis, H_1 indicates the alternative hypothesis, μ_1 indicates the first population mean, and μ_2 indicates the second population mean.

The statistical tests of hypotheses are based on the fundamental that if the samples have significant evidence against the null hypothesis (H_0), then H_0 is rejected in favor of the alternative hypothesis (H_1). Then the question is how significant is significant, when do we say there is enough evidence, the answer is based on the idea of Type I error, the probability of rejecting H_0 when in fact it is true. The power of the test is determined by the rate of rejection of H_0 when it should be rejected. In other words, how well our test sees that $H_0 \neq H_1$.

p-value

The observed level of significance (or the Type I error) of a test is known as the *p*-value of the test. This is the probability of rejecting H_0 when it is in fact true. In our study we use a 5% level of significance. This however, is just one of the many common levels of significance commonly used.

Parametric Tests

[Type text]

1. According to Reinard (2006), when the two population distributions are normal, the population variances σ_1^2 and σ_2^2 are unknown and unequal, the test statistic is

$$T = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \text{ where } \bar{x} = \frac{\sum_{i=1}^m x_i}{m}, \bar{y} = \frac{\sum_{i=1}^n y_i}{n}, s_1^2 = \frac{\sum_{i=1}^m (x_i - \bar{x})^2}{m-1}, s_2^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}, \text{ and } T$$

has a t -distribution with degrees of freedom $df = \frac{(A+B)^2}{\frac{A^2}{m-1} + \frac{B^2}{n-1}}$, where $A = \frac{s_1^2}{m}$ and $B = \frac{s_2^2}{n}$.

2. According to Tanis and Hogg (2008), when the two population distributions are normal, the population variances σ_1^2 and σ_2^2 are unknown but equal, the test statistic is:

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}}, \text{ where } s_p = \sqrt{\frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}} \text{ and } T \text{ has a } t\text{-distribution with}$$

$m+n-2$ degrees of freedom.

3. According to Tanis and Hogg (2008), when the two population distributions are not assumed as normal, the population variances σ_1^2 and σ_2^2 are unknown, and the sample sizes n_1 and

n_2 are large, the test statistic is: $Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}$, where Z is the standard normal variate.

Note that $\mu_1 - \mu_2 = 0$ for all three cases above, as per the null hypothesis. But in general it is not necessarily zero as if we want to test that one mean is at least an amount higher than the other then $\mu_1 - \mu_2$ is that least amount, and so on. The cases for known variances are not considered as they are not common in practice.

In this paper we will consider the first test and the third test and denote as TD and ZD , respectively. We also will consider TP when the tests are computed as in TD and ZD but the p -value is computed by considering all permutations of the data. For larger samples, TP uses random permutations instead of all possible permutations. The corresponding p -values are denoted as PTD , PZD , and PTP for the t -test, normal test, and the respective permutation test, respectively.

[Type text]

Non-parametric Tests

Wilcoxon Rank-Sum Test

In Higgins (2004) the method to perform the Wilcoxon rank-sum test is computed as follows. Let m be the sample size of the one group or treatment, and n be the sample size of another. Combine $m+n$ observations into one group, and rank the observations from smallest to largest. Let 1 be the rank of the smallest observation, 2 the rank of the next smallest observation, and so on. It is common to have ties among observations in a data set; that is, one or more observations may have the same value. In this case, the assignment of ranks to the observations is ambiguous. To resolve this ambiguity, the average rank is assigned to the tied observations. Find the observed rank sum W of treatment 1 (Note we may analyze either treatment 1 or treatment 2 due to the equivalency of the statements $\mu_1 = \mu_2$ and $\mu_2 = \mu_1$). Then the p -value of the test is computed either by using the distribution of all possible permutations of the ranks or by using normal approximation for larger samples. For the two sided test considered here

$$WR = \text{maximum}(R - W, W),$$

where R is the sum of the ranks for the combined sample.

Permutation Distribution

In Higgins (2004) the method to perform the permutation distribution test follows. Find all possible permutations of the ranks in which m ranks are assigned to treatment 1 and n ranks are assigned to treatment 2.

For each permutation of the ranks, find the sum of the ranks for treatment 1 (or treatment 2).

Determine the two sided p -value as

$$PWR = \frac{\text{number of maximum}(R - U, U) \geq WR}{\binom{m+n}{m}},$$

where U is the sum of the ranks for treatment 1 (or treatment 2) for a permutation.

When the sample sizes are so large that all permutations cannot be performed within a reasonable time period, random permutations for a reasonable number (10,000 or 100,000) of times can be performed depending on time and computational facility.

Large Sample Approximation

[Type text]

According to Higgins (2004), for larger samples with sample size 10 or greater, such permutations can be considered large,

$$Z = \frac{W - E(W)}{\sqrt{V(W)}}$$

follows approximate standard normal distribution and hence can be used to obtain an

approximate p -value. Where $E(W) = m\mu$, $V(W) = \frac{mn\sigma^2}{m+n-1}$, μ is the mean for all ranks for the

combined sample irrespective of whether there is any ties, and σ^2 is the population variance for all ranks for the combined sample irrespective of whether there is any ties. Without ties,

$\mu = \frac{m+n+1}{2}$ and $\sigma^2 = \frac{(m+n-1)(m+n+1)}{12}$. Let the large sample approximate p -value for the

Wilcoxon Rank Sum test be denoted as PWZ .

van der Waerden Score Test

The process of this test is exactly similar to the Wilcoxon Rank Sum test where the ranks are replaced by the van der Waerden scores. In Higgins (2004) the van der Waerden scores are defined by

$$V_{(i)} = \Phi^{-1}\left(\frac{i}{m+n+1}\right)$$

where Φ^{-1} denotes the inverse of the cdf of the standard normal distribution. The test statistic is the sum of the van der Waerden scores for treatment 1 (or treatment 2). Then the p -value is computed using the methods as described for the Wilcoxon Rank Sum test by using the van der Waerden scores instead of the ranks. Let the permutation p -value for the van der Waerden score test be denoted as PVS and the large sample approximate p -value for the van der Waerden score test be denoted as PVZ .

Exponential Score Test

The process of this test is exactly similar to the Wilcoxon Rank Sum test where the ranks are replaced by the Exponential scores. The Exponential scores are defined by

$$\frac{1}{m+n}, \frac{1}{m+n} + \frac{1}{m+n-1}, \frac{1}{m+n} + \frac{1}{m+n-1} + \frac{1}{m+n-2}, \dots$$

in Higgins (2004). The test statistic is the sum of the Exponential scores for treatment 1 (or treatment 2). Then the p -value is computed using the methods as described for the Wilcoxon Rank Sum test by using the Exponential scores instead of the ranks. Let the permutation p -value

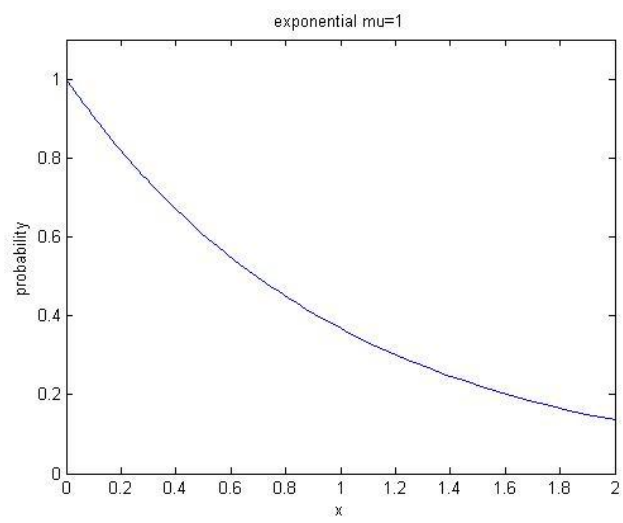
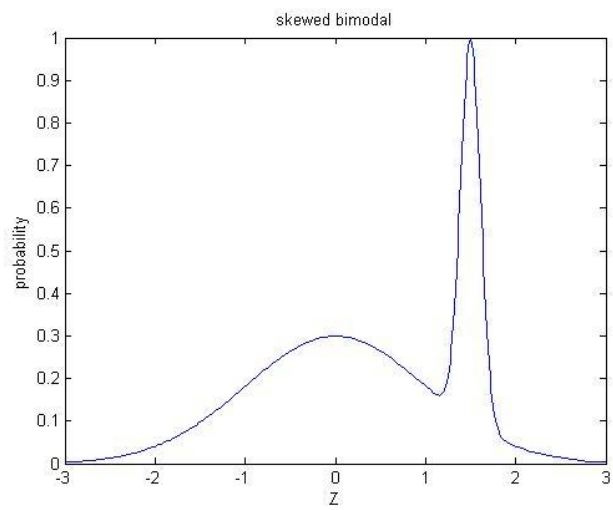
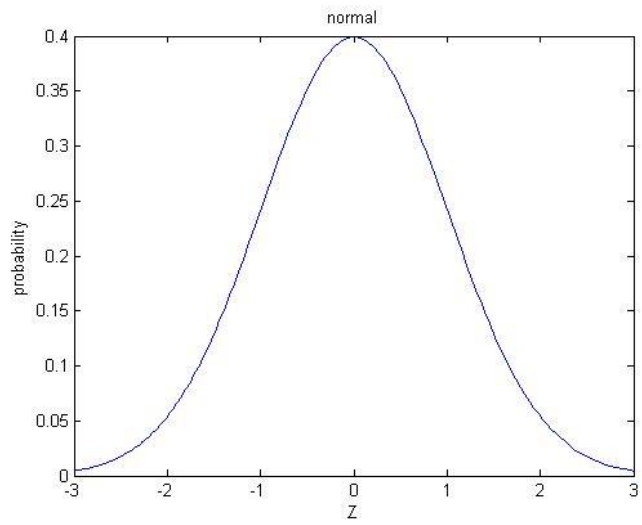
[Type text]

for the van Exponential score test be denoted as *PES* and the large sample approximate *p*-value for the van der Waerden score test be denoted as *PEZ* .

There are certain parameters under which parametric methods have been suggested to be superior to nonparametric methods. Similarly, there are instances where nonparametric methods are suggested over their parametric counterparts. According to Warner (2007), nonparametric methods should be used when the sample size is small, whereas parametric methods should be used when the sample size is large. Also when there is an outlier in the data, nonparametric methods are said to be preferable. According to Tanis and Hogg (2008), when the population distribution is normal and the sample size *n* is as small as 4 or 5 the normal test should a very adequate approximation.

I also tested some parameters not considered or addressed by statisticians to see if they suggest one method or the other. One of the parameters that will be tested is if different distributions have any effect on the performance of the two methods. The following three graphs illustrate the different distributions used. Different variances are also adjusted to see if any effects make themselves apparent. The distance between means is also changed, to see if the methods equivalently pick up on the more severe difference.

[Type text]



[Type text]

Figure A) Distribution Examples

Simulation Study

To investigate how the tests are related to the estimates of the Type I error, 1000 samples of sizes 5, 8, 11, and 15 are selected from independent normal populations with different means and variances. All nine p -values mentioned above (PTD , PZD , PTP , PWR , PWZ , PVS , PVZ , PES , and PEZ) are computed and the numbers of p -values less than or equal to 0.05 are recorded. The choices are: (i) Population 1: Normal with mean 1 and variance 1; Population 2: Normal with mean 1 and variance 1, (ii) Population 1: Normal with mean 1 and variance 1; Population 2: Normal with mean 1 and variance 1 with an outlier. The proportions of rejections are displayed in Table 1. The values displayed in Table 1 represent the rate at which the tests said the means were different when in fact they were the same. Each of the tests was performed on these two different distribution comparisons for the sample sizes 5, 8, 11, and 15.

Table 1: Estimates of the Level of Significance

n	PTD	PZD	PTP	PWR	PWZ	PVS	PVZ	PES	PEZ
				$N(1,1)$	$N(1,1)$				
5	0.053	0.089	0.056	0.037	0.066	0.040	0.066	0.060	0.037
8	0.054	0.072	0.057	0.054	0.054	0.054	0.054	0.060	0.049
11	0.042	0.065	0.043	0.046	0.046	0.046	0.046	0.054	0.047
15	0.041	0.051	0.041	0.035	0.036	0.035	0.036	0.050	0.041
				$N(1,1)$	$N(1,1)$	w/outlier			
5	0.013	0.036	0.051	0.030	0.063	0.031	0.063	0.057	0.030
8	0.004	0.018	0.032	0.030	0.030	0.030	0.030	0.029	0.020
11	0.014	0.019	0.036	0.039	0.039	0.039	0.039	0.039	0.034
15	0.019	0.029	0.043	0.041	0.041	0.041	0.041	0.048	0.041

To investigate the powers of the tests, samples are generated from the populations having different means. The choices are: (i) Population 1: Normal with mean 1 and variance 1; Population 2: Normal with mean 3 and variance 1, (ii) Population 1: Normal with mean 1 and variance 1; Population 2: Normal with mean 5 and variance 2, (iii) Population 3: Normal with mean 1 and variance 1; Population 2: Normal with mean 2 and variance 1, (iv) Population 1: Exponential with mean 1/3; Population 2: Normal with mean 1 and variance 1, (v) Population 1: Exponential with mean 1/3; Population 2: Exponential with mean 1, (vi) Population 1: Skewed bimodal with mean 3/8 and variance 7/9; Population 2: Normal with mean 0 and variance 1, (vii) Population 1: Skewed bimodal with mean 3/8 and variance 7/9; Population 2: with mean 3 and variance 1. Then for each of the choices proportion of rejections are computed and displayed in Table 2. The values displayed in Table 2 represent the rate at which the tests said the means were

[Type text]

We now analyze the various scenarios and compare the effectiveness of the parametric and non-parametric tests. We will compare populations which share different distributions, populations that have different respective distributions, populations with different variances, different populations, populations with different means, and treatments with extreme outliers. We will observe how quickly the tests are picking up on the fact that $H_0 : \mu_1 = \mu_2$ when it is the case.

We begin with two populations each having a normal distribution. One of the samples has a mean of 1 and a variance of 1. The other has a mean of 1 and a variance of 1. Since the means are equal we are computing the level of significance of the tests. We can see from Table 1 that *PZD* or the normal test had slightly higher levels of significance for all four of the populations sizes. However this difference was not significant. The decision made of rejecting or accepting H_0 depends entirely on your desired level of significance. No test drastically stood out such that a majority of commonly used levels of significance would result in different test yielding different results. All of the tests picked roughly 5% for a level of significance except *PZD* when $n=5$, however, even that was off by less than 4%. Additionally the large sample approximation of the exponential scores test or *PEZ* picked a low level of significance when the sample size $n=5$. The data discussed is plotted in the following graph (Figure 1).

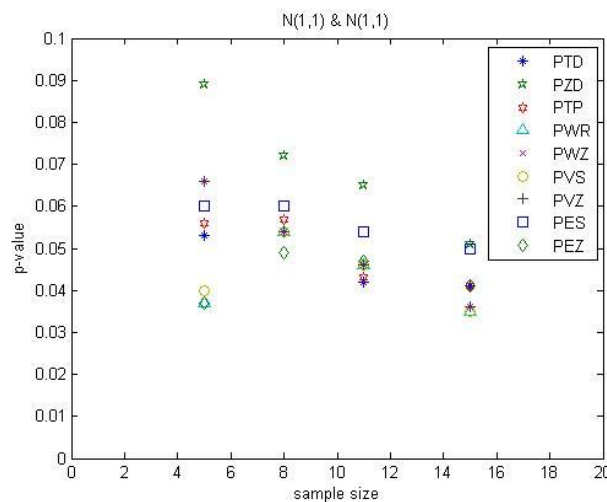


Figure 1: Type 1 Error; $N(1,1)$ vs $N(1,1)$

Now we observe the results of similarly constructed populations with the addition of outliers. Again, since the means are equal we compute the levels of significance. It is apparent from the data displayed in Figure 2 that the scores were on average lower than in Figure 1, this means that the tests were, on average, more effective in determining that H_0 is true. When the sample size $n=5$, *PWZ*, *PVS*, and *PES*, all picked values greater than 5%, while the rest picked lower values. When the sample size was greater, however, all the tests performed similarly

[Type text]

picking value lower than 5%. While the observed levels of significance are somewhat greater for the nonparametric methods, they still generally resulted in the same conclusion of rejecting $H_0 : \mu_1 = \mu_2$. The data discussed is plotted in the Figure 2.

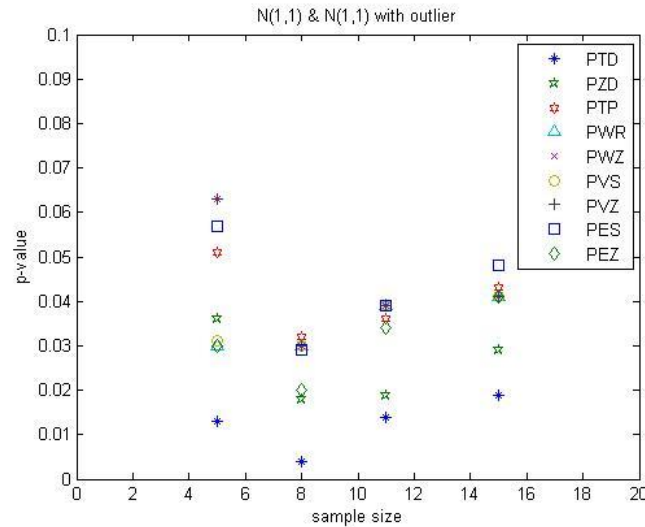


Figure 2: Type 1 Error; $N(1,1)$ vs $N(1,1)$ with outlier

When finding the levels of significance in both cases the methods did not differ too greatly. While in certain circumstances some tests had a p -value greater than 5%, the tests that had a p -value less than 5% were not far below this level of significance. When considering the differences between the tests we observed that on average the difference between the parametric and nonparametric methods was rather small. Since there was not a great deal of difference in the performance of the tests when considering the different styles of distributions and the sample sizes, there was no single method of test, parametric or nonparametric, that clearly performed better than the rest. We shall soon see that, when we dive into observing the power of the tests, the similarities become even more apparent

We now consider how effective the tests were in determining when $H_0 : \mu_1 = \mu_2$ is not true. This first simulation compares two normal populations each having a variance of 1, and means of 1 and 3, respectively. When the sample size $n=5$, PZD had a slightly greater power than the rest, while the other tests performed very similarly when testing the power. When the sample size increased there was very little difference between any of the test's performance. Since there was no significant difference between any of the tests for all four of the sample sizes, the test performed equally. The data discussed is plotted in the following graph.

[Type text]

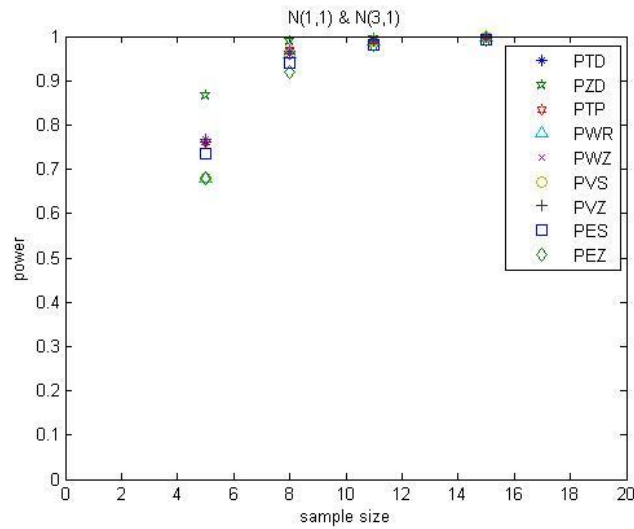


Figure 3: Type 2 Error; simulation 1, $N(1,1)$ vs $N(3,1)$

In the second simulation we analyze two normal populations, population 1 with a mean of 5 and variance 1, and population 2 with mean 5 and variance 2. Each of the tests picked up on this increased difference in means rather effectively. As the sample size increases this becomes even more apparent. This is especially true when the sample size $n=15$. In this case all of the tests had identical values.

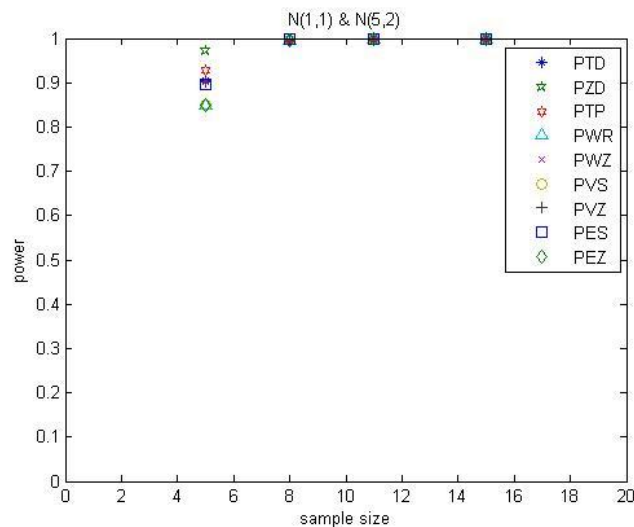


Figure 4: Type 2 Error; simulation 2, $N(1,1)$ vs $N(5,2)$

For the third simulation we analyze two normal populations each having a variance of 1, and means of 1 and 2 respectively. The normal test of PZD picked a slightly higher value for the two lesser of the four sample sizes. The other test performed similar to each other for each of the

[Type text]

sample sizes. When the sample size was greater, *PZD* was performing closely to the other eight tests.

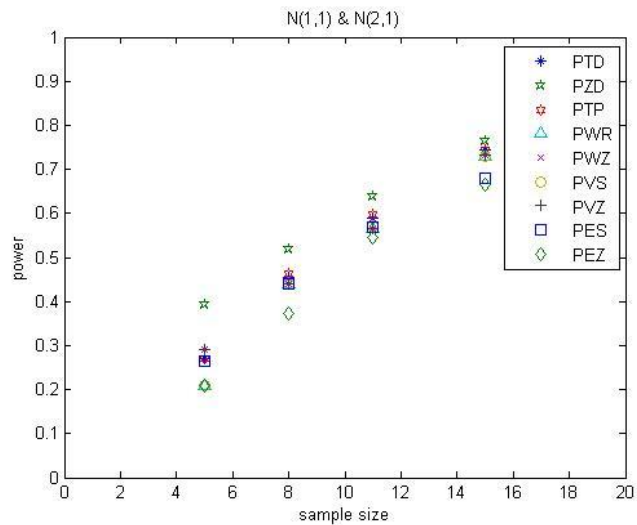


Figure 5: Type 2 Error; simulation 3, $N(1,1)$ vs $N(2,1)$

In the fourth simulation we change the distribution of one of our samples to exponential and give it a mean of $1/3$, the second population has normal distribution with a mean of 1 and variance 1. Similarly to the previous scenarios, the tests gave approximately the same result for all the sample sizes, with the differences between the tests decreasing as the sample size increased.

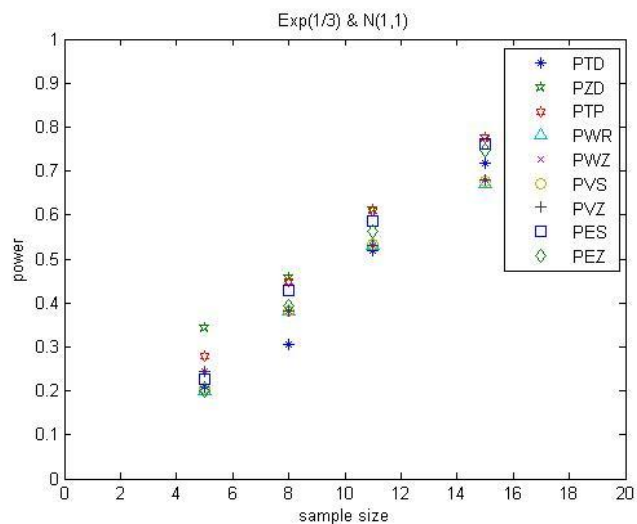


Figure 6: Type 2 Error; simulation 4, $Exp(1/3)$ vs $N(1,1)$

[Type text]

The fifth simulation compares two exponential distributions with means $1/3$ and 1 , respectively. In a slight change of pace, none of the tests stood out either above or below for any of the sample sizes in determining when $H_0 : \mu_1 = \mu_2$ is false. When the sample size $n=5$ the tests all have values close to 20%-25%. Each of the tests had almost identical values for higher three sample sizes.

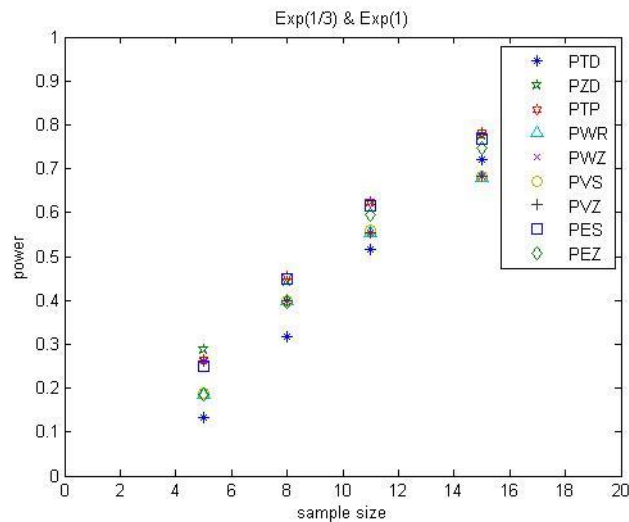


Figure 7: Type 2 Error; simulation 5, $Exp(1/3)$ vs $Exp(1)$

In the sixth and seventh simulations we compared skewed bimodal distributions with normal distributions. In both of the trials the skewed bimodal distribution had a mean of $3/8$ and variance of $7/9$, while the normal distributions had means 0 and 3 respectively, and in both cases variance of 1 . In the sixth simulation for all four of the sample sizes the tests all performed similarly, picking values approximately 8% apart or less. They also stayed below 20% in all of the cases. In the seventh trial however, the tests all had values 85% or high, while still maintaining a maximum difference of 10%.

[Type text]

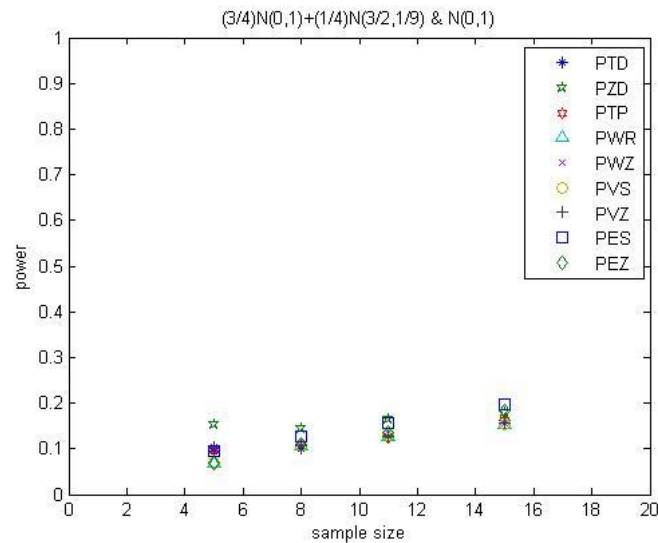


Figure 8: Type 2 Error; simulation 6, $(3/4)N(0,1)+(1/4)N((3/2),(1/9))$ vs $N(0,1)$

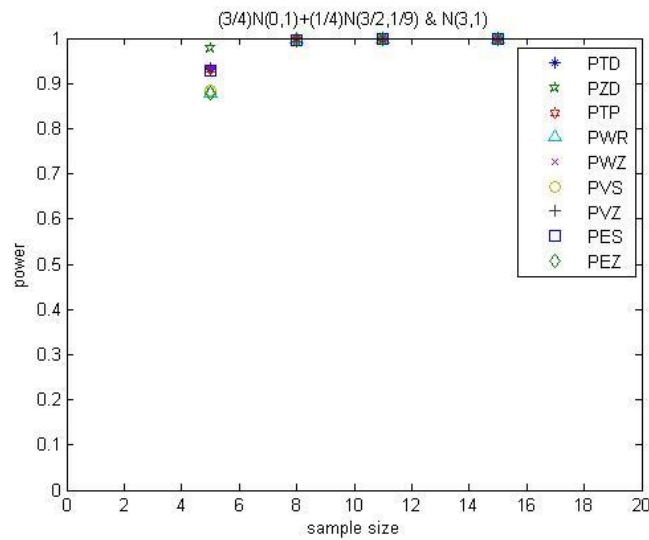


Figure 9: Type 2 Error; simulation 7, $(3/4)N(0,1)+(1/4)N((3/2),(1/9))$ vs $N(3,1)$

While there were instances where one of the tests had a slightly higher or lower value for a certain set of parameters, when there was a difference it was not large enough to be considered significant. In finding both the power and the level of significance, none of the tests truly “outperformed” the others for any particular set of parameters. When finding the observed level of significance, the nonparametric tests did prove to be consistently more effective than the parametric tests. However, this difference in effectiveness or performance was not enough to influence the decision of whether or not to reject $H_0 : \mu_1 = \mu_2$. Consequently, when we consider

[Type text]

the set of parametric test against the set of nonparametric tests we did not observe that one set or the other had a significantly higher power or more accurately picked the level of significance.

Contrary to accepted set of criteria for determining which to use, our research did not find a specific set of parameters for which parametric tests are the proper choice over nonparametric. A small sample size had a small effect on the performance of the tests, however when the size increased, the tests performed almost equivalently. This is the opposite of what the accepted notion of the performance of the parametric methods versus nonparametric methods. Changing the variance also seemed to have no effect. When the difference between the means was greater, both sets of tests, parametric and nonparametric, picked up on this difference similarly. Even when comparing different distributions types, the tests performed relatively similar to each other.

Since there was no clear scenario when parametric methods outperformed nonparametric methods or visa versa, the research was inconclusive. None of the tested parameters had an effect significant enough to cause noticeable change in the outcome. Thus, the choice of parametric or nonparametric seems to be left to the preference of the person analyzing the population data.

Bibliography

Higgins, Jams J. Introduction to Modern Nonparametric Statistics. Pacific Grove, CA: Brooks/Cole-Thompson, 2004.

Hogg, Robert V., and Tanis, Elliot A. A Brief Course in Mathematical Statistics. Upper Saddle River, NJ: Pearson Prentice Hall, 2008.

Reinard, John C. Communication Research Statistics. London, UK: Sage Publications, 2006.

Warner, Rebecca M. Applied Statistics: From Bivariate Through Multivariate Techniques. London, UK: Sage Publications, 2007.