Permutation Tests for Comparing Two Populations

Ferry Butar Butar, Ph.D. †
Jae-Wan Park ‡

Abstract

Permutation tests for comparing two populations could be widely used in practice because of flexibility of the test statistic and minimal assumptions. The Wilcoxon sum rank test is more powerful than a t test statistic for moderate and large sample sizes for heavier tailed distributions. Using a Resampling Stats, this test is easy to implement and a significance level is exact when calculating all possible permutations. The approximate significance level can be used when the numbers of permutations are very large.

Introduction

Suppose a researcher wants to know whether a new experimental drug relieves symptoms attributable to the common cold. The response variable may be the time until the cold symptoms go away. If we let μ_1 be the mean time until cold symptoms go away for individuals who take the drug and μ_2 be the mean time until symptoms go away for individuals who take placebo, then the hypothesis could be H_0: $\mu_1 = \mu_2$ and the alternatives could be H_a: $\mu_1 < \mu_2$ or H_a: $\mu_1 > \mu_2$ or H_a: $\mu_1 \neq \mu_2$. The first alternative means that a drug is effective since the mean time until cold symptoms go away is less than for individuals who take the drug than for those who do not take the drug. In parametric setting, there are several assumptions for this test to be valid. First, the two samples come from populations with normal density. Second, the samples must be independent. If both population variances are known then the test statistics is given by

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}.$$

When population variances are unknown but the sample sizes for each population are greater than 30 then one usually uses *Welch's approximate t* which is a *t* test statistics that can be calculated similar to Z test statistics above by replacing S_1^2 and S_2^2 for σ_1^2 and σ_2^2. Another assumption is when both variances are unknown but they are equal. Since the variances are equal and we wish to test both population means are equal, then the natural way to estimate the variance is to combine the sample, called *pooled variance*. The pooled variance is computed by finding a weighted average of the samples variances, i.e, $S^2_p=(n+m-2)^{-1}$ $((n-1)S_1^2+(m-1)S_2^2)$. And the test statistic is just replaced Z with t and σ_1^2 and σ_2^2 with S^2_p. In practice we do not know whether the variances are actually equal, thus usually we have to check this by testing equality of the variances using *F* test statistic. We can imagine that there is no guarantee that all assumptions above are satisfied. If these assumptions are not

valid then we can still test similar hypothesis with nonparametric test. In the literature, there are many different methods for testing equality of two populations. In this paper, we will use a permutation method.

Hypothesis Testing for Equal Treatment Effect

In a nonparametric statistic, there is no parameter to be tested. Let $F_1(x)$ be the cumulative distribution function (cdf) of population 1 and $F_2(x)$ be the cdf of population 2. Then the null hypothesis is H_0: $F_1(x) = F_2(x)$. In this case the two distributions are identical under null hypothesis. Here it does not say the means are equal but the variances are different. When the two treatments are the same under the null hypothesis, meaning that the distribution of the observations is the same. The alternative hypothesis is given by H_a: $F_1(x) \leq F_2(x)$ with at least one x for strict inequality. The statement above is the same as H_a: $\mu_1 > \mu_2$ in parametric case, since observations for treatment 1 tend to be larger than observations for treatment 2. We can also have the alternative hypothesis H_a: $F_1(x) \geq F_2(x)$. For two sided alternative the alternative hypothesis is H_a: $F_1(x) \leq F_2(x)$ or, $F_1(x) \geq F_2(x)$ with strict inequality for at least one x.

Permutation Test

Permutation tests also known as randomization tests. It is widely used in nonparametric statistics where a parametric form of the underlying distribution is not specified. Consider sample of m observations from treatment 1 and n observations from treatment 2. Assume that under the null hypothesis there is no difference between the effect of treatment 1 and treatment 2. Then any permutation of the observations between the two treatments has the same chance to occur as any other permutation. The steps for a two-treatment permutation test:

- Compute the difference between the mean of observed data, called it D_{obs}.
- Create a vector of $m+n$ observations.
- Select at random experimental units to one of the two treatments with m units assigned to treatment 1 and n units assigned to treatment 2.
- Permute the $m+n$ observations between the two treatments so that there are m observations for treatment 1 and n observations for treatment 2. The number of possibilities are $\binom{m+n}{n} = \frac{(m+n)!}{m!n!}$.
- For small sample sizes, obtain all possible permutations of the observations; for large sample sizes, obtain a random sample of predetermined, R, permutations.
- For each permutation of the data, calculate the difference between the mean of treatment 1 and mean of treatment 2, called it, D.

- For the upper tailed test, compute *p-value* as the proportion of *D* greater than or equal D_{obs} i.e., $P-value=\frac{no.of\ D's \geq D_{obs}}{\binom{m+n}{n}}$
- If p-value less than or equal to the predetermined level of significance α then we reject H_0.

This permutation test is very flexible. One can choose a test statistic suited to the task at hand. Instead of using the difference between means of treatment 1 and mean of treatment 2 as a test statistic, one can also use the sum of either treatment 1 or 2. If there is an outlier in the data set one can use difference between median as a statistic instead of the difference between mean. One may also use trimmed mean as a statistic by deleting equal numbers of the smallest observations and the largest observations. (Higgins, 2004).

How to decide which one of the mean, median, or trimmed mean to employ? It depends on the knowledge about the population from which data come from. In practice, one can use the difference between means when the data is approximately normal density; use the difference between medians, if the distributions of observations are asymmetric; and if the distribution is symmetric with some unusual large and small observations, then one can utilizes a trimmed mean as a statistic. These nonparametric methods do not require analytical derivation of test statistic under the null hypothesis. Again there is a relaxation in choosing the test statistic. With this relaxation, this permutation test has advantageous over a parametric test. Permutation tests can be applied to continuous, ordinal, or categorical data, to values of normal or non-normal density. Whenever a parametric test works, a permutation test also works. These permutation methods have wide range of applications. Permutation methods can be applied whenever parametric statistical methods fail (Good, 1994).

Now instead of permuting the original observations, one can permute the ranks of the observations. Let $X_1,X_2,...,X_N$ $(N=n+m)$ be the combined observations. The rank of X_i among the *N* observations, $R(X_i)$, is $R(X_i)$ = number of $X_j's \leq X_i$. If no two observations have the same value then let 1 be the rank of the smallest observation, 2 be the rank of the next smallest observation, and so on. For example, let the observations to be: 6 5 8 9 10; and their ranks are: 2 1 3 4 5. In case of tie observations, one can adjust the rank. For example the observations 3,3,4,5,5,5,5. Their ranks are 1.5, 1.5, 3, 5.5, 5.5, 5.5, 5.5. The first two ranks are computed by averaging rank of 1 and 2. If one uses the sum of ranks for one treatment the test statistic is called the Wilcoxon rank-sum test, *W*. To calculate a permutation test, combine the $m+n$ ranks. Permute the ranks among the two treatments in which *m* ranks are assigned to treatment 1 and *n* ranks are assigned to treatment 2. For a small sample sizes *m* and *n*, obtain all possible permutations of the ranks, for large sample size obtain a random sample of *R* permutations. For each permutation of the ranks, compute the sum of the ranks, *W*, for one treatment. The p-value is a fraction of the sum of the ranks for one treatment, *W*, greater than or equal to sum of the observed ranks, W_{obs}.

Properties of the Tests

The only assumption we have is the distribution of treatment 1 under the null hypothesis is the same as the distribution of treatment 2. In practice the fewer assumptions and limitations, the broader the applications. The question is how powerful is this test? Is this test more powerful than the parametric counterpart? How robust is the test? That is if there is a violation of the underlying distribution how sensitive is it? What are the effects of outliers or extreme values, especially for a small sample sizes (Good, 1994).

In selecting a statistical method, a statistician will pay very close attention to the significance level and the power of the test. The significance level of a test, denoted by α, is the probability of making type I error, that is, the probability of deciding erroneously on the alternative hypothesis when, in fact, the null hypothesis is true. Probability of type II error is the probability of failing to reject the null hypothesis when, in fact, the null hypothesis is false. The power of a test denoted by β, is the complement of probability of type II error, that is, the probability of deciding the alternative hypothesis when the alternative hypothesis is true.

We would like the significance level to be very small or close to zero, and the power of the test to be very big or close to one. In practice we fix significance level, for example, 0.001-.10 and then maximize the power of the test. At the same significance level α, a test is said to be a most powerful test if a test, at specified significance level, is more powerful against a specific alternative than all other tests. A test is said to be a uniformly most powerful test, if a test at specified significance level, is more powerful against all alternatives than all other tests. In practice, it is unusual to know the distribution of the variable(s) or its variance. A test is said to be *exact*, if probability of making type I error is exactly a significance level α. For a test to be exact, a sufficient condition for the combined observations can be permuted is exchangeability. The observations $x_1,x_2,...,x_N$ are exchangeable if, for example, probability of any particular joint outcome, say $x_1+x_5+x_7=x$, is the same regardless of the order of the observations considered. Therefore independent identically distributed, sampling with replacement, and a dependent normal with constant variance and constant covariance are exchangeable.

Example: Consider the illustration mentioned in section 1. Let the alternative hypothesis be the mean time until cold symptoms go away is less than for individuals who take the drug than for those who do not take the drug. The observations for two treatments are as follows:

Table 1
Hours until the symptoms go away

Drug	Dale (36)	Kenneth (60)	Mathar (39
Placebo	Butar (37)	Honjo (55)	Park (70)

The combined samples are 36 39 55 60 70 73. If there is no difference between the two treatments one can expect that the total observations for the treatment 1

is the same as for the treatment 2. If the total observations for the treatment 1 is smaller than for the treatment 2, on the average, then the alternative hypothesis is true since less time for the treatment 1 to recover from the cold.

Table 2
Permutation distribution of Treatment 1 and Treatment 2

Number	Drug (Rank)	Placebo (Rank)	Difference Bet. means	Sum of drug (Sum of Ranks)	Diff bet. medians
1	36 39 55 (1 2 3)	60 70 73 (4 5 6)	-24.33	130	-31 (6)
2*	36 39 60 (1 2 4)	55 70 73 (3 5 6)	-21.00	135	-31 (7)
3	36 39 70 (1 2 5)	55 60 73 (3 4 6)	-14.33	145	-21 (8)
4	36 39 73 (1 2 6)	55 60 70 (3 4 5)	-12.33	148	-21 (9)
5	36 55 60 (1 3 4)	39 70 73 (2 5 6)	-10.33	151	-15 (8)
6	36 55 70 (1 3 5)	39 60 73 (2 5 6)	-3 67	161	-5 (9)
7	36 55 73 (1 3 6)	39 60 73 (2 4 5)	-1.67	164	-5 (10)
8	36 60 70 (1 4 5)	39 55 73 (2 3 6)	-0.33	166	5 (10)
9	36 60 73 (1 4 6)	39 55 70 (2 3 5)	1 67	169	5 (11)
10	36 70 73 (1 5 6)	39 55 60 (2 3 4)	8.33	179	15 (12)
11	39 55 60 (2 3 4)	36 70 73 (1 5 6)	-8.33	154	-15 (9)
12	39 55 70 (2 3 5)	36 60 73 (1 4 6)	-1.67	164	-5 (10)
13	39 55 73 (2 3 6)	36 60 70 (1 4 5)	0.33	167	-5 (11)
14	39 60 70 (2 4 5)	36 55 73 (1 3 6)	1.67	169	5 (11)
15	39 60 73 (2 4 6	36 55 70 (1 3 5)	3.67	172	5 (12
16	39 70 73 (2 5 6)	36 55 60 (1 3 4)	10.33	182	15 (13
17	55 60 70 (3 4 5)	36 39 73 (1 2 6)	12.33	185	21 (12)
18	55 60 73 (3 4 6)	36 39 70 (1 2 5)	14.33	188	21 (13
19	55 70 73 (3 5 6)	36 39 60 (1 2 4)	21.00	198	31 (14)
20	60 70 73 (4 5 6)	36 39 55 (1 2 3)	24.33	203	31 (15)

In table 2, there are 20 possible two-sample set from the list of permutations distribution. The second data set with asterisk from table 2 above is the observed data from the example. It has a difference of means of -21. The p-value for the lower-tail is the probability of observing a difference of means of -21 or less under the assumption that the treatments do not differ. In this case there are only two numbers less than or equal to -21 out of 20 difference means (that is -21 and -24.33), therefore its p-value is 2/20 = 0.10. If one uses the sum of the treatment 1 (sum of rank of treatment 1) as a statistic, the p-value is the proportion of number less than or equal to 135 (7). In this case is 2/20=0.10. This p-value is *exact* since we calculate it from a permutation distribution, hence it is the exact significance level, not an approximation.

Testing for Deviances

The test explained above is design to distinguish between the effects of two treatments, whether observations from one treatment tend to be larger than observations from another treatment or vice versa. In this section we are interested in the variability of the observations for the two treatments. This variability is very important in quality control. Even though the data have the correct mean or median, it is possible to have an excessive variability within the data. The engineer has to fix this to reduce the variability and the data should be closed to the center. Suppose we are testing that there is no variability between treatment 1 and treatment 2, or H_0: $\sigma_1 = \sigma_2$. If we assume that both samples come from normal distribution then the usual test is $F=S_1^2/S_2^2$, which follows an F distribution with $m-1$ and $n-1$ degrees of freedom. This test is not valid if the underlying distribution is not normal. Higgins (2004) suggested to find the deviances of each treatment i.e., $dev_{i1}= X_i - \mu_1$ and $dev_{j2}= X_j - \mu_2$. The test statistic is the ratio of mean deviance (RMD) as $RMD = \dfrac{\sum_{i=1}^{m} | dev_{i1} | / m}{\sum_{j=1}^{n} | dev_{j2} | / n}$.

If the means are unknown, then replace the means with the medians of the samples. The steps for the permutation test on deviances are similar to the permutation test for two treatment effect. First, compute the test statistic for the original observations, then permute (rearrange) the $m+n$ observations between treatments so that m observations for treatment 1 and n observations for treatment 2. Find the medians for each treatment and estimated deviances then compute the statistic from the sample just permuted. Compare this value to the value obtained for the observed data. If the computed value from the permutation is greater than or equal to the statistic from the original observations, then count 1. Repeat permutation and computation for a number of times. If the sample sizes are small obtain all possible permutations and if sample sizes are large obtain permutations by randomly selecting from all possible permutations. Reject the null hypothesis if the proportion of 1 from the above is less than or equal to predetermined significance level. Higgins (2004) calculated medians only once from the original observations. He then permuted

deviances before calculating statistic, RMD. Our's is different than the Higgins'. Before calculating any statistic from a permutation, we permuted the observations, and then find medians from the permuted data in order to finally calculate deviances and then use a formula RMD above.

Example: This example is from McClave et. al. (1997) textbook page 390. Tests of product can be completely automated or they can be conducted using human inspectors or human inspectors aided by mechanical devices. Although human inspection is frequently the most economical alternative, it can lead to serious inspection error problems. To evaluate the performance of inspectors in a new company, a quality manager had sample 12 novice inspectors evaluate 200 finished products. The same 200 items were evaluated by 12 experienced inspectors. The following table lists the number of inspection errors made by each inspector. The manager believe that the variability of inspection errors was lower for experienced inspectors than for novice inspectors.

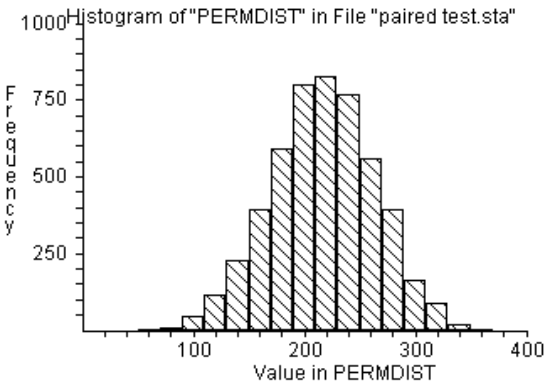
Table 3:

Novice Inspectors						Experienced Inspectors					
30	35	26	40	36	20	31	15	25	19	28	17
45	31	33	29	21	48	19	18	24	10	20	21

If we calculate all possible permutations then we have $\binom{24}{12} = 2{,}704{,}156$ arrangements. Keller-McNulty and Higgins (1987) concluded that there is enough to resample *R=1,600*, since nothing to be gained if taking sample more than 1600. We use *R=1500* in our permutation in order to get an approximation of p-value. We use a software Resampling Stats (2000). Resampling Stats is a simple, but powerful software language that can solve complex problems both in probability and statistics. Since means of the two treatments are unknown, we use the sample medians which are 32, and 19.5, respectively. Calculate the absolute value of the observed minus the median for each treatment. The ratio of median deviance for the original observations $R\hat{M}D$ is 0.6375. The 1st, 2.5th, 5th, 10th, and 20th percentiles of the permutation distribution based on 1,000 randomly selected permutations are found to be .5, .5471, .5913, .6789 and 0.79104. Thus the statistic is significant at the 10% level but not at the 5% level. And the approximate p-value is 0.072. So there is no variability of inspection errors for experienced inspectors and novice inspectors. The histogram of $R\hat{M}D$ shown below is the distribution of the ratio of deviation of errors by experienced inspectors and novice inspectors.

Power Functions of the Tests

In this section we will compare the power of test for Wilcoxon's rank sum test with t-distribution using the permutation method under various nonnormal distributions. We will consider uniform, exponential, log-normal, Poisson, Pareto, and Weibul distributions. We will use small (*m=5, n=5*), moderate (*m=n=20*) and large sample sizes (*m=n=30*). The simulation is conducted as follows: 1) Two independent samples of size *m* and *n* are randomly selected from probability distribution above. 2) We add a constant



to each observation for treatment 1, for example, so that $\mu_1 > \mu_2$. 3) Calculate the observed value of *t*, and the Wilcoxon statistic. 4) Use 1000 resamples of the data to determine a p-value of the permutation test. 5) If a p-value less than 0.05 then reject null hypothesis using a test at the 5 % level. 6) Repeat steps 1-5 for a number of times, say 2000. 7) The power is the proportion of times in the 2000 that are rejected. Again, complete steps 1-7 by increasing the value of a constant in step 2 until you get a wide range of power functions. Figure 1-5 are the graphs of the power functions. It is been known when the distributions of the data are normal with unknown, but equal variances, under a t-test for difference between two samples, t is unbiased test. That is, t has the greatest power and correct of probability of a Type I error. If the underlying distributions are not normal then the power of t is not optimal.

For our simulation, we compare the Wilcoxon's sum rank test with the t-test. Results for small samples (see figure 1-5): t-test is uniformly better than the Wilcoxon's rank sum test for uniform, exponential, and Weibull distributions. Even though the t-test has advantage over the Wilcoxon's, the difference between a power is not significant. There is no clear choice between t-test and Wilcoxon's for Lognormal, and Pareto. For Poisson, most of the time the power of the Wilcoxon's is above of t-test except in two points.

For moderate sample sizes the results are as follows: The uniform is the only distribution in which t-test is more powerful than the Wilcoxon's, again the difference is not very significant. There is no difference between t-test and the

Wilcoxon for Weibull distribution. The Wilcoxon's rank sum test is better than t-test for exponential, lognormal, Pareto, and Poisson. The Wilcoxon's test replaces the original observations by ranks. However, the observations that are unusually large compared to the rest of the data can affect the t-test. Exponential, lognormal, and Pareto are heavy-tailed distribution that can have unusually extreme observations. Whenever the Wilcoxon rank-sum test is uniformly better than t-test, the power of the test is very significant. For the Pareto, for example, let the difference between the means is .50, then the power of Pareto under the Wilcoxon is 86.95% while the power of t-test is 40.70%.

Results for large samples: Some practitioners mistakenly assumed under central limit theorem a power of t-test is uniformly better than the Wilcoxon. However, in our simulation only uniform distribution has the edge for t-test, with no difference for a Weibull. The Wilcoxon rank sum statistic has bigger power for Pareto, lognormal, exponential, and Poisson distributions than that of t-statistic.

Based on our simulation we conclude that for light-tailed distribution such as uniform distribution, the t-test is better than the Wilcoxon's for small, moderate, and large samples. For small sample size, symmetric and light-tailed, t-test is better most of the times. The Wilcoxon sum rank test is more powerful than t-test for moderate and large sample sizes most of the time, when the distributions of the data come from heavy-tailed distributions.

Figure I
One-tailed Power Functions of the Two Independent Mean

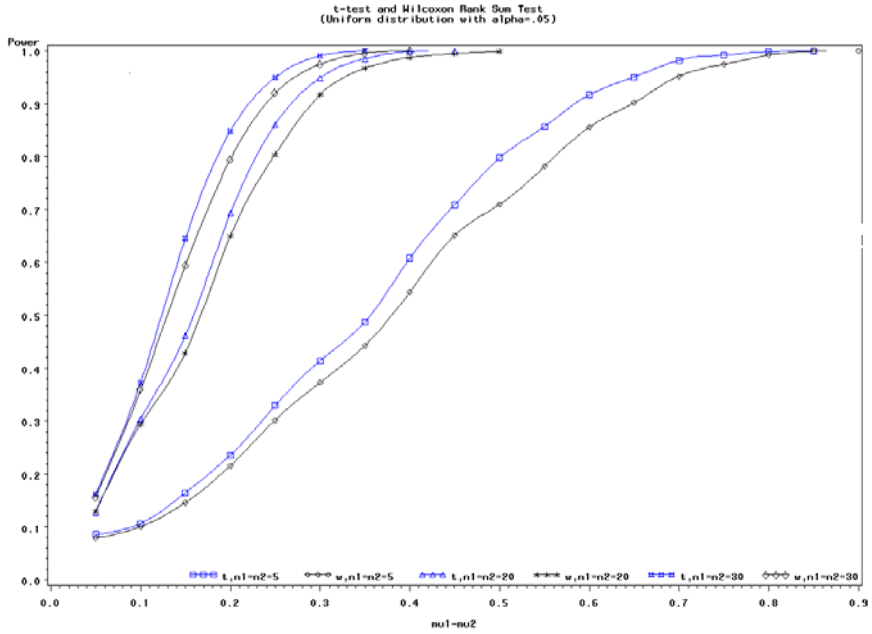


Figure II
One-tailed Power Functions of the Two Independent Mean

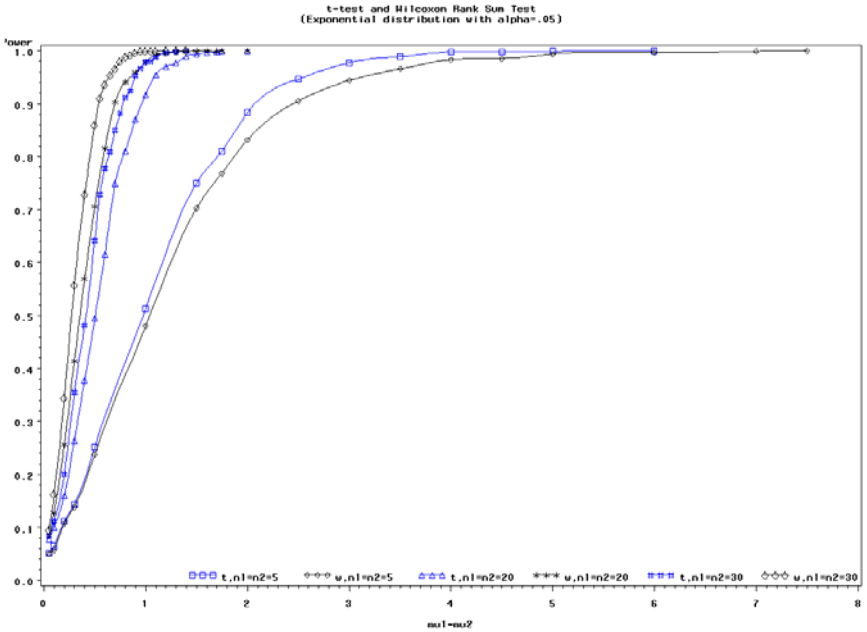


Figure III
One-tailed Power Functions of the Two Independent Mean

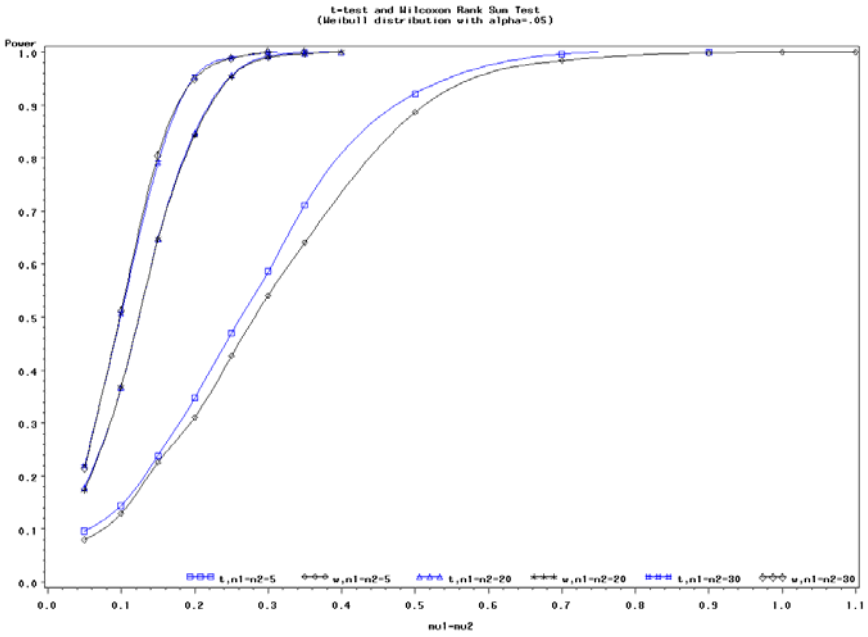


Figure IV
One-tailed Power Functions of the Two Independent Mean

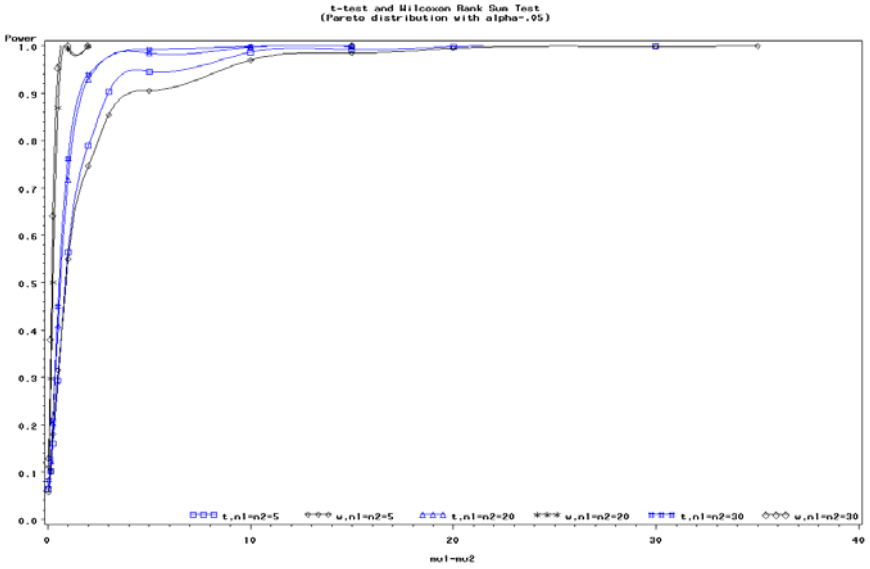
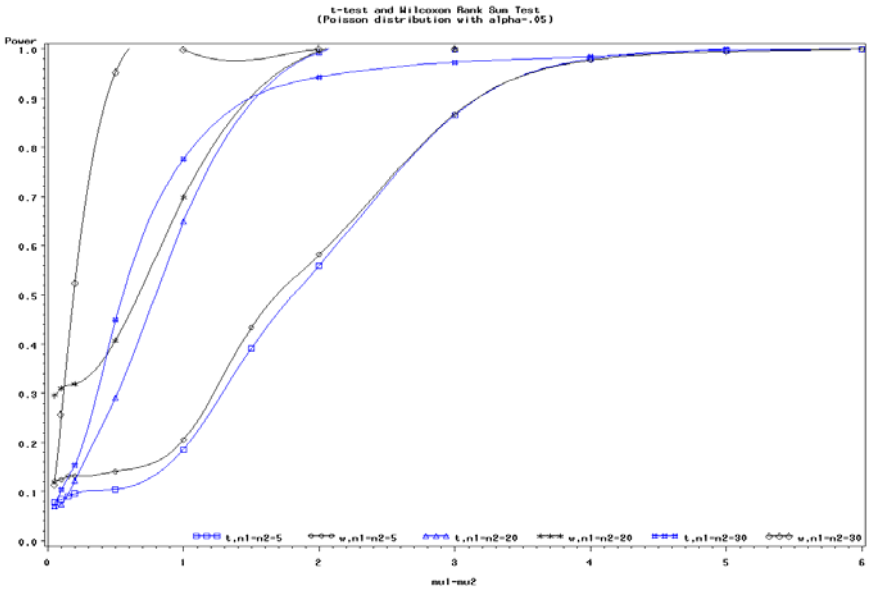


Figure V
One-tailed Power Functions of the Two Independent Mean



† *Ferry Butar Butar, Ph.D.*, Sam Houston State University, Huntsville, Texas
‡ *Jae-Wan Park*, Sam Houston State University, Huntsville, Texas

References

Blair, R.C., and Higgins, J.J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of student's t statistic under various nonnormal distributions, *Journal of Educational Statistics*, 5, 309-335.

Conover, W. J. (1999). Practical Nonparametric Statistics, third ed., John Wiley & Sons, New York.

Good, P. (1994). Permutation Tests: A practical guide to resampling methods for testing hypothesis, Springer, New York.

Hicks, C.R., Turner, K.V. (1999). Fundamental Concepts in the Design of Experiments, Fifth ed., Oxford University Press, New York.

Higgins, J.J. (2004). Introduction to Modern Nonparametric Statistics, Duxbury.

Resampling Stats User's Guide. (2000). Resampling Stats, Inc.", Vol 5.0.2, Arlington, VA.

Siegel, S. and Castellan, N.J. (1998). Nonparametric Statistics for the Behavioral Sciences", second ed. McGraw Hill, Boston.