Comparison of the Power of the Paired Samples using Permutation Tests

Ferry Butar Butar, Ph.D. †
Ananda Bandulasiri, Ph.D. ‡

Abstract

The *t* test is classic for testing a paired comparison when the distribution of difference scores from a random sample are normally distributed. For unspecified distributions, the sign test or the Wilcoxon signed rank test can be utilized. Without knowledge of the underlying distributions, the permutation test can be utilized both for the original observations as well as their ranks. For the permutation test, the significance level is exact when calculating all possible permutations. The approximate significance level is used when the numbers of permutations are very large. A simulation study is conducted to compare the power of the *t*, signed, and signed rank tests.

Introduction

One way to do an experiment is to assign a group of subjects to one particular treatment and to assign another group of subjects to another treatment. In such an experiment, the experimenter would like to determine whether there is no treatment difference between the two groups. Another way of comparing two treatments is by pairing the subjects. Here, the experimenter wants to assign two treatments to the same subject, one before some circumstance and the other after. Another matching possibility is the selection of pairs of students with the same ability and motivation. Such selection should be based pairing subjects that are of the highest similarity.

The usual method of analyzing data from two related samples is by applying Student's *t* test to the difference in average scores. To use the *t* statistic, an experimenter can assume that the difference scores are an independent random sample from a normal distribution. When these assumptions are satisfied, the *t*-test is unbiased and is the most powerful test to use. In practice, the *t*-test assumptions on the observations are sometimes unrealistic. Sometimes the differences between scores are not represented by scores, but rather as signs or ranks.

In the next section, we will explain the sign test, which is applicable to both quantitative and qualitative observations. Then, we will explain the matched pairs analogue with scores by using a permutation test. Under the null hypothesis, one can permute the scores as either positive or negative. We will also explain the Wilcoxon rank sum test. If the sample size is small, one can permute all possible permutations. If the sample is moderately large, then one can select a random sample of permutations. A large sample approximation will be utilized for large sample size. Finally, we will use simulation for comparing the power of the tests.

Hypothesis Testing

Consider a bivariate random sample of size n, (X_1,Y_1), (X_2,Y_2),..., (X_n,Y_n). The comparison is made for each pair of (X_i,Y_i),i=1,2,...,n. Let "+" be the event of $X_i < Y_i$ or $X_i - Y_i < 0$ (the positive difference between $Y_i - X_i$), let "-" be the event that $X_i > Y_i$ or $X_i - Y_i > 0$, and "0" be the event $X_i = Y_i$ or $Y_i - X_i = 0$. The null hypothesis to be tested is H_0: $P(X < Y) = P(X > Y)$ and the two-tailed alternative hypothesis is H_a: $P(X < Y) \neq P(X > Y)$. The lower-tailed alternative is H_a: $P(X < Y) < P(X > Y)$, which is the number of positive is less than the number of negative or where *X* tends to be smaller than *Y*. The upper-tailed alternative is H_a: $P(X < Y) > P(X > Y)$, which is the number of positive is more than the number of negative or where *X* tends to be greater than *Y*. In the case of testing the location of the parameter or the median, then the null hypothesis becomes H_0: $E(X - Y) = 0$. The median of *X* equals the median of *Y* against the two tailed alternative H_0: $E(X - Y)$. If *X* and *Y* have a different median, the lower tailed alternative is H_0:$E(X - Y) > 0$ and the upper tailed alternative is H_0:$E(X - Y) < 0$. See (Conover, 1999).

Sign Test

Dating back to 1710, the sign test is the oldest test of all nonparametric tests. This test is simple and easy to use. The sign test can be applied to both numerical and ordinal measurements. For this test, the differences between $Y_i - X_i$ are random variables with symmetric distribution. The distribution of a random variable *X* is symmetric about $x=0$ if $P(X \leq -x) = P(X \geq x)$ or $F(x) = 1 - F(-x)$. If we omit the ties from the difference between $Y_i - X_i$, the sign test is used to compare the probability of positive difference with negative difference. Under the null hypothesis, we expect that the number of positive pairs of $X_i < Y_i$ are equal to the number of negative pairs of $X_i > Y_i$. Thus, when the null hypothesis is true we will have about half of the difference scores to be positive and about half to be negative. In a Bernoulli trial, we may call each positive pair a "success" and a negative pair as a "failure". Since $P(X_i < Y_i) + P(X_i > Y_i)=1$ and under the null hypothesis, $P(X_i < Y_i) = P(X_i > Y_i)$ hence H_0: $P(+) = P(X_i < Y_i) = 1/2$. Note that in terms of testing the medians, the null hypothesis becomes the median difference between X and Y which is zero. See (Conover, 1999, and Siegel & Castellan, 1988).

Now let the test statistic S_+ , be the number of + pairs; that is S_+ is the number of pairs (X,Y) where X_i is less than Y_i. Thus, for the lower-tailed test, p-value is given by

$$P(S_+ \leq t) = \sum_{i=0}^{t} \binom{n}{i} \left(\frac{1}{2}\right)^n$$

Example

Below is the grade of my students from one of my elementary statistics classes in Spring semester 2005. Let X_i be the first exam and Y_i be the second exam. Consider the hypothesis that the grade for the second exam is higher than that of the first exam. The grade is as follows:

Name	Xi	Yi	Yi - Xi	Name	Xi	Yi	Yi - Xi
Justin	46	45	-	Charmain	81	66	-
Travis	65	86	+	Alexis	93	90	-
Vanessa	62	82	+	Brandon	46	51	+
Shelton	75	65	-	Tonya	57	70	+
Jessica	83	93	+	Garrett	59	47	-
Rebekah	82	78	-	Brittany	70	86	+
Talley	72	45	-	Matthew	74	88	+
Adam	64	63	-	Stephanie	80	54	-
Bradley	64	63	-	Kelli	81	60	-
Jeremy	66	76	+	Megan	84	79	-
Shannon	59	78	+	Marissa	82	82	0
Schimank	59	80	+	Melinda	96	92	-
Robin	78	75	-	Lindsey	59	67	+
Priscilla	63	60	-	Cody	91	97	+
Robert	73	93	+	Doyle	56	40	-

From the above table, there are 16 positive, 13 negative, and 1 tie. We discard a tie, and use the remaining sample of 29. The p-value is $P(S_+ \leq 13)=\sum_{i=0}^{13}\binom{29}{i}\left(\frac{1}{2}\right)^{29} = 0.3555$. Since the p-value is big, there is no difference between the first and the second exam scores.

Large Sample Approximation for Sign Test

If the sample size is large, one can use the standard normal distribution to approximate binomial probability. Recall that the mean of binomial is $np = n/2$, and the variance is $np(1-p) = n/4$. Thus, the value of z is given by $z = (S_+ - \mu) / \sigma = (S_+ - .5n) / (.5\sqrt{n})$. Note that one needs to make an adjustment from discrete probability (binomial) to a continuous distribution (normal) and which is called continuity correction. Therefore, the formula becomes $z = (S_+ \pm .5 - .5n)/(.5\sqrt{n})$, where ± 0.5 is used for the continuity correction.

From the previous example, using the normal approximations, we have $z = -0.56$ and $P(z < -0.56) = 0.2877$.

The Wilcoxon Signed Ranks Test

The sign test discussed in the previous section analyzes the data based on the sign of the differences within pairs, which include positive, negative, and zero. If we are considering the sizes of the positive differences relative to the sizes of negative differences, a more powerful test can be used. The Wilcoxon signed ranks test is considering both the sign and the magnitude of the differences.

The practitioner can tell the sign of the difference between any pair, and rank the differences in order of absolute size. Let $D_i = Y_i - X_i$ be the difference score for any matched pair, representing the difference between the pair's scores under two treatments X and Y. When the two scores of any pair are equal, that is $D_i = X_i - Y_i = 0$ (no difference between two treatments is observed for that pair), omit these pairs from further analysis and the sample size is reduced accordingly. Thus, *n* is the number of matched pairs minus the number of pairs for which $D_i = 0$.

To use the Wilcoxon signed rank test, ranks from 1 to n are assigned to these pairs according to relative size of $|D_i|$; the rank of 1 is given to the smallest $|D_i|$, the rank of 2 to the next smallest, etc, with n assigned to the largest $|D_i|$. If two or more D_i's are equal to each other, assign D to each of these which is the average rank that would have otherwise been assigned. For example D_i's are -2, -2, 2, 2. Assign the tied values the same rank, that is the average rank (1+2+3+4)/4 = 2.5 to each of the four D_i's. Then to each rank, assign -2.5, -2.5, for the negative D_i's and 2.5, 2.5 for the positive D_i's. The next D_i in order would receive the rank of 5 since the rank of 1, 2, 3, 4 have already been assigned.

In this test we assume that the distribution of the differences is symmetric. It is easy to see symmetry in discrete distribution, since the left half of the graph of the probability function is the mirror image of the right half. For example sign test with p=.5 is symmetric.

Definition: The cumulative distribution function (cdf) of any random variable X is symmetric at about x=b, for some constant b, if the probability of $X \leq b-x$ equals the probability of $X \geq b+x$ for every value of x.

Example: Let a random variable *X* of the continuous type have a probability distribution function (pdf) whose graph is symmetric with respect to $x = b$. If the mean value of *X* exists, then E(X) = b.

Proof: $E(X-b) = \int_{-\infty}^{\infty} (x-b) f(x)dx = \int_{-\infty}^{b} (x-b) f(x)dx + \int_{b}^{\infty} (x-b) f(x)dx$. Let y = b - x in the first term and let z = x - b in the second term on the right hand side of the equal sign. Thus, E(X-b)= - $\int_{\infty}^{0} y f(b-y)dy + \int_{0}^{\infty} z\, f(b+z)dz$. Due

to symmetry $f(b-y) = f(b+z)$, thus E(X - b)= -

$\int_0^\infty y f(b-y)dy + \int_0^\infty z\, f(b+z)dz = 0$. Therefore, E(X) = b.

As explained in the section of hypothesis testing, we want to compare a subject before and after training. Let D be the random variable of the difference scores for any pair (X,Y). Its cumulative distribution function is F(d). The null hypothesis to be tested is H_0: F(d) = 1 - F(-d) and the alternative hypothesis is given by H_a: F(d) < 1 - F(-d) or H_a: F(d) > 1 - F(-d) or H_a: F(d) ≠ 1 - F(-d). The first alternative says that the differences tend to fall more to the negative side of 0. A special case of the alternative is a location parameter where F(d) = G(d - δ), where G(d) is a distribution symmetrical about 0. Thus the hypothesis becomes H_0: δ =0 and Ha: δ > 0 or Ha: δ < 0 or Ha: δ ≠ 0. This δ can be the median or the mean of the differences.

If H_0 is true, the treatment *X* and treatment *Y* are equivalent, that is the samples of *X* and *Y* are both from populations with the same medians, or the numbers of positive are the same as the numbers of negative. For the null hypothesis to be true, we would expect to find some of the larger positive D_i tends to favor treatment *X* and some tend to favor treatment *Y*. That is, when there is no difference between treatment *X* and treatment *Y*, some of the larger ranks would come from positive D_i's whereas others would come from negative D_i's. Thus if we summed those ranks having positive signs and summed those ranks having negative signs, we would expect the two sums to be approximately equal when the null hypothesis is true. When the sum of the negative ranks is very much different from the sum of the positive ranks, we would conclude that the treatment *X* is different from the treatment *Y*, and thus we would reject the null hypothesis. We reject H_0 if either the sum of the ranks for the negative D_i's or the sum of the ranks for positive D_i's are too small or too big. (See Conover, 1999; Hollander & Wolfe, 1999; and Siegel & Castellan, 1988).

Permutation Test for Rank Observations

Assume that there is no tie in the difference between scores. Under the null hypothesis, for each pair, any difference is as likely to be a positive as it is to be a negative, that is for each pair X > Y is the same chance as X < Y. So there are 2^n possible permutations. The steps for the Wilcoxon signed rank test are as follows:

- Determine the sign rank and calculate SR_{+obs}, the sum of positive signed ranks for the original observations
- Examine all 2^n possible permutations, obtain the signed ranks, and calculate statistic SR_+
- Compare each SR_+ statistic computed above with SR_{+obs}. For the upper tailed test, whenever $SR_+ \geq SR_{+obs}$ record 1.

- The upper p-value is the total number of 1's divided by 2^n. For the two-tailed test , the p-value is two times the one-tailed p-value.

If it is not possible to obtain all possible permutations, one can randomly select sample permutations, say 1000 or 2000. Let R_i be the ranks of absolute difference. Consider independent random variables $U_1, U_2,..., U_n$ where $U_i = 0$ with probability 0.5, and $U_i = 1$ with probability 0.5.

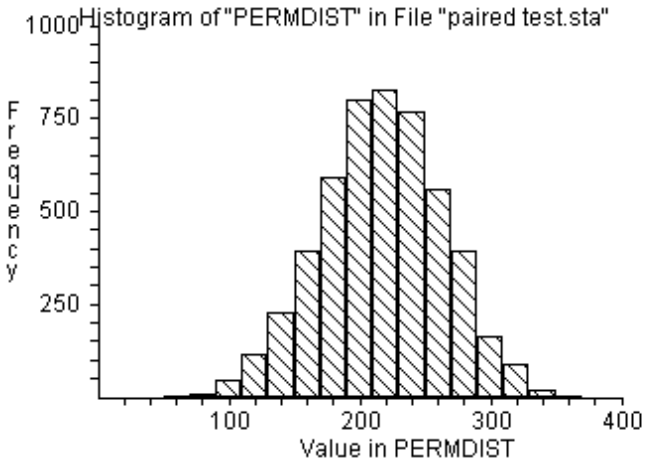
Then, randomly selected permutations can give the value of SR_+ as follows; $SR_+ = \sum_{i=1}^{n} U_i R_i$. (See Good, 1994; and Higgins, 2004).

Example

In the previous example , there are 2^{29} = 536, 870,912 possible permutations of the data. Using Resampling Stats software we take a random sample of 5000 from all possible permutations. The approximate one tail p-value is 0.2384 . For the pair t-test, $\bar{d}$ =2.3448, and s_d=17.3056. The p-value for the paired t-test is P(t > 29*2.3448/17.3056) = P(t > 0.7297) = 0.2358. The frequency distribution and histogram of the permutation distribution is shown below.

Frequency of Permutation distribution

Bin Center	Frequency	Percent	Cumulative percent
60	1	0.0	0.0
80	6	0.1	0.1
100	46	0.9	1.1
120	114	2.3	3.3
140	226	4.5	7.9
160	393	7.9	15.7
180	592	11.8	27.6
200	801	16.0	43.6
220	826	16.5	60.1
240	767	15.3	75.4
260	562	11.2	86.7
280	394	7.9	94.6
300	161	3.2	97.8
320	86	1.7	99.5
340	20	0.4	99.9
360	5	0.1	100.0



Large Sample Approximation for signed ranks test

For a large sample size, we can use the standard normal table to approximate the p-value.

The mean of SR_+ is $\mu = E(SR_+) = \frac{1}{2}\sum_{i=1}^{n} i = \frac{n(n+1)}{4}$, and the variance of SR_+ is

$$Var(SR_+) = \sum_{i=1}^{n} i^2\, Var(U_i) = \sum i^2[E(U_i^2) - E^2(U_i)]$$

$$= \sum_{i=1}^{n} i^2 (0^2.\frac{1}{2} + 1^2.\frac{1}{2} - [0.\frac{1}{2} + 1.\frac{1}{2}]^2)$$

$$= \sum_{i=1}^{n} i^2 (\frac{1}{2} - \frac{1}{4})$$

$$= \frac{1}{4}\sum_{i=1}^{n} i^2$$

$$= \frac{1}{4}\frac{n(n+1)(2n+1)}{6}$$

$$= \frac{n(n+1)(2n+1)}{24}$$

Thus the z score is $z = \dfrac{(SR_+ \pm .5) - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$.

From the previous example we have, $SR_+=251.5$, $E(SR_+)=217.5$, and $Var(SR_+)=2138.75$. Thus $z = \frac{251.5 - 0.5 - 217.7}{\sqrt{2138.75}} = 0.72$. The p-value = $P(z > 0.72) = 0.2358$.

Theorem: For any positive integers n, and k, $\sum_{r=0}^{k-1} \binom{k}{r} S(n,r) = (n+1)^k - 1$, where $S(n,r) = 1^r+2^r+3^r+...+n^r$, r=0,1,2,...

Lemma: $S(n,0) = 1^0+2^0+3^0+...+n^0 = n$,
$S(n,1) = 1^1+2^1+3^1+...+n^1 = n(n+1)/2$,
$S(n,2) = 1^2 + 2^2 + 3^2+...+n^2 = n(n+1)(2n+1)/6$,
$S(n,3)=1^3+2^3+3^3+...+n^3= n^2(n+1)^2/4$.

Proof: (See Freund)

Permutation Test for the Original Observations

There are two different methods to determine the test for the permutation distributions. One method is to find the mean of the differences, and the other method is to either obtain the sum of positive differences, SO_+ , or the sum of negative differences, SO_. Both methods are equivalent, since SO_+ + S0 is equal to the sum of all observations. Thus, the mean of the differences is just the average of the sum of the positive and the negative. Below are the steps for computing p-values using the permutation distributions.

- Obtain the differences from observations, D_i's and compute the mean of the differences, $\overline{D}_{obs}$.
- For the differences of n samples, determine the number of possible outcomes of plus and minus signs to the $|D_i|$'s, 2^n.
- For each of the 2^n, compute statistic $\overline{D}$
- Compare each statistic $\overline{D}$ computed above with $\overline{D}_{obs}$. For the upper tailed test, whenever $\overline{D} \geq \overline{D}_{obs}$, record 1.
- The upper p-value is the total number of 1's divided by 2^n. For the two-tailed test, the p-value is two times the one-tailed p-value.

If it is not possible to obtain all 2^n permutation distributions, we can select random sample permutations, 2000 for example. Let the random variables $W_1,W_2,...,W_n$ be independent, where W_i = -1 with probability 0.5 and W_i =1 with probability 0.5. Then a randomly selected permutation can give the mean value of the differences, $\overline{D}$, as follows $\overline{D} = n^{-1}\sum_{i=1}^{n} W_i \mid D_i \mid$. The permutation distribution of S_+ is similar to the SR_+, except R_i is replaced by $|D_i|$. (See Higgins, 2004).

For a large sample size, one can approximate the p-value using standard normal approximation. Calculate the variance of the mean difference which is $n^{-2}\sum_{i=1}^{n} | D_i |^2$. Then use z score to standardize the $\overline{D}$. Note that the expected value of the mean differences is zero since $E(W_i) = 0$. Again using example 3.1, we have $S_+=68$ and $Var(S_+) = 5658$. Thus p-value = $P(z > 68/\sqrt{5658}) = P(z > 0.90) = 0.1841$. Therefore, there is no sufficient evidence to conclude that the mean score of exam 2 is greater than the mean score of exam 1.

Example

When we apply a paired t-test, the mean of the difference is 1.1034, the standard deviation is 14.1708, the test statistic is $t = 0.4193$, the one-tailed p-value is 0.3391. Since the number of permutations are too large, we then randomly sampled 5000 from all possible permutations. The upper-tailed p-value is 0.3518.

Power of the Tests

Butar Butar and Park (2008) have been successfully comparing the power of the tests for two independent samples. They simulated data from the uniform, exponential, chi-square, lognormal, and the Pareto distributions. Here we would generated the observations from the uniform, exponential, chi-square, lognormal, and the Pareto distributions mentioned above. But it was not possible that the distribution of the differences of the data could be a Uniform, Exponential, chi-square, lognormal, and a Pareto. Here we would compare the pair of the t-distribution to the sign and the Wilcoxon signed rank test for each distribution mentioned above. We will generate the observations from each marginal distribution, not from their difference score. All populations were generated by using the Matlab. The following is how we generate the data. Consider a group of n subjects that have been tested at two different times. Assume the first data generated were a "before" score of an individual is represented as

$$y_{i1} = \theta_i + \varepsilon_{i1}, \quad i = 1,2,...,n,$$

where y_{i1} is the score of the ith subject at the first testing period, θ_i is the true score of the ith subject's score, and ε_{i1} is the random error of the ith subject's score at the first testing period. An individual's "after" score is represented as

$$y_{i2} = \theta_i + \varepsilon_{i2}, \quad i = 1,2,...,n,$$

where y_{i2} is the score of the ith subject at the second testing period, θ_i is the true score of the ith subject's score, ε_{i2} is the random error of the ith subject's

score at the second testing period, and μ is a treatment effect, which is common to all subject's "after" scores. The difference between the "after" and "before" test score is given by:

$$y_{i2} - y_{i1} = \mu + \varepsilon_{i2} - \varepsilon_{i1}, \; i=1,2,...,n.$$

An individual score of ε_{i1}, ε_{i2} was generated from the Uniform, Exponential, chi-square, lognormal, or the Pareto. The value of μ will be increased from 0 to a larger number so that the power of the test would be closer to one for both tests.

We will use small (n=10), moderate (n=30) and large sample sizes (n=50). Following Butar Butar and Park, the simulation is conducted as follows: 1) Two samples of size n each are randomly selected from the probability distribution above. 2) Add a constant, μ, to each difference observation of treatment 2 and treatment 1. 3) Calculate the observed value of the t, the Wilcoxon signed, and the Wilcoxon signed rank statistic. 4) Resample the data 1000 times to determine the p-value of the permutation test. 5) If a p-value is less than 0.05, then reject the null hypothesis using a test at the 5 % significance level. 6) Repeat steps 1-5 for a number of times, say 2000. 7) The power is the proportion of rejections. Again, complete steps 1-7 by increasing the value of a constant, μ, in step 2 until one gets a wide range of power functions. (See also Blair and Higgins, 1985).

Figure 1-5 are the graphs of the power functions. Figure 1 shows the uniform distribution. Except for n=10, where the t-distribution looks best for n=30 and n=50. For the exponential, chi-square, lognormal, and the Pareto distributions (figures 2 - 5), both the sign and the Wilcoxon rank signed test are uniformly better than the t-test.

Conclusion

We have shown three different methods for paired comparisons. The sign test is the simplest and it is easy to use. The only requirement for the sign test to work, is that the ordinal measurement is within pairs; that is if the score of one member of a pair can be ranked as less than the score of the other member of the same pair. When the order of magnitude becomes a concern, then we recommend the use of the Wilcoxon signed rank test. Also in the Wilcoxon signed rank test, the ordinal scale is both within and between pairs observations. The permutation test can be used for both original observations, as well as for the rank of the data. It is a powerful test compared to the parametric t-test. When the data are heavy-tailed, the permutation test is much more efficient than the t-test. From the simulations, we found that it is better to utilize permutation tests rather than the regular t-distribution for the skewed distributions; we see evident from the power comparisons (see figures 2 – 5).

Figure 1. One–tailed Power Functions of the pair Means

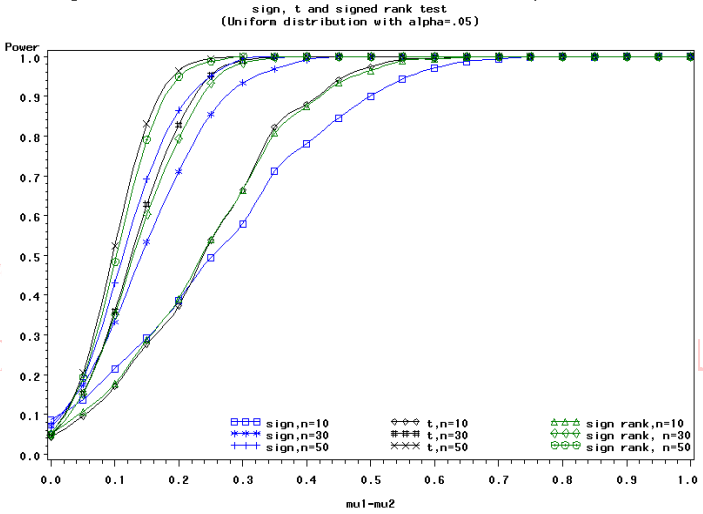


Figure 2. One–tailed Power Functions of the pair Means

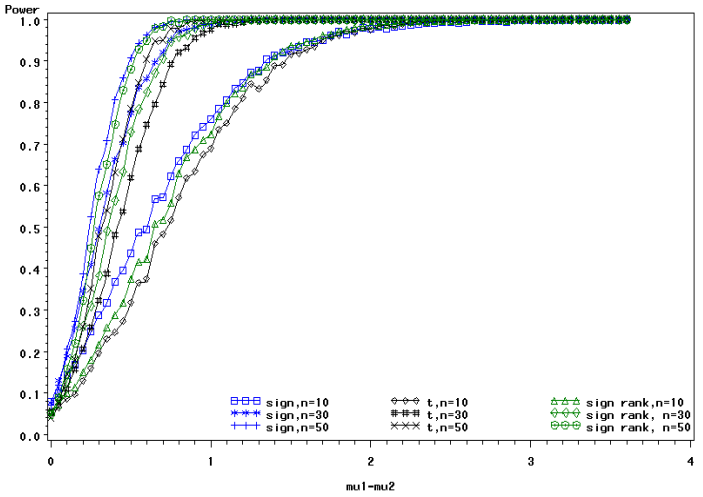


Figure 3. One–tailed Power Functions of the pair Means

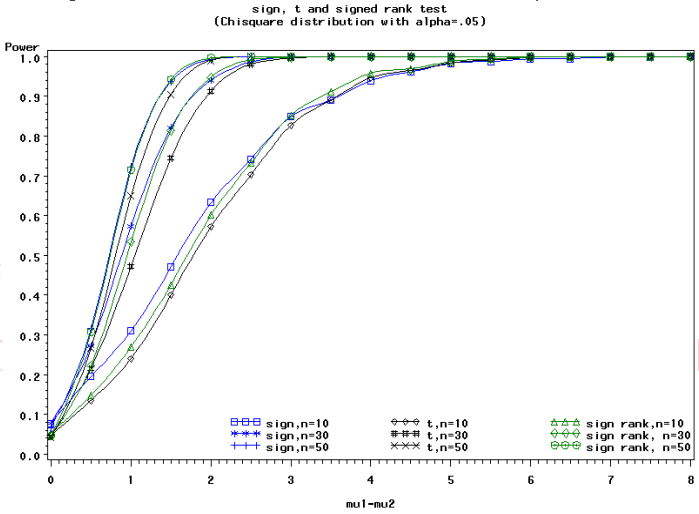


Figure 4. One–tailed Power Functions of the pair Means

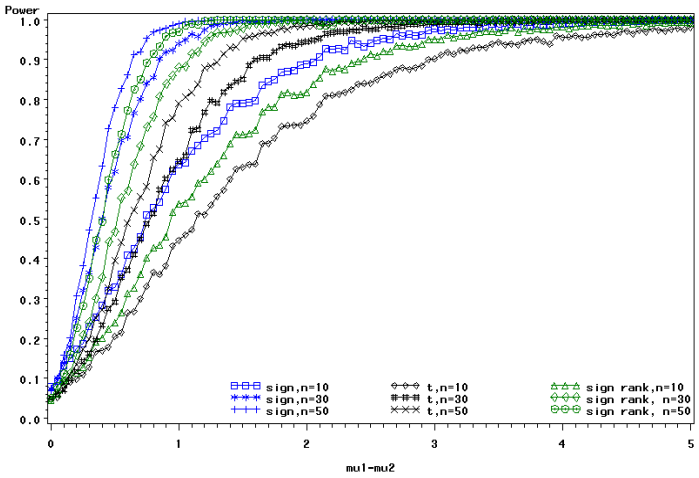
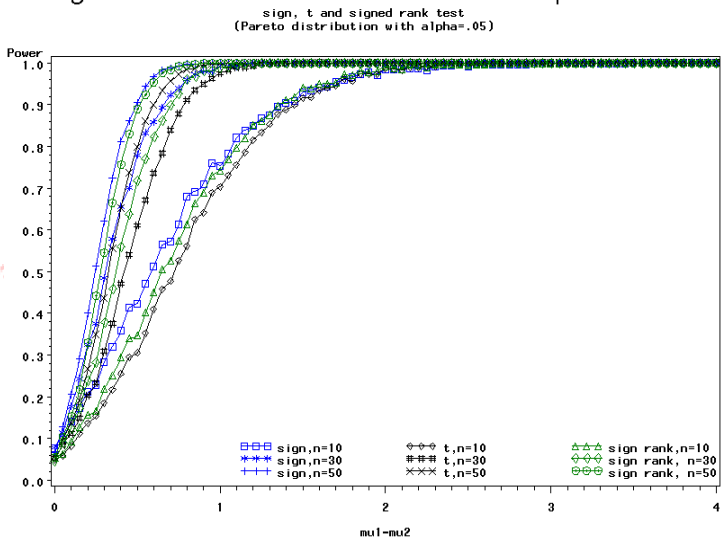


Figure 5. One–tailed Power Functions of the pair Means



† *Ferry Butar Butar, Ph.D.*, Sam Houston State University, TX, USA
‡ *Ananda Bandulasiri, Ph.D.,* Sam Houston State University, TX, USA

References

R. C. Blair, and J. J. Higgins, A comparison of the Power of the Paired Sample T-test to that of Wilcoxon's signed-Ranks test under various Population Shapes, Psychological Bulletin, 97, 119-128, 1985.

F. Butar Butar, and Jae-Wan Park, Permutation tests for comparing two populations, Journal of Mathematical Science & Mathematics Education V3 No. 2 pp 19 – 30, 2008.

Freund, J. E, Mathematical Statistics, fifth edition, Prentice Hall, New jersey, 1992.

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

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

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

M. Hollander, and D. A. Wolfe, Nonparametric Statistical Methods, Second ed., Wiley, New York, 1999.

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

S. Siegel, and N. J. Castellan, Nonparametric Statistics for the Behavioral Sciences, second ed. McGraw Hill, Boston, 1988.