Journal of Modern Applied Statistical Methods

Volume 6 | Issue 1 Article 10

5-1-2007

Comparison of the t vs. Wilcoxon Signed-Rank Test for Likert Scale Data and Small Samples

Gary E. Meek
Black Hills State University

Ceyhun Ozgur
Valparaiso University

Kenneth Dunning
The University of Akron

Recommended Citation
Meek, Gary E.; Ozgur, Ceyhun; and Dunning, Kenneth (2007) "Comparison of the t vs. Wilcoxon Signed-Rank Test for Likert Scale Data and Small Samples," *Journal of Modern Applied Statistical Methods*: Vol. 6 : Iss. 1 , Article 10.
DOI: 10.22237/jmasm/1177992540

Journal of Modern Applied Statistical Methods
May, 2007, Vol. 6, No. 1, 91- 106

Comparison of the t vs. Wilcoxon Signed-Rank Test for Likert Scale Data and Small Samples

Gary E. Meek
Black Hills State University

Ceyhun Ozgur
Valparaiso University

Kenneth Dunning
The University of Akron

The one sample t-test is compared with the Wilcoxon Signed-Rank test for identical data sets representing various Likert scales. An empirical approach is used with simulated data. Comparisons are based on observed error rates for 27,850 data sets. Recommendations are provided.

Key words: Nonparametric, Wilcoxon's signed-rank test, one sample t-test, Likert scales, Type I and Type II error rates.

Introduction

There has been disagreement since the 1940s concerning the use of the t-test versus its nonparametric equivalents when the assumptions of the t-test may not be valid, particularly those of normality. Similarly, controversies have raged at various times over the past 60 years about the use of classical or parametric procedures versus distribution-free or nonparametric procedures when the level of measurement is less than interval. The discussions in the literature began with Stevens (1946) and Siegel (1956) who stated that the level of measurement attained in the data should be a major factor in test selection. Siegel (1956) took a definite stance that nonparametric procedures should be utilized whenever the level is no more informative than ordinal.

In the behavioral sciences, particularly in psychology, Baggaley (1960) and Binder (1984) fueled the fire started by Stevens (1946). The extensive use of Likert scales in the behavioral sciences continues to make test selection a debatable issue. The debate is not restricted to the social sciences, because Likert scales also are widely used in opinion-based research in marketing, human resource management and other areas of business as well as in education and nursing. The liveliness of the discussions surrounding this issue in presentations at various conferences provided the motivation for this investigation.

Comparisons of distribution-free and parametric procedures initially were based upon theoretical considerations involving asymptotic relative efficiency (ARE), which is a large sample property. It pertains to the limit of the ratio of the sample sizes required to attain a specified power as the alternative, or true value, approaches the value under the null hypothesis and the sample size goes to infinity. Although the ARE is theoretically appealing, infinite sample sizes are difficult to obtain in practice.

According to Conover (1999) and Siegel (1956), the ARE of the one sample Wilcoxon signed-rank (WSR) test for location as compared with the one sample t-test for a normal population is 0.955. Conover (1999) stated that if the underlying population is uniformly distributed the ARE is 1.0 and for most non-normal populations exceeds 1.0, but is never less than 0.864.

Gary Meek is retired after 40+ years in academia. He has 40+ peer-reviewed publications. His doctorate in statistics is from Case Western Reserve University. Email address: meek@mato.com. Ceyhun Ozgur is Professor of Information/Decision sciences, and has published about two dozen refereed journal articles. Email address: Ceyhun.Ozgur@valpo.edu. Kenneth A. Dunning is Professor of Management / Information systems. He specializes in six-sigma applications. Email address: dunning@uakron.edu

The ARE is an important theoretical consideration for comparing the theoretical power of two different statistical procedures. However, it is considered to be of limited value when working with small samples. Sawilowsky (1990) stated that at best, Monte Carlo studies have shown that ARE may be indicative of the promise of relative power of non-parametric procedures versus their parametric counterparts for small samples.

Conover (1999) pointed out that the t-test is more powerful than its rank-sum alternatives when populations are normally distributed. However, as most statisticians would agree, normality is a very difficult property to obtain. Sawilowsky and Blair (1992) demonstrated that when populations are not normally distributed, the Wilcoxn rank-sum procedure is more powerful than the t-test.

For the correlated layout, Siegel (1956, p. 83) stated that, for small sample sizes, the efficiency of the Wilcoxon signed-rank test to the t-test is near 95 percent. Most textbooks that include the Wilcoxon signed-rank procedure and discuss its assumptions versus the t-test recommend using the Wilcoxon signed-rank test in small sample situations whenever there is any question about normality or an interval level of measurement, but symmetry is reasonable and the differences are ordinal. Therefore, the purpose of this article is to compare the performance of these two procedures with Likert scale data in small sample situations where the assumptions of normality and interval measurement, are not satisfied.

Some Literature Review for Independent and Dependent Tests with Ordinal Scaled Data

There is limited evidence of practical comparisons of parametric versus nonparametric procedures based on the actual scale of measurement available in the data. The term practical is used because discussion in the literature, Stevens (1946), Siegel (1956), Baggaley (1960), Binder (1984) and Conover (1999), has historically been predicated on philosophical issues or asymptotic properties. Sawilowsky (1991) presented an excellent summary of the level of measurement issue and the weak measurement versus strong statistics controversy.

Some studies that have considered scale of measurement in comparing parametric vs. nonparametric tests are Blair and Higgins (1985), Nanna and Sawilowsky (1998), Nanna (2002), and two preliminary studies done by Meek, et al., (2000) and (2001).

Blair and Higgins (1985) used Monte Carlo methods with ten theoretical distributions to compare power of the paired samples t-test and the Wilcoxon signed ranks test for paired data, utilizing samples of 10, 25 and 50. They found the paired t-test to have a slight power advantage over the Wilcoxon procedure under normal and uniform distributions but little or no advantage under the other distributions for n = 10 and none at the larger sample sizes.

The first study by Meek, et al., (2000) used an identical approach to that utilized in this paper but compared the two independent samples t-test to the Mann-Whitney procedure under various combinations of Likert scales and sample sizes. Their findings indicated that, for small samples, there appeared to be little difference in precision between the t-test and the Mann-Whitney for data collected on a Likert scale. More germane to this article, the second article by Meek, et al., (2001) used a similar approach to compare the performances of the Wilcoxon signed-rank test and the t-test with Likert scales but was limited by having only slightly more than 2400 cases, and therefore, the results are not discussed further here.

Nanna and Sawilowsky (1998) compared the power of the independent samples t-test to that of the Wilcoxon rank-sum procedure with actual data sets measured on an ordinal scale. Their data were based on Functional Independence Measure (FIM) scores in medical rehabilitation. FIM scores used a 7-point Likert scale and often are highly skewed. Nanna and Sawilowsky (1998) found that the Wilcoxon rank-sum procedure had higher power than the t-test test for almost all combinations of sample size and alpha level examined. Nanna (2002) found that the rank transformation procedure provided an increase in power over Hotelling's T^2 when testing for equality of centroids using Likert scale data. Nanna (2002) used essentially the same FIM data sets as Nanna and Sawilowsky (1998).

Many of the current textbooks that include coverage of the Wilcoxon signed-rank procedure are quite limited in their discussions of its assumptions and make no specific recommendations for its use compared to the t-test other than to indicate it should be used if the assumptions of normality and interval measurement are questionable, particularly in small sample situations. Although robustness of the t-test is often cited as a reason for choosing t over the Wilcoxon signed rank and other nonparametric procedures, Bradley (1980) found that both the Z test and t-test were very non-robust for L-shaped distributions when comparing average p-values to nominal alphas.

Sawilowsky (1990) stated that the concept of robustness relates to both Type I and Type II errors and that choosing a test procedure requires one to consider other issues and properties too. Sawilowsky (1991) pointed out that there are no hard and fast rules for choosing between parametric and nonparametric procedures and Sawilowsky (2005) presented a summary of misconceptions regarding such choices. Heeren and D'Agostino (1987) found the independent samples t-test to be robust with ordinal data while Sawilowsky and Blair (1992) found that the t-test was reasonably robust when sample sizes were equal and at least of size n = 30 per group.

Several current statistics texts in the business field were reviewed to determine how they presented nonparametric versus parametric procedures. Anderson, et al., (2005), Moore, et al., (2003) and Newbold (1995) did not mention the assumptions underlying the Wilcoxon signed-rank test and made no recommendations regarding its use. Bowerman, et al., (2007) stated that when n is small, the distribution is non-normal and the measurement is ordinal the t-test is not valid and the Wilcoxon signed-rank test should be used. Keller (2005), Berenson, et al., (2004) and Chou (1989) made statements similar to those of Bowerman, et al., (2007). Keller (2005, p.738) further stated that the t-test cannot be used if the data are ordinal, thus eliminating its use with Likert scales. Doane and Seward (2007) recommended the use of the Wilcoxon signed-rank test in small sample situations because it is free of the normality assumption, uses ordinal data, is robust to outliers and has fairly good power over a range of non-normal population shapes. Conover (1999) differed, and stated that, as does the t-test, the Wilcoxon signed-rank test requires an interval scale of measurement that also should eliminate its use with a Likert scale. Siegel (1956) specified that the Wilcoxon signed-rank test requires a level of measurement that is between ordinal and interval, called an ordered metric scale.

All of the textbooks cited above stress that the basis for their recommendations is to be able to calculate an exact probability of making a Type I error. If the assumptions underlying any procedure are questionable then it is not possible to do so. However, it is seldom possible to completely verify that all assumptions of any procedure are totally satisfied and, sometimes, it is of more interest to protect against a Type II error than against a Type I error.

Simulation of the Data

In order to generate data that would be typical of Likert scale responses from distributions with specified means, the simulations were obtained using the method detailed in the study by Meek, et al., (2000). That is, binomial distributions were used to generate integer results from a population whose range was 0 to k-1 and had a mean of μ-1. These distributions, and the resulting data, were then shifted one unit to the right to obtain a range of observed values from 1 to k with a population mean of μ. Data were generated to represent five-point and seven-point Likert scales. A total of 27,850 simulations were conducted with 8,750 (31.4%) of them representing symmetric distributions. Because the one sample Wilcoxon signed-rank procedure is a test of the population median the symmetric cases are the only ones where it is truly appropriate, assuming that a Likert scale truly generates ordinal data. Based on Doane and Seward's (2007) statement that the Wilcoxon signed-rank procedure is fairly robust to non-normal or asymmetric shapes it should be reasonable for use on the majority of the remaining cases, too.

In addition to the level of measurement's being ordinal, at best, the underlying distributions used to generate observations were discrete, though infinite, and the actual distributions were skewed in slightly more than two-thirds of the cases rather than symmetric. Thus, in all cases, the basic assumptions of the t-test were violated while in approximately 69% of the cases at least one assumption for the Wilcoxon signed-rank test was violated. Of the 27,850 data sets on which comparisons were made 11,350 represent a five-point Likert scale and 16,500 a seven-point Likert scale with varying sample sizes of 5, 10 and 15 for both scales.

Experimental Design

Comparisons of the one sample t-test and the Wilcoxon signed-rank test for location were based on corresponding p-values and the number of incorrect decisions that resulted from each. The p-values were calculated for each test procedure's results using Minitab® and the numbers of rejections and non-rejections at various nominal significance levels were tabulated for combinations of scale size, sample size, hypothesized mean and actual mean. The numbers of rejections for each test procedure were determined by comparing the p-values to nominal significance levels of 0.01, 0.05 and 0.10. The absolute differences between the hypothesized and actual means that were evaluated were 0.0, 0.5 and 1.0. Differences greater than 1.0 were not considered because both tests were rejecting H_o with sample sizes of 10 and 15 approximately 90% of the time at that difference at the 0.10 level using a 5-point scale. A similar percentage of rejections occurred for the 7-point scale when the sample size was 15.

Two-way contingency tables were constructed for each combination by numbers of rejections and non-rejections versus the test procedure used. The Chi-square test of association was used to test for a relationship between the statistical decision and the procedure used. It is recognized that the use of the Chi-square test is questionable since both the t-test and the Wilcoxon signed-rank test were run on the same samples. The Chi-square test results do help to highlight disparities in the numbers of rejections between the two procedures. Tables were constructed identifying for which combinations significant differences occurred and at what level. It should be noted that Chi-square tests could not be run for combinations having alphas of 0.01 and 0.05 when n is five because the theoretical (expected) number of rejections by the Wilcoxon signed-rank test is zero in those cases.

Another, and possibly more informative, way of comparing the two procedures is to look at their corresponding error rates. Thus, tables were constructed to compare the error rates of the two test procedures for all of the various combinations indicated above. Because the majority of samples simulated were from asymmetric distributions a separate table was constructed showing error rates for the procedures when the actual distributions were symmetric. In a very limited number of cases (eighteen) the Wilcoxon signed-rank procedure had more rejections than the t-test. These are tabulated.

Results

One of the assumptions underlying the Wilcoxon signed-rank procedure is that the distribution is symmetric. Several comparisons were made for which this assumption is violated; for example, data on the 7-point scale when the actual mean is 2.0 or when it is 6.0. These simulations and corresponding tests were conducted to see what happens in that situation. It is recognized that any results under those conditions are questionable but, in terms of actual errors, are useful because Doane and Seward (2007) indicated that the Wilcoxon signed-rank test is robust to non-normal, and somewhat asymmetrical, population shapes. In fact, the assumptions underlying the t-test are violated in every situation because there is neither an underlying normal distribution nor an interval level of measurement. Even so, the results indicate that, in almost every case when the null hypothesis was false, the t-test performed as well or better than the Wilcoxon signed-rank test. There were a total of 13 cases in which the Wilcoxon signed-rank test rejected more times than the t-test when H_o was false and in only one of those was the difference significant, and that was at the 0.10 level.

The results of comparing the numbers of rejections for the two procedures using contingency tables are presented in Tables 1 through 6. In each of those tables, the first two columns represent the values for the hypothesized mean and the actual mean, respectively. Columns 3 and 4 identify the sample size and the number of samples generated for that combination of hypothesized and actual means. The alpha values listed at the tops of columns 5, 6 and 7 represent the nominal significance levels at which the t and Wilcoxon signed-rank tests were run. Except for Tables 1 and 4, the last three columns in each table give the levels at which the Chi-square tests comparing corresponding results of the t and Wilcoxon signed-rank procedures were significant. In Tables 1 and 4, columns 5 and 6 have asterisks entered because the Wilcoxon signed-rank test cannot reject at alphas of 0.05 and 0.01 when $n = 5$.

A brief explanation of the entries in columns 5, 6 and 7 follows. For example, an entry of NS in the column headed by $\alpha = 0.10$ indicates that the numbers of rejections by t and the Wilcoxon signed-rank test were not significantly different at a nominal alpha of 0.10 for the combination of hypothesized and actual means listed for that row. Similarly, an entry of 0.05 under the column headed by $\alpha = 0.01$ indicates that the numbers of rejections by the two tests were significantly different at the 0.05 level of significance for the set of means in that row.

As an example, in the second row of Table 3, below, where the hypothesized mean is 1.5 and the actual mean is 2.0 the t-test rejected 139 times (not presented) out of 200 runs at 0.10 while the Wilcoxon signed-rank test rejected 125 times (not presented) at 0.10. Thus, t and Wilcoxon signed-rank test each failed to reject at that level 61 and 75 times, respectively. Casting those values into a contingency table using tests as columns and decisions as rows results in a calculated Chi-square value of 2.18 which is not significant at $\alpha = 0.10$. This is the significance level, NS, entered in row 2 and column 7 of Table 3. Other than the asterisks in Tables 1 and 4, already explained above, all other entries in columns 5, 6 and 7 of Tables 1 through 6 were obtained similarly. It should be noted that corrections for 1 degree of freedom for the Chi-square test are not incorporated in Minitab®.

Tables 1, 2 and 3 correspond to data generated for a 5-point Likert scale while Tables 4, 5 and 6, given below, are for data generated on a 7-point Likert scale. All of the significant differences between the numbers of rejections for the two procedures for the 5-point scale correspond to more rejections by the t-test than by the Wilcoxon signed-rank test. From Table 1, it is seen that the t-test rejected the null hypothesis significantly more times than the Wilcoxon signed-rank test in 18 of the 25 comparisons, or 16 of 23 if we ignore the cases where H_o corresponds to a boundary value.

As the sample size increases, the cases where the numbers of rejections for the two procedures differ significantly drops correspondingly for significance levels of 0.10 and 0.05, to 8 and 12, respectively, out of 23 for $n = 10$ and 3 and 5, respectively, out of 23 for $n = 15$. However, they stay about the same for 0.01, 18 and 17 for $n = 10$ and $n = 15$, respectively. These numbers ignore boundary value cases.

As with the 5-point scale, ignoring boundary values, in the 7-point scale we see that at $\alpha = 0.10$ the significant differences decrease from 25 to 4 as the sample size increases from 5 to 15. Correspondingly, at 0.05 and 0.01, the significant differences decrease from 13 to 3 and 27 to 21, respectively, as n increases from 10 to 15.

A better way to compare the two test procedures, rather than looking at significant differences between the numbers of rejections, is to look at their estimated Type I and Type II error rates. These are presented below for all distributions in Table 7.

In Table 7, it is obvious that the Wilcoxon signed-rank procedure protects better against a Type I error because its average Type I error rate was always less than that of the t-test, whose Type I error rate exceeded the nominal significance level five times with the 7-point scale. It should be noted though that, except for $n = 15$ with the 7-point scale, the actual Type I error rate for the t-test was closer to the nominal level in all other comparisons. The average Type

Table 1: Five-point scale comparison of t test and Wilcoxon signed-rank test for a sample size of 5

Hypothesized Mean	Actual Mean	N	# of runs	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$
1.0[1]	2.0	5	100	*	*	.01
1.5	2.0	5	100	*	*	NS
2.0	2.0	5	100	*	*	.01
2.5	2.0	5	100	*	*	NS
3.0	2.0	5	100	*	*	.01
1.5	2.5	5	100	*	*	NS
2.0	2.5	5	100	*	*	.01
2.5	2.5	5	100	*	*	NS
3.0	2.5	5	100	*	*	.01
3.5	2.5	5	100	*	*	.10
2.0	3.0	5	300	*	*	.01
2.5	3.0	5	300	*	*	.10
3.0	3.0	5	300	*	*	.01
3.5	3.0	5	300	*	*	.05
4.0	3.0	5	300	*	*	.01
2.5	3.5	5	100	*	*	.05
3.0	3.5	5	100	*	*	.01
3.5	3.5	5	100	*	*	NS
4.0	3.5	5	100	*	*	.01
4.5	3.5	5	100	*	*	NS
3.0	4.0	5	100	*	*	.01
3.5	4.0	5	100	*	*	.10
4.0	4.0	5	100	*	*	.01
4.5	4.0	5	100	*	*	NS
5.0[1]	4.0	5	100	*	*	.01

[1] When the hypothesized value equals a boundary value a better test would be to reject H_o if any other value occurs in the sample.

* H_o cannot be rejected at a significance level of 0.05 or 0.01 for samples of size 5 using Wilcoxon signed-rank test.

Table 2: Five-point scale comparison of t test and Wilcoxon signed-rank test for a sample size of 10

Hypothesized Mean	Actual Mean	n	# of runs	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$
1.0[1]	2.0	10	100	.01	.01	.01
1.5	2.0	10	100	.05	.05	NS
2.0	2.0	10	100	NS	.10	NS
2.5	2.0	10	100	.01	NS	NS
3.0	2.0	10	100	.01	.10	NS
1.5	2.5	10	100	.01	NS	NS
2.0	2.5	10	100	.05	NS	.10
2.5	2.5	10	100	NS	NS	NS
3.0	2.5	10	100	.01	.10	.10
3.5	2.5	10	100	.01	NS	NS
2.0	3.0	10	300	.01	.01	NS
2.5	3.0	10	300	.01	NS	NS
3.0	3.0	10	450	.05	.05	.05
3.5	3.0	10	300	.01	.10	.10
4.0	3.0	10	300	.01.	.01	.05
2.5	3.5	10	100	NS	NS	NS
3.0	3.5	10	100	.01	.05	NS
3.5	3.5	10	100	NS	NS	NS
4.0	3.5	10	100	.10	.05	.10
4.5	3.5	10	100	.01	NS	NS
3.0	4.0	10	100	.01	.01	.05
3.5	4.0	10	100	.05	NS	NS
4.0	4.0	10	100	NS	.10	.10
4.5	4.0	10	100	.05	NS	NS
5.0[1]	4.0	10	100	.01	.05	.05

[1] When the hypothesized value equals a boundary value a better test would be to reject H_o if any other value occurs in the sample.

Table 3: Five-point scale comparison of t test and Wilcoxon signed-rank test for a sample size of 15

Hypothesized Mean	Actual Mean	n	# of runs	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$
1.0[1]	2.0	15	200	.01	NS	NS
1.5	2.0	15	200	.10	.10	NS
2.0	2.0	15	200	NS	NS	NS
2.5	2.0	15	200	.10	NS	NS
3.0	2.0	15	300	.01	NS	NS
1.5	2.5	15	100	.05	NS	NS
2.0	2.5	15	100	.01	NS	NS
2.5	2.5	15	100	NS	NS	NS
3.0	2.5	15	100	.01	NS	NS
3.5	2.5	15	100	.05	NS	NS
2.0	3.0	15	300	.01	NS	NS
2.5	3.0	15	300	.01	NS	NS
3.0	3.0	15	400	NS	.05	NS
3.5	3.0	15	300	.01	.05	.05
4.0	3.0	15	300	.01	NS	NS
2.5	3.5	15	100	.01	NS	NS
3.0	3.5	15	100	.01	NS	NS
3.5	3.5	15	100	NS	NS	NS
4.0	3.5	15	100	.01	.10	NS
4.5	3.5	15	100	.05	NS	NS
3.0	4.0	15	100	.01	.10	NS
3.5	4.0	15	100	.05	NS	NS
4.0	4.0	15	100	NS	NS	.05
4.5	4.0	15	100	NS	NS	.10
5.0[1]	4.0	15	100	.01	NS	NS

[1] When the hypothesized value equals a boundary value a better test would be to reject Ho if any other value occurs in the sample.

Table 4: Seven-point scale comparison of t test and Wilcoxon signed-rank test for a sample size of 5

Hypothesized Mean	Actual Mean	N	# of runs	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$
1.0[1]	2.0	5	100	*	*	.01
1.5	2.0	5	100	*	*	NS
2.0	2.0	5	100	*	*	.01
2.5	2.0	5	100	*	*	NS
3.0	2.0	5	100	*	*	.01
1.5	2.5	5	100	*	*	NS
2.0	2.5	5	100	*	*	.01
2.5	2.5	5	100	*	*	NS
3.0	2.5	5	100	*	*	.01
3.5	2.5	5	100	*	*	NS
2.0	3.0	5	100	*	*	.01
2.5	3.0	5	100	*	*	NS
3.0	3.0	5	100	*	*	.01
3.5	3.0	5	100	*	*	NS
4.0	3.0	5	100	*	*	.01
2.5	3.5	5	100	*	*	.10
3.0	3.5	5	100	*	*	.01
3.5	3.5	5	100	*	*	NS
4.0	3.5	5	100	*	*	.01
4.5	3.5	5	100	*	*	.05
3.0	4.0	5	300	*	*	.01
3.5	4.0	5	300	*	*	.05
4.0	4.0	5	300	*	*	.01
4.5	4.0	5	300	*	*	..05
5.0	4.0	5	300	*	*	.01
3.5	4.5	5	100	*	*	NS
4.0	4.5	5	100	*	*	.01
4.5	4.5	5	100	*	*	NS
5.0	4.5	5	100	*	*	.01
5.5	4.5	5	100	*	*	NS
4.0	5.0	5	100	*	*	.01
4.5	5.0	5	100	*	*	NS
5.0	5.0	5	100	*	*	NS
5.5	5.0	5	100	*	*	NS
6.0	5.0	5	100	*	*	.01
4.5	5.5	5	100	*	*	.05
5.0	5.5	5	100	*	*	.01
5.5	5.5	5	100	*	*	NS
6.0	5.5	5	100	*	*	.01
6.5	5.5	5	100	*	*	NS
5.0	6.0	5	100	*	*	.01
5.5	6.0	5	100	*	*	NS
6.0	6.0	5	100	*	*	.01
6.5	6.0	5	100	*	*	NS
7.0[1]	6.0	5	100	*	*	.01

[1] When the hypothesized value equals a boundary value a better test would be to reject H_o if any other value occurs in the sample.

* Ho cannot be rejected at a significance level of 0.05 or 0.01 for samples of size 5 using Wilcoxon signed-rank test.

Table 5: Seven-point scale comparison of t test and Wilcoxon signed-rank test for a sample size of 10

Hypothesized Mean	Actual Mean	N	# of runs	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$
1.0[1]	2.0	10	100	.01	.01	NS
1.5	2.0	10	100	NS	NS	NS
2.0	2.0	10	100	NS	.05	NS
2.5	2.0	10	100	.01	NS	NS
3.0	2.0	10	100	.01	.05	NS
1.5	2.5	10	100	.01	NS	NS
2.0	2.5	10	100	NS	.05	NS
2.5	2.5	10	100	NS	NS	NS
3.0	2.5	10	100	.01	.05	NS
3.5	2.5	10	100	.01	NS	NS
2.0	3.0	10	100	.01	.05	NS
2.5	3.0	10	100	NS	NS	NS
3.0	3.0	10	100	NS	NS	NS
3.5	3.0	10	100	NS	NS	NS
4.0	3.0	10	100	.01	.05	NS
2.5	3.5	10	100	.01	.01	NS
3.0	3.5	10	100	NS	.10	NS
3.5	3.5	10	100	NS	NS	NS
4.0	3.5	10	100	.01	.05	NS
4.5	3.5	10	100	.01	NS	NS
3.0	4.0	10	200	.01	.01	.10
3.5	4.0	10	200	.01	NS	NS
4.0	4.0	10	200	NS	NS	NS
4.5	4.0	10	200	.05	NS	NS
5.0	4.0	10	200	.01	NS	NS
3.5	4.5	10	100	.01	NS	NS
4.0	4.5	10	100	.05	NS	NS
4.5	4.5	10	100	NS	NS	NS
5.0	4.5	10	100	.05	NS	NS
5.5	4.5	10	100	.01	NS	NS
4.0	5.0	10	200	.01	.01	NS
4.5	5.0	10	200	.01	NS	NS
5.0	5.0	10	200	NS	NS	.10
5.5	5.0	10	200	NS	NS	NS
6.0	5.0	10	200	.01	.05	.05
4.5	5.5	10	100	.01	NS	NS
5.0	5.5	10	100	.01	NS	NS
5.5	5.5	10	100	NS	NS	NS
6.0	5.5	10	100	.10	.10	NS
6.5	5.5	10	100	.01	NS	NS
5.0	6.0	10	100	.01	NS	NS
5.5	6.0	10	100	.01	NS	NS
6.0	6.0	10	100	NS	NS	.10
6.5	6.0	10	100	NS	NS	NS
7.0[1]	6.0	10	100	.01	.05	.05

[1] When the hypothesized value equals a boundary value a better test would be to reject H_o if any other value occurs in the sample.

Table 6: Seven-point scale comparison of t test and Wilcoxon signed-rank test for a sample size of 15

Hypothesized Mean	Actual Mean	N	# of runs	$\alpha = .01$	$\alpha = .05$	$\alpha = .1$
1.0[1]	2.0	15	100	.01	NS	NS
1.5	2.0	15	100	NS	NS	NS
2.0	2.0	15	100	NS	NS	NS
2.5	2.0	15	100	NS	NS	.10
3.0	2.0	15	100	.10	NS	NS
1.5	2.5	15	100	NS	NS	.10
2.0	2.5	15	100	.05	NS	NS
2.5	2.5	15	100	NS	NS	NS
3.0	2.5	15	100	NS	NS	NS
3.5	2.5	15	100	.10	NS	NS
2.0	3.0	15	100	.01	NS	.10
2.5	3.0	15	100	NS	NS	NS
3.0	3.0	15	100	NS	NS	NS
3.5	3.0	15	100	NS	NS	NS
4.0	3.0	15	100	.01	NS	NS
2.5	3.5	15	100	.05	NS	NS
3.0	3.5	15	100	.05	NS	NS
3.5	3.5	15	100	NS	NS	NS
4.0	3.5	15	100	NS	NS	NS
4.5	3.5	15	100	.05	NS	NS
3.0	4.0	15	300	.01	NS	NS
3.5	4.0	15	300	.05	NS	NS
4.0	4.0	15	300	NS	.10	NS
4.5	4.0	15	300	.05	NS	.10
5.0	4.0	15	300	.01	.05	NS
3.5	4.5	15	100	NS	NS	NS
4.0	4.5	15	100	.10	NS	NS
4.5	4.5	15	100	NS	NS	NS
5.0	4.5	15	100	.05	NS	NS
5.5	4.5	15	100	.10	.01	NS
4.0	5.0	15	100	.05	NS	NS
4.5	5.0	15	100	NS	NS	NS
5.0	5.0	15	100	NS	NS	NS
5.5	5.0	15	100	NS	NS	NS
6.0	5.0	15	100	.01	NS	NS
4.5	5.5	15	100	.10	NS	NS
5.0	5.5	15	100	NS	NS	NS
5.5	5.5	15	100	NS	NS	NS
6.0	5.5	15	100	NS	NS	NS
6.5	5.5	15	100	NS	NS	NS
5.0	6.0	15	100	.05	NS	NS
5.5	6.0	15	100	.10	NS	NS
6.0	6.0	15	100	.05	NS	NS
6.5	6.0	15	100	NS	NS	NS
7.0[1]	6.0	15	100	NS	NS	NS

[1] When the hypothesized value equals a boundary value a better test would be to reject H_o if any other value occurs in the sample.

Table 7: Average error rates (%)

				WSR test average @ α of			t-test average @ α of		
$\Delta = \mu_o - \mu_a$	n	Error Type	Runs	0.01	0.05	0.10	0.01	0.05	0.10
Five	**Point**	**Scale**							
-1.0[1]	5	II	600	*	*	74.2	89.0	59.0	38.2
-0.5	5	II	700	*	*	88.4	98.4	88.1	76.4
0.0	5	I	700	*	*	1.9	0.1	3.7	9.0
+0.5	5	II	700	*	*	87.6	98.4	87.0	76.4
+1.0[1]	5	II	600	*	*	73.2	87.3	59.0	36.5
-1.0[1]	10	II	600	84.7	32.2	14.7	47.5	17.5	9.0
-0.5	10	II	700	99.1	83.6	68.4	92.3	75.4	63.7
0.0	10	I	850	0.0	1.9	5.1	1.2	4.1	9.5
+0.5	10	II	700	96.9	78.4	65.0	88.0	70.1	57.6
+1.0[1]	10	II	600	82.5	30.5	11.8	48.0	19.2	7.3
-1.0[1]	15	II	600	39.3	7.0	2.2	20.0	4.2	1.0
-0.5	15	II	800	89.8	62.0	46.6	79.9	54.8	41.9
0.0	15	I	900	0.3	2.1	6.0	0.9	4.6	8.9
+0.5	15	II	800	87.6	60.5	44.8	76.4	53.4	40.0
+1.0[1]	15	II	700	38.1	7.3	3.3	16.9	4.6	1.7
Seven	**Point**	**Scale**							
-1.0[1]	5	II	1000	*	*	76.0	89.1	70.3	51.5
-0.5	5	II	1100	*	*	88.7	97.7	87.9	78.6
0.0	5	I	1100	*	*	3.5	0.6	**5.4**	**11.5**
+0.5	5	II	1100	*	*	89.5	97.8	88.9	79.8
+1.0[1]	5	II	1000	*	*	77.2	90.3	70.4	53.8
-1.0$_{(1)}$	10	II	1000	88.2	44.2	28.1	61.7	35.4	23.1
-0.5	10	II	1100	98.8	83.2	69.6	91.6	76.8	63.7
0.0	10	I	1100	0.0	3.0	6.5	0.9	4.7	**10.2**
+0.5	10	II	1100	98.4	84.9	72.9	92.6	79.2	68.6
+1.0[1]	10	II	1000	89.7	47.1	29.4	66.9	39.0	23.9
-1.0[1]	15	II	1000	58.0	21.9	12.3	42.8	17.3	10.4
-0.5	15	II	1100	93.1	72.5	58.5	86.5	67.1	53.0
0.0	15	I	1100	0.5	4.8	9.3	1.8	**6.4**	**11.4**
+0.5	15	II	1100	91.8	73.3	58.7	87.5	69.4	54.8
+1.0[1]	15	II	1000	60.1	22.3	11.4	43.7	17.3	9.6

[1] If the hypothesized value equals a boundary value the result was deleted since a better test is to reject H_o if any value other than μ_o occurs in the sample.

*Ho cannot be rejected at a nominal α value of either 0.01 or 0.05 for a sample of size 5.

Bold-faced entries indicate cases where the nominal α was exceeded.

II error rates using the t-test were lower than those using the Wilcoxon signed-rank test for every set of mean differences. Of the total of 630 combinations of mean comparisons and significance levels, the Wilcoxon signed-rank test rejected more times than the t-test when H_o was false in only 13 combinations and the difference was significant, at 0.10, in only one of those. Because many of these cases involve distributions that are not symmetric and means are compared rather than medians it may not be fair to compare the Wilcoxon signed-rank test's Type II error rates to those of the t-test, even though at least two of the t-test's assumptions are violated in every case. Therefore, error rates for cases involving only symmetric distributions, where means and medians are the same, are presented in Table 8 below.

As can be seen in Table 8 the pattern of error rates is very similar to that shown in Table 7 for all distributions. That is, even though the Wilcoxon signed-rank test's assumptions are satisfied in all cases, assuming data generated on a Likert scale can be considered ordinal, and the t-test's assumptions are not satisfied in any cases the average Type II error for the t-test is smaller than that of the Wilcoxon signed-rank test. As before, the Wilcoxon signed-rank test protects better against a Type I error with smaller average error rates. Surprisingly, and contrary to popular belief, the t-test, even though its assumption are violated, appears to protect substantially better against Type II errors for sample sizes of 5 and for larger mean differences at a significance level of 0.01. This phenomenon occurred for both symmetric and non-symmetric distributions. For the cases involving distributions that were not symmetric there did not seem to be any substantial differences between distributions that were skewed to the left from those that were skewed to the right.

In all, 630 combinations of means, sample sizes, Likert scales and nominal α-levels were run. The Wilcoxon signed-rank test rejected more times than the t-test in only eighteen of those combinations and only once did it do so significantly. The combination for which the Wilcoxon signed-rank test rejected significantly more times than the t-test at a level of 0.10 was: $n = 15$, 7-point scale, hypothesized mean of 2.5 vs. actual mean of 2.0 and a nominal α of 0.10. Cases for which Wilcoxon signed-rank test rejections exceeded t rejections are given in Table 9. Of the 18 cases in Table 9, five correspond to cases where H_o was true and 13 to cases where H_o was false.

Table 8: Average error rates (in %) for symmetric distributions

					WSR test error @ $\alpha =$			t-test error @ $\alpha =$		
Scale	N	Runs	$\Delta=\mu_o-\mu_a$	**Type** Error	0.01	0.05	0.10	0.01	0.05	0.10
5	5	300	-1.0	II	*	*	87	92	62	41
5	5	300	-0.5	II	*	*	87	99	92	82
5	5	300	0.0	I	*	*	0	0	4	10
5	5	300	0.5	II	*	*	84	98	87	78
5	5	300	1.0	II	*	*	84	87	59	35
5	10	300	-1.0	II	89	35	16	49	20	12
5	10	300	-0.5	II	99	81	67	92	76	70
5	10	450	0.0	I	0	2	5	1	4	9
5	10	300	0.5	II	97	78	67	90	72	61
5	10	300	1.0	II	88	33	12	53	19	7
5	15	300	-1.0	II	45	7	2	23	4	1
5	15	300	-0.5	II	92	67	52	84	61	48
5	15	400	0.0	I	0	2	6	1	5	8
5	15	300	0.5	II	90	64	49	78	56	42
5	15	300	1.0	II	42	7	3	18	5	2
7	5	300	-1.0	II	*	*	86	91	71	51
7	5	300	-0.5	II	*	*	85	98	89	77
7	5	300	0.0	I	*	*	1	1	5	11[1]
7	5	300	0.5	II	*	*	90	98	94	84
7	5	300	1.0	II	*	*	91	95	75	62
7	10	200	-1.0	II	92	49	28	64	36	20
7	10	200	-0.5	II	99	85	72	91	79	66
7	10	200	0.0	I	0	4	7	1	6[1]	10
7	10	200	0.5	II	98	85	75	94	82	69
7	10	200	1.0	II	93	38	27	66	35	21
7	15	300	-1.0	II	65	23	15	44	18	12
7	15	300	-0.5	II	94	75	62	89	71	56
7	15	300	0.0	I	0	3	9	1	5	11[1]
7	15	300	0.5	II	93	73	61	88	69	54
7	15	300	1.0	II	65	23	11	45	16	8

[1] Identifies cases where the nominal α was exceeded.

* Indicates cases where n is too small for Wilcoxon signed-rank test to reject at the nominal significance level.

Table 9: Cases where Wilcoxon signed-rank rejections exceed t rejections

Scale	N	Runs	Actual μ	Hypoth. μ	Nom. α	WSR rej.	t-test rej.	Sig.
5	5	100	2.5	2.5	0.10	6	2	NS
5	10	300	3.0	2.5	0.10	99	91	NS
5	15	200	2.0	2.5	0.10	115	114	NS
5	15	100	2.5	2.5	0.10	6	4	NS
5	15	100	4.0	3.5	0.10	66	65	NS
7	5	100	2.5	2.5	0.10	9	8	NS
7	5	100	5.5	5.5	0.10	9	8	NS
7	5	100	6.0	6.5	0.10	14	11	NS
7	10	100	2.0	2.5	0.10	55	50	NS
7	10	100	5.5	4.5	0.10	64	63	NS
7	10	100	6.0	5.5	0.10	49	48	NS
7	15	100	2.0	2.5	0.10	67	55	0.10
7	15	100	2.0	2.5	0.05	51	47	NS
7	15	100	2.0	3.0	0.05	84	81	NS
7	15	100	2.5	2.5	0.10	17	13	NS
7	15	100	5.5	4.5	0.10	88	87	NS
7	15	100	5.5	4.5	0.05	81	79	NS
7	15	100	6.0	5.5	0.05	42	39	NS

Conclusion

Based on the 27,850 simulations conducted in this study, of which 8,750 involved symmetric distributions, it appears that the t-test may be preferred over the signed-rank procedure, even for very small sample sizes, unless it is imperative that one be able to calculate the exact probability of committing a Type I error. As in the Meek, et al., (2000) and (2001) studies, the Blair and Higgins (1985) study, and the Nanna (2002) study the level of measurement does not appear to be an important factor in test selection, at least in the case of a Likert scale. A more important consideration, at least with respect to the one sample test of location, is which error is more critical to guard against. The limitations of this study are that all data were generated from binomial distributions, the assumptions for the t-test are violated in all cases and the symmetry assumption of the signed-rank test is violated in 69% of the cases. Even with these limitations the t-test showed a lower average Type II error rate across all of the sample sizes that were used in the study. Contrary to what was expected, based on the literature, the t-test was much better at protecting against a Type II error for the sample size of five than was the signed-rank test, even when the Wilcoxon signed-rank test's assumptions were all satisfied. As the sample size increased the number of significant differences between the two procedures decreased dramatically for the 0.10 and 0.05 significance levels, to the point that the tests had similar error rates for those significance levels when n = 15. Although the results of this study seem to be in conflict with those of Blair and Higgins (1985), Nanna and Sawilowsky (1998), and Nanna (2002) their studies involved testing either two populations or the multivariate case and used different underlying distributions.

In summary, the results of these simulations indicated:

1. Except for a sample size of 5, the numbers of significant differences were fewest at a nominal α of 0.10 while significant differences decreased for both 0.10 and 0.05 as the sample size increased, but not for 0.01;

2. The t-test tended to have a higher Type I error rate, but closer to the nominal value, on average, while the Wilcoxon signed-rank test had a higher Type II error rate;

3. There did not appear to be any dramatic differences between error rates when the distributions were symmetric as opposed to being asymmetric;

4. The t-test actually appears to reject false hypotheses better; i.e., to have higher power, than the Wilcoxon signed-rank test when the sample sizes are small, even though its assumptions are violated in every case; and,

5. This study appears to contradict statements and recommendations about the use of the t-test vs. the Wilcoxon signed-rank test in small sample applications involving these particular non-normal distributions and ordinal data.

Further study needs to be done using different types of underlying distributions to generate the data to determine if these results might be attributed to having used a binomial generator. Additional points that might be considered in the future are other Likert scales, such as a 9-point, and ordinal measurements that do not correspond to Likert scale data. Regardless of this study's limitations it is quite surprising to find that all of the recommendations in the literature for using Wilcoxon's signed-rank procedure over the t-test, particularly with small sample sizes and Likert scale data, appear to be groundless, even when the t-test's assumptions are violated. Under conditions similar to the ones in this study it seems the only justification for using the Wilcoxon signed-rank procedure over the t-test is that it be imperative that an exact Type I error be able to be calculated.

References

Anderson, D., Sweeney, D. & Williams, T. (2005) *Statistics for business and economics* (9th ed.). Mason, OH: Southwestern Publishing Co.

Baggaley, Andrew R. (1960). Some remarks on scales of measurement and related topics. *The Journal of General Psychology, 62,* 141-145.

Berenson, M., Levine, D. & Kreihbel, T. (2004). *Basic business statistics: concepts and application*, (9th ed.). Englewood Cliffs, NJ: Prentice-Hall Publishing Co.

Blair, R. C., & Higgins, J. J. (1985). Comparison of the power of the paired samples t test to that of Wilcoxon's signed-ranks test under various population shapes. *Psychological Bulletin*, *97*(1), 119-128.

Bowerman, B. & O'Connell, R. (2007). *Business statistics in practice*. New York: McGraw-Hill Irwin Publishing Co.

Bradley, J. V. (1968). *Distribution free statistical tests*. Englewood-Cliffs, NJ: Prentice-Hall Publishing Co.

Bradley, J. V. (1980). Nonrobustness in one-sample Z and t tests: A large-scale sampling study. *Bulletin of the Psychonomic Society*, *15*, 29-32.

Chou, Y. L. (1989). *Statistical analysis for business and economics*. New York: Elsevier Science Publishing Co.

Conover, W. J. (1971). *Practical nonparametric statistics*. New York: John Wiley & Sons.

Doane, D. & Seward, L. (2007) *Applied statistics in business and economics*. New York: McGraw-Hill Irwin Publishing Co.

Harris, R. J. (1975). *A primer of multivariate statistics*. New York: Academic Press.

Heeren, T.& D'Agostino, R. (1987). Robustness of the two independent samples when applied to ordinal scale data. *Statistics in Medicine*, *6*, 79-90

Keller, G. (2005). *Statistics for management and economic* (7th ed.). Belmont, CA: Thomson Publishing Co.

Meek, G., Ozgur, C. & Dunning, K. (2000). Does scale of measurement really make a difference in test selection: An empirical comparison of t test vs. Mann Whitney. *Proceedings of the 2000 National Annual Meeting of the Decision Sciences Institute*, 951-953.

Meek, G., Ozgur, C. & Dunning, K. (2001). Does scale of measurement really make a difference in test selection: The one-sample case, t vs. Wilcoxon signed-rank test. *Proceedings of the 2001 National Annual Meeting of the Decision Sciences Institute*, 1280 (abstract).

Moore, D., McCabe, G., Duckworth, W. & Sclove, S. (2003) *The practice of business statistics: using data for decisions*. New York, NY: W. H. Freeman and Co.

Nanna, M. J. (2002). Hotellings T^2 vs the rank transform with real Likert data. *Journal of Modern Applied Statistical Methods*, *1*(1) 83-99.

Nanna, M. J., & Sawilowsky, S. S. (1998). Analysis of Likert scale data in disability and medical rehabilitation research. *Psychological Methods*, *3*(1), 55-67.

Newbold, P. (1995). *Statistics for business and economics* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Russel, C. J. & Bobko, P. (1992). Moderated regression analysis and Likert scales: Too coarse. *Journal of Applied Psychology, 77,* (3), 336-342.

Sawilowsky, S. S. (1991). Comments on using alternatives to normal theory statistics in social and behavioral sciences. *Canadian Psychology*, *34*(4), 432-433.

Sawilowsky, S. S. (2005). Misconceptions leading to choosing the t test over the Wilcoxon Mann-Whitney test for shift in location parameter. *Journal of Modern Applied Statistical Methods*, *4*(2), 598-600.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research*, *60*(1), 91-126.

Sawilowsky, S. S. & Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin***,** *111*, 352-360

Siegel, S. (1956) *Nonparametric Statistics for the Behavioral Sciences*. New York, NY: McGraw-Hill.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677-680.