ROBUSTNESS AND POWER OF THE T, PERMUTATION T AND WILCOXON TESTS

by

MICHELE WEBER

DISSERTATION

Submitted to the Graduate School

Of Wayne State University,

Detroit, Michigan

In partial fulfillment of the requirements

For the degree of

DOCTOR OF PHILOSOPHY

2006

MAJOR: EDUCATIONAL EVALUATION AND RESEARCH

Approved by:

Stevo Sawilowsky 9/14/06

Advisor Date

Donald R Marcoulides

Gregg Jacobs

UMI Number: 3225885

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

ACKNOWLEDGEMENTS

I want to thank all the people involved in my endeavor from the bottom of my heart. In the first place, I give thanks to the Lord Almighty for, without Him, nothing is possible. My husband Rolf Weber played a tremendous role in my accomplishment: A special thanks for his immense help and support. I also want to express my gratitude to Brian Dates who told me about this program and for his support. I am also grateful for my committee: Dr. Donald Marcotte, Dr. Gail Fahoome, Dr. Royce Hutson who agreed to be a member of my committee so late in the game and, Dr. Shlomo Sawilowsky who was such a great advisor. To all the friends and family who are not named, many thanks for all the support and prayers.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Statistics are fundamental for the development of systematic inquiry in the physical, behavioral and social sciences such as in biology, medicine, psychology, education, social work, and business (Fisher, 1958; Wilcox, 1996). Among the reasons that make them essential, statistical methods assist researchers in finding meaningful descriptions of populations and their subsets, and provide methods for comparing groups in the attempt to discover effective treatments (Weinbach & Grinnell, 1997; Wilcox, 1996). Two primary types of statistical methods are parametric and nonparametric.

Parametric Methods

Parametric tests are one of the most common choices of statistical tests used in research. They investigate hypotheses using probabilities in which the population parameters fulfill specific premises in regards to the samples drawn from that population (Kerlinger & Lee, 2000). They depend on certain assumptions that need to be met (Hinkle, Wiersma & Jurs, 2003; Kerlinger & Lee, 2000; Sawilowsky & Fahoome, 2003). These requirements assume random assignment and random selection of the data (Berger, Lunneborg, Ernst & Levine, 2002; Kerlinger & Lee, 2000; Ludbrook & Dudley, 1998; Neyman & Pearson, 1928; Weinbach & Grinnell, 1997). It is also assumed that the observations are independent of each other, the variance is

homogeneous, and that the population is normally distributed (Bradley, 1968; Fahoome & Sawilowsky, 2000; Kerlinger & Lee, 2000; Zimmerman, 1998).

When these assumptions are met, parametric tests provide important information in regards to the analysis of the data at hand. They also allow the testing of complex hypotheses (Hinkle et al., 2003). Examples of parametric tests include the t-test, analysis of variance, and ordinary least squares multiple regression (Hair, Anderson, Tatham & Black, 1998; Kerlinger & Lee, 2000).

Robustness indicates the insensitivity against small deviations in the assumptions (Huber, 1977). A robust procedure is designed to reduce the sensitivity of the parameter estimates to failure in the assumptions (Huber, 2003). A test is said to be robust if the Type I error is close to the pre-determined α (Man, Wang & Wang, 2000). Type I error is defined as the likelihood to reject a true null hypothesis (Hinkle, et al., 2003).

Type II error is the probability of not rejecting a false null hypothesis (Hinkle, et al., 2003). It is inversely related to the Type I error (Hair, et al., 1998). Statistical power is the probability of rejecting a false null hypothesis (Hinkle, et al., 2003; Hogg & Craig, 1995). Power relates to the Type II error (Hair, et al., 1998; Hinkle, et al., 2003; Hogg & Craig, 1995). It is determined as a function of the significance level α set by the researcher for a Type I error, the sample size used in the analysis and the examined effect size, which is the estimation of the degree to which the studied phenomenon exists in the population (Hair, et al., 1998).

As assumptions, such as normality and homogeneity, are violated, parametric tests become less robust and less powerful (Man, et al., 2000). The two sample tests of difference in means, the t-test, is the uniformly most powerful unbiased test, but only when its underlying assumptions such as normality have been met (Bradley, 1968; Good, 1994; Kerlinger & Lee, 2000). Under certain non-normal situations and conditions, it is still robust in terms of Type I error and powerful (Boneau, 1960; Glass, Peckham & Sanders, 1972; Sawilowsky & Blair, 1992; Zimmerman, 1998). However, , when the equal variance assumption is violated, the t-test does not maintain its significance level (Rogan & Keselma, 1977; Zaremba, 1965; Zimmerman, 1996; Zimmerman, 1998). Similarly, the test is less robust for testing two tailed hypotheses than one-tailed tests, and for small samples sizes if the group with the smaller n has the larger variance or the reverse, i.e. a larger n with a smaller variance (Sawilowsky & Blair, 1992).

Nonparametric Methods

A nonparametric procedure tests hypotheses without any appeal to population parameters (Sawilowsky & Fahoome, 2003). According to Walsh (1968), a test is considered nonparametric when its Type I error properties are satisfied when assumptions such as normality do not hold. According to Sawilowsky (1990), there are at least three types of nonparametric tests, which include categorical, sign and ranks. The Spearman's rank correlation coefficient, the Chi-Square and the Wilcoxon sign-rank tests are examples of

nonparametric tests (Hinkle et al., 2003; Neave & Worthington, 1988; Wilcox, 1996).

Nonparametric tests have been used in applied data analysis in increasing frequencies in the past half century when the assumptions for the parametric tests were not met. The reason is because they are good alternatives to parametric methods under non-normality (Lehmann, 1975; Marascuilo & McSweeney, 1977). Blair and Higgins (1985b) and Sawilowsky (1990) demonstrated that they are not merely alternate to parametric methods but they actually outperform their parametric counterparts under many non-normal situations when testing for shift in location. However, nonparametric tests should be used because of their own merits, meaning their use should not be restricted to situations where parametric assumptions fail.

Although nonparametric tests do not depend on normality, the random data selection, the independence of observations and the continuous distribution of the data remain assumptions (Kerlinger & Lee, 2000; Sawilowsky & Fahoome, 2003). Nonparametric tests are by definition robust with regard to Type I error. The Wilcoxon rank sum test, for example, is more powerful than the parametric t-test when the population's distribution is non-normal (Blair & Higgins, 1985; Hodges & Lehmann, 1956; Sawilowsky, 1990; Sawilowsky & Fahoome, 2003; Zimmerman & Zumbo, 1989). Indeed, it can be up to four times more powerful than the t-test when the data is sampled from an exponential distribution (Sawilsowsky & Blair, 1992).

With some rank nonparametric methods, it has been incorrectly assumed that these tests lack statistical power and there is some loss of information when the data are being transformed to ranks (Adams & Anthony, 1996; Borg, 1987; Chase, 1976; Garrett, 1966). However, Wolfowitz (1949) stated that this possible loss of information due to the ranking is not important because this information is not available in the actual data. Moreover, Blair and Higgins (1985b) and Blair, Higgins and Smitley (1980) showed that converting the data into ranks does, in fact, not lose the important information (Sawilowsky, 1990). Indeed, it has been demonstrated through Monte Carlo studies that, under non-normal situations, nonparametric rank tests perform better than their parametric counterparts in terms of statistical power.

Permutation Techniques

Modern permutation tests were first developed at least by 1935 by Fisher in experiments conducted with soil (Mielke & Berry, 2001). The permutation technique was further developed by Pitman in 1937 and 1938 (Mielke & Berry, 2001). The advance in computing technology has favorably impacted the use of permutation tests because they provide a cost effective and time efficient method in generating permutations (Edgington, 1995; Good, 1994).

Permutation tests can be used for testing of hypotheses where a specific relationship between two sets of random variables exists (Edgington, 1995; Noreen, 1989). There are different types of permutation tests: (1) the exact permutation test involves the calculation of the chosen test statistic for

all possible permutations between the observed data sets, (2) the approximate randomization, also known as resampling or sampling permutation, examines a relatively smaller subset of all possible permutations, and (3) the moment approximation test involves the continuous probability density function which is based on the exact lower moments of the chosen test statistic fitted to the discrete permutation distribution (Mielke & Berry, 2001).

Permutation techniques have fewer assumptions than parametric methods (Heckel, Arndt, Cizadlo, Andreasen, 1996; Manly, 1995). The random selection continues to be a requirement (Kempthorne, 1966) for generality. Although Noreen (1989) stated that the samples do not need to be randomly selected to produce internally valid results, the lack of randomization precludes making inferences to a population. Permutation techniques are distribution free statistics and can be used for normal and non-normal distributions (Bradley, 1968; Edgington, 1995; Maritz, 1981; Mielke & Berry, 2001). For independent tests, the observations also need to be independent of each other (Good, 1994; Maritz, 1981). The observations have to be exchangeable, i.e. they are identically distributed and the probability of any outcome is independent of the observations' order (Boik, 1987; Commenges, 2003; Good, 2002; Lu, Chase & Li, 2001; Opdyke, 2003). Other requirements are the continuity of the distributions (Edgington, 1995) and homogeneity of variance (Boik, 1987) when using a permutation test for differences in location parameters.

Permutation techniques create their own sampling distribution from the data at hand (Manly, 1995; McArdle & Anderson, 2004). By reshuffling the available data between the sets as many times as possible, a distribution of possible outcomes is created, generating these data's own sampling distribution (Adams & Anthony, 1996). According to Sawilowsky (1990), Lehmann and Stein (1959) showed that permutation methods are as efficient and as powerful as the t-test under normality. Overall, permutation tests are as powerful as their parametric counterparts when the parametric assumptions are met (Noreen, 1989). Lehmann and D'Abrera (1975) stated that permutation tests are more robust than parametric tests under non-normality.

An example of how the permutation method may be carried out is as follows: After two or more sets of data are collected and divided into a treatment and a control group, a test statistic is chosen. The data are, then, permuted between the two groups with all possible n!/(m! k!) combinations (n = total number of observations, m = number of treatment group observations, k = number of control group observations, n = m + k and n! = n (n-1) (n-2) …1) and the test statistic value is computed and recorded for every resulting permutation.

The test statistic values are then sorted from the highest to the lowest value and, for example, the 95^{th} percentile is determined. This percentile is the chosen critical value, which represents the P-value (Adams & Anthony, 1996; Edgington, 1995; Good, 1994). The test statistic original value prior to the permutations is then compared to that critical value. If the original statistic

value prior to the permutations it is greater than or equal to the critical value (or in certain cases of test statistics – less than or equal to), the null hypothesis is rejected.

Permutation methods offer advantages over parametric procedures. There is no need to refer to statistical tables in order to retrieve critical values because the permutation test gives the critical value based on the data at hand (Edgington, 1995; Mielke & Berry, 2001). This permutation technique is also used to obtain the exact p-value or P-value (Good, 1994). Permutation tests offer an advantage dealing with possible outliers by most likely detecting the difference in means with outliers (Edgington, 1995).

Nevertheless, the amount of computation involved in permutation techniques requires access to fairly fast computers (Noreen, 1989). Although books were available, researchers needed to have a minimum knowledge of programming languages such as Fortran 90 and PASCAL to write the programs required for these computer-intensive methods (Noreen, 1989). However, these disadvantages are no longer a major issue for researchers who want to apply permutations in their analyses. In fact nowadays, a few programming software are readily available, such as StatXact, Strata 9 and exact modules in SPSS. But an inconvenience remains, which relates to the amount of time required for all possible permutations. Indeed, as the sample size increases, the number of permutations increases exponentially which can become an unpractical amount of time.

Edgington (1995) and Good (1994) posited that a strength of permutation is to analyze data at hand. Mielke and Berry (2001, p.3) stated that permutation tests are “data-dependent” tests.

Nevertheless, from a research design perspective, it may create a selection bias, which indicates that the selection may be done deliberately, possibly affecting the generalization of the results (Kerlinger & Lee, 2000). Campbell and Stanley (1963) explained that with selection bias, the possibility of the effects could only be seen for that particular population. This is a valid point in regard to the internal validity, addressing the question “Did the intervention bring about the outcome?” but it precludes consideration of external validity, which is the question “Are the results generalizeable?” External validity leads to generalizability (Campbell & Stanley, 1963). Without external validity, the results obtained cannot be generalized beyond the data at hand.

Adams and Anthony (1996), Edgington (1995), Good (1994) and other authors elaborated on the advantages that permutation tests have over the parametric tests but they did not expand much on their disadvantages. Thus, it is important to reiterate the assumptions that permutation tests require, such as random selection, exchangeable observations and continuous data.

Statement of the Problem

Permutation tests maintain the Type I error to the significance level of nominal α (Edgington, 1995; Mielke & Berry, 2001, Sawilowsky, 2004) even for nonrandom selected samples. Many authors, such as Edginton (1995) and

Good (1994) assume that they are superior in terms of comparative power as compared with nonparametric procedures. However, there is no evidence to support this assertion.

Through a Monte Carlo study, Type I error rate and power are investigated. Different theoretical distributions and real education and psychology data sets are used to evaluate the Type I error and power properties of the t-test, permutation t-test and the Wilcoxon test.

Importance of the Problem

In practicality, the rationale for selecting a particular method for the statistical analysis resides in finding an effect in a treatment, as subtle as it could be if one exists. The detection of the possible effects is quantified by the power of the test. As the power increases, there is a better chance of finding even a slight effect. Although permutation techniques have excellent properties with regard to robustness to Type I error, they are not as powerful as other nonparametric methods under non-normality. Therefore, permutation techniques may not detect small effects in treatments. Instead, the Wilcoxon test would be a more suitable procedure if the nature of the treatment changes the mean of two independent samples.

Historically and incorrectly, nonparametric tests were used for the main purpose of cleansing the data to obtain the appropriate Type I error. Nonparametric methods and permutation techniques do not assume that the data are samples from normal distributions (Nunally, 1975, 1978; Sawilowsky, 1990) and they are capable of preserving Type I error rates to a nominal level

of significance (Sawilowsky & Fahoome, 2003). With the advent of computers, permutation tests were found useful with regard to these two properties under non-normality. The comparison of nonparametric methods and permutation techniques was not pursued to examine which of these two methods has better power.

It is interesting to note that Sir Ronald Fisher, who wrote most of the original power tables with Yates, did not focus his attention to the power issue in permutation tests. The reason is that Sir Ronald Fisher did not support the concept of alternative hypotheses, Type I or Type II errors (Huberty, 1987; Huberty, 1993; Nix & Barnette, 1998), which are necessary to calculate the power for permutation tests. Sir Ronald Fisher developed the significance-testing concept where the determined P-value was the criteria to accept or reject the null hypothesis (Huberty, 1993). Jerzy Neyman and Egon Pearson developed the hypothesis testing with the use of the alternative hypothesis and the specification of α (Huberty, 1987; Huberty, 1993).

Another possible explanation may be that computers, at their times, were too slow to allow simulations of all possible permutations for the development of permutation tests power tables.

Limitations of the Study

Although most occurring situations in education, psychology and related fields reflect a change in variance and shift in means, an important assumption for parametric, nonparametric and permutation tests is the homogeneity of variance. Thus, the scope of this study solely focuses on a shift in means. For

studies where there is a shift in location and scale, Sawilowsky (2002) discussed alternatives to the t-test.

In summary, the use of parametric tests such as the t-test is the best option for the analysis of data under normality. In social and behavioral sciences, the parametric methods are, however, limited, as the assumption of normality is not always met. Nonparametric tests were initially used when the parametric assumptions were not satisfied. They are now used because of their own merits and properties such as their robustness and higher power than their parametric counterparts. Nevertheless, their application is still restricted by underlying requirements such as homogeneity of variance. Permutation tests are alternative techniques for both parametric and nonparametric methods. However, they are also limited to certain assumptions such as random selection in light of generalization of the findings. Under non-normality, they have excellent properties such as being distribution-free, exact and robust which makes their use much more appealing. However, their power is in question compared to the power of nonparametric tests, which brings about the issue of finding possible treatment effects when one exists, no matter how small. Moreover, using the data at hand limits the generalizability of the study.

CHAPTER II

LITERATURE REVIEW

Although combinatorics has been a formal discipline for centuries, Fisher (1935) was one of the earliest to introduce permutation techniques in the format of the null hypothesis through his experiment with the soils. He demonstrated that permutation tests are based on and work well with the actual data instead of the data transformed into ranks, although it was tedious computing the calculations manually. With the evolution of computer technology, computer intensive methods, such as permutation tests, are easy to perform, readily available, and can avoid problematic assumptions that are essential with the conventional parametric methods (Edgington, 1995; Fisher, 1960).

Controversies in the Literature

Adams and Anthony (1996), Lehmann (1986) and Ludbrook and Dudley (1998) asserted that permutation tests have higher power than other nonparametric tests because of the use of actual data instead of ranks. But in a letter to the editor of *The American Statistician* journal, Langbehn, Berger, Higgins, Blair and Mallows (2000) commented that power is not lost during the analysis of ranked data but it may actually be increased. Langbehn, Berger, Higgins, Blair and Mallows (2000) concluded that the Wilcoxon test, among other ranked-based tests is, in many cases, more powerful than the

permutation test and t-test, although the permutation version of a parametric test, such as the t-test, has similar power as the t-test it is based on.

Nonparametric tests are as powerful, if not more powerful than parametric and permutation tests because ranking does not lose information as suspected (Sawilowsky, 1990; Wolfowitz, 1949). Under normality, it had been demonstrated that permutation tests are almost as powerful as the t-test (Good, 1994; Albert, Bickel & van Zwet, 1976; Lehmann & Stein, 1959; Sawilowsky, 1990), which is found to be the most powerful parametric test (Bradley, 1968; Good, 1994). In regard to the robustness, Rao and Sen (2002) stated that permutation methods are more robust than parametric tests, although Boik (1987) and Manly (1995) demonstrated that permutation tests may not necessarily be robust when the samples come from distributions with the same mean but different variance.

Randomization tests and permutation tests are used as interchangeable terms (Heckel et al., 1998; Good, 1994; Lu, Chase & Li, 2001; Kerlinger & Lee, 2000; Rosenberger & Rukhin, 2003). However, Edgington (1995) stated that these two methods are not identical. Edgington (1995) explained that randomization tests are permutation tests with random assignment added to the permutations. Permutation tests do not provide hypothetical outcomes for the same subjects but rather the outcomes of two sets of subjects (treatment and control group) randomly selected from identical populations (Edgington, 1995). Randomization tests generate hypothetical

outcomes for the same set of subjects under alternative random assignment (Edgington, 1995).

The terms “nonparametric” and “distribution-free” are also not interchangeable, although many authors confuse the two. According to Boik (1987), Bradley (1968), Fahoome and Sawilowsky (2000), and Good (1994), nonparametric relates to the fact that the sampled population distribution is not specified, and distribution-free means that the test’s significance level is independent of the form of the distribution of the population.

Permutation Techniques

Assumptions of Permutation Tests

Permutation tests are based on the assumption that the selection of the samples is random, although this fact cannot always be verified (Kempthorne, 1966). According to Edgington (1995) and Noreen (1989), the random selection is not imperative for permutation tests unless the reason for the experiment is inference to the population. Indeed, random selection provides generality (Hunter & May, 1993), which is referring to the implication of the samples’ results to the population.

Although it was said that permutation tests are valid for any type of sample, regardless how it was selected (Lu, et al., 2001), Edgington (1995), Feinstein (1985), Fisher (1935) and Good (1994) do not agree with this statement, especially with regard to random selection. The samples have to be randomly assigned to groups (Kempthorne, 1966). Random assignment is a requirement for cause / effect inference (Hunter & May, 1993). Edgington

(1995) demonstrated that, with random assignment, randomization can provide statistical inferences about the experimental subjects. Randomization methods provide statistical inferences about treatment effects with a minimum of assumptions, which is random assignment (Edgington, 1995).

For permutation techniques on independent tests, exchangeability of the data is required, i.e., for categorical data, the rows and columns scores need to be mutually independent of each other (Commenges, 2003; Good, 1994; Good, 2002; Hunter & May, 1993; Lu, et al., 2001; Opdyke, 2003). In other words, there is no relationship between the scores (Mielke & Berry, 2001; Good, 1994). Explicitly, the probability of any outcome is the same if the observations are exchangeable, regardless of the order the observations are performed (Lu et al., 2001). The observations have to be independent of each other (Good, 1994; Good 2002; Lu, et al., 2001) to be considered exact tests (Berger, Lunneborg, Ernst & Levine, 2002; Lu et al., 2001; Manly, 1995).

Homogeneity of variance is also a requirement (Boik, 1987). It means that the variances of the populations are equal (Hinkle, et al., 2003; Kerlinger & Lee, 2000; Wilcox, 1996), i.e., the variance within and between the groups is statistically the same. Boneau (1960) found that the heterogeneity of variance affects significance tests negatively (Kerlinger & Lee, 2000). Boik (1987) demonstrated that, if the data are heterogeneous in terms of group variance, permutation tests are less robust than parametric tests.

Permutation tests are distribution free and thus do not assume distributional properties of the population (Bradley, 1968; Good, 1994; Mielke

& Berry, 2001). It is an important property for social and behavior sciences because the populations' characteristics from which the samples are drawn are not always known (Edgington, 1995; Good, 2002; Mielke & Berry, 2001). However, the distributions need to be continuous so that the probability of tied ranks is insignificant. A population's distribution is continuous when, for every assumed continuous X, there is a Y on the population distribution (Hinkle, et al., 2003). Moreover, the difference between populations requires being of a uniform shift (Edgington, 1995), not affecting the distribution shape.

Permutation tests are exact statistics (Adams & Anthony, 1996; Berger, 2001; Berger et al., 2002; Bradley, 1968; Manly, 1995; Opdyke, 2003). Walsh (1965) explained that a procedure is exact when the relevant properties are specifically determined. For example, an exact significance test gives a precisely determined level of significance (Walsh, 1965). An exact test provides an actual significance level exactly the same as the nominal significance level (Maritz, 1995). Maritz (1995, p.1) defined that a distribution free method is "valid under minimal assumptions about the underlying distributional form." Feinstein (1985) posited that, with permutation tests, the distributions are determined directly from the data at hand.

Approximate Randomization, Randomization and Permutation Tests

The number of permutations may be overwhelmingly large when the sample is large, making exact randomization not feasible (Edgington, 1995; Noreen, 1989). Approximate randomization is used to decrease the large number of computations required for the permutations. It is a procedure where

a smaller number of permutations is performed by randomly shuffling one variable relative to another avoiding all possible permutations of the variables (Noreen, 1989). It is usually accomplished through a random number generator program. As the number of shuffles approaches the exact randomization procedure, the approximation becomes better as it gets closer to the accurate probability distribution of the exact randomization (Noreen, 1989). The approximate randomization maintains all the properties that the exact randomization has.

Edgington (1995) gave some propositions for the use of randomization: the experimental results from permutations based on random assignment alone can be applicable to nonrandom experiments; a randomization test is valid if the probability of obtaining an exact p-value is as small as the chosen level of significance under the truth of the null hypothesis; it is the negation of the null hypothesis that is provided statistical support by the smallness of the P-value; and permutation tests assist with the understanding of similar tests that require more requirements.

Edgington (1995) explained that the inferences are restricted to the data at hand when random selection is not required. Fisher (1935) posed the random selection problem as a theoretical issue since permutations were not as readily achievable at his time. Fisher (1960) explained that statisticians are aware that only a limited amount of information is available through the finite body of data. Therefore, the amount of information cannot be increased by rearranging the data. Permutations are limited to the data at hand (Feinstein,

1985), because the comparison of the observed values and the permutations are done by reshuffling and relabeling the same data (Good, 1994), which brings up the issue of generalizability. As the current data determine the sampling distribution statistics, randomizations depend on the quality of the scores or data at hand.

t-tests

The parametric Student's t-test is robust with respect to Type I error for departure from normality when the samples are equal, greater than about 25 or 30, and the variances are equal (Sawilowsky & Blair, 1992; Wilcox, 1996). Generally when the samples are unequal and the tested two groups are from different distributions, the t-test performs unsatisfactory in controlling the Type I error, even under normality (Wilcox, 1996). Permutations methods are found to control for the Type I error in the Student's t-test (Edgington, 1995; Mielke & Berry, 2001), i.e., permutations on the t-test decrease the Type I error rate to a nominal alpha.

Although the t-test has optimal power properties under normality (Blair & Higgins, 1980), it can be inadequate when departure of normality or when heavy tailed distributions occur (Wilcox, 1996). Permutation tests and nonparametric tests are more powerful than the t-test when the data are skewed (Edgington, 1995; Lu, et al. 2001).

Rasmussen (1985) conducted a study to show that if the data were to be cleaned from outliers prior to the t-test, the t-test and the Wilcoxon test would actually show the same power. In his study, Rasmussen (1985) showed

that the t-test corrected for outliers has power advantages over the Wilcoxon test. However, it is important to note that Rasmussen (1985) did not perform any data cleansing before running the Wilcoxon test. It is known that the Wilcoxon method ensures the appropriate handling of the outliers and other abnormal data in regards to Type I error but for maximum power, it requires some data adjustment. Thus in Rasmussen's (1985) study, finding that these two statistics have the same power is not reasonable because of the unequal treatment of the competitor.

In summary, the debate on nonparametric tests being less powerful than their parametric counterparts seems unsettled and contentious. Sawilowsky (1990) and colleagues continue to argue about the veracity of these statements.

Certain terms, such as randomization and permutation, and distribution-free and nonparametric, have been used interchangeably, which is incorrect. It appears that terms are employed with different understanding of their definitions. Although permutation tests require both random assignment and selection, the former requirement is for cause/effect inferences of the results and random selection is necessary for generalizability.

Edgington (1995), Feinstein (1985), Good (1994), and Mielke and Berry (2001) amongst others agreed that permutation tests depend on the data at hand. Permutation tests also have assumptions, as parametric and nonparametric tests do. It has been stated that permutation tests maintains the Type I error to a nominal significance level, its power is still fairly lower

compared to other nonparametric methods such as the Wilcoxon test's power under non-normality.

CHAPTER III

METHODOLOGY

The goal of this study is to investigate how permutation tests behave with regard to the Type I error and power under normal and non-normal situations. The parametric test, the independent Student's t-test, and the nonparametric test, the Wilcoxon test, will be used to compare them with their respective permutation tests. To accomplish this purpose, a Monte Carlo simulation is conducted using the independent Student's t-test, according to the formula**:**

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right)}{\sqrt{\left[\frac{\sum X_1^2 - \frac{\left(\sum X_1\right)^2}{N_1} + \sum X_2^2 - \frac{\left(\sum X_2\right)^2}{N_2}}{N_1 + N_2 - 2}\right]\left[\frac{1}{N_1} + \frac{1}{N_2}\right]}}$$

and the Wilcoxon rank-sum statistic according to the formula:

$$U = W - \frac{n(n+1)}{2}$$

where W is the sum of the ranks for sample of size n.

The Monte Carlo study consists of multiple parts. The first section deals with the conformity of the Type I error to α with the permutation method. At first, the permutation method is applied with a t-test under normality comparing two small equal random samples containing 10 numbers in each group, n_1 = n_2 = 10. These random numbers are obtained from Rangen 2.0, which is a collection of subroutines used to generate pseudo-random numbers

(Sawilowsky & Fahoome, 2003). The Rangen 2.0 shows a normal distribution as follows:

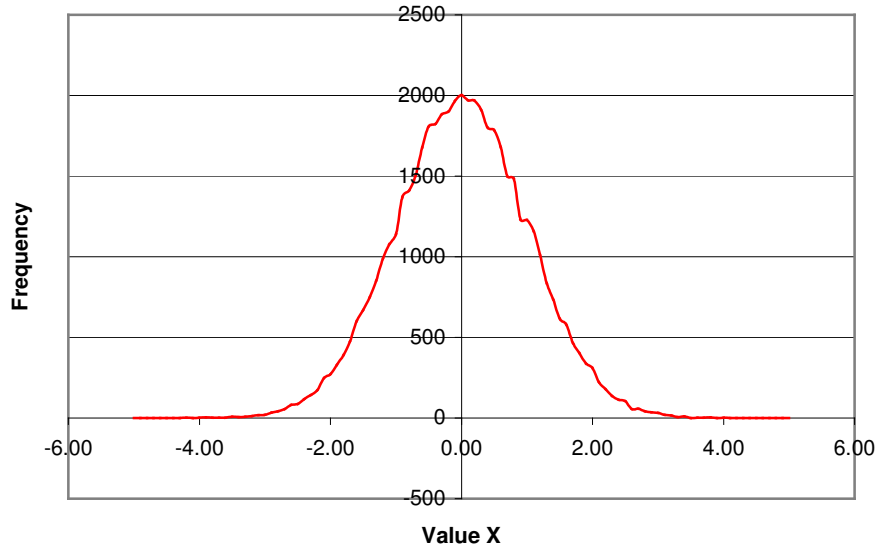

Figure 1: Normal Distribution from Pseudo-Random Data.

The procedure is as follows: the t-test value is calculated and compared to the critical t value for all possible permutations between the two samples. In this case, the one-tail t-test with a degree of freedom of 18 and $\alpha = .05$ gives a critical t-value of 1.734. The proportion of the permutations that have t-test values greater than or equal to 1.734 divided by the amount of all permutations is then the p-value. One million simulations of sample sizes $n_1 = n_2 = 10$ are generated in order to determine the percentage of the pairs whose original t-value (without applying permutation) is bigger than or equal to the critical t-value. Out of the million simulations 1,500 times were used for the permutation t-test with different random samples of two sets of 10 numbers in order to verify that the Type I error occurs at α or .05.

The same procedure is accomplished with two other distributions. One of the distribution is the exponential with μ = σ = 1, that is reflected in Figure 2, what the obtained pseudo-random numbers gave.

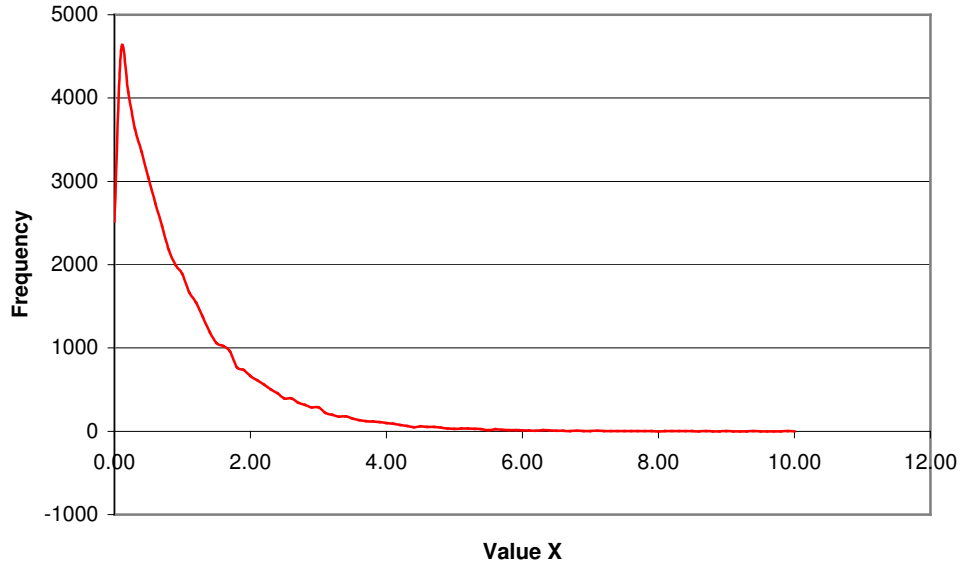

Figure 2: Exponential Distribution (μ = σ = 1) from Pseudo-Random Data.

The other distribution used is the Chi-square with degree of freedom of 6. This Chi-square with degree of freedom of 6 was selected arbitrarily as this distribution appeared to be neither too skewed nor too symmetric, as reflected in Figure 3.

A real data set, the Multimodal-Lumpy data set (Sawilowsky, Blair & Micceri, 1990) is also used for the procedure. This real data set represents a distribution commonly found in the area of education and psychology (Micceri, 1989). Figure 4 reflects the appearance of the data set distribution.

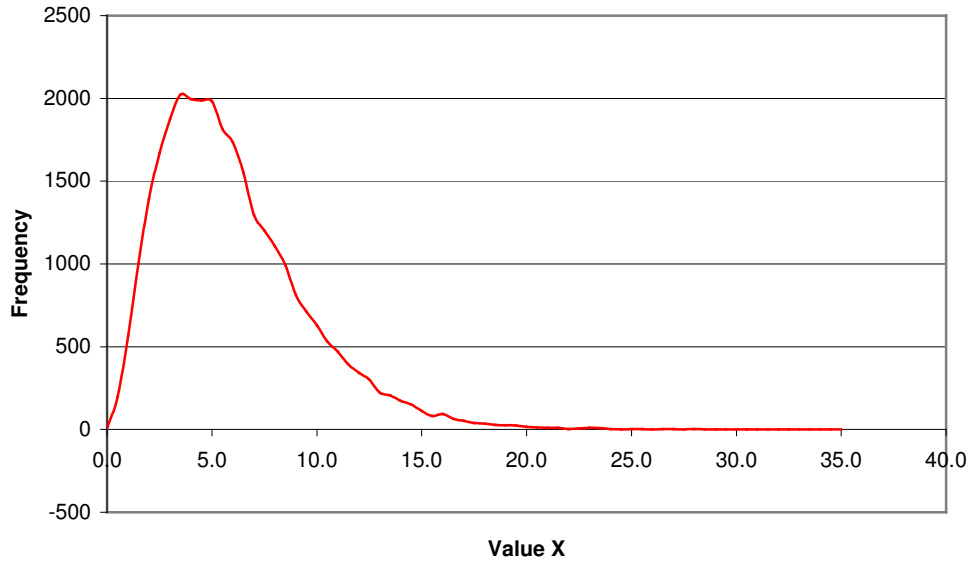

Figure 3: Chi-square Distribution (df = 6) from Pseudo-Random Data.

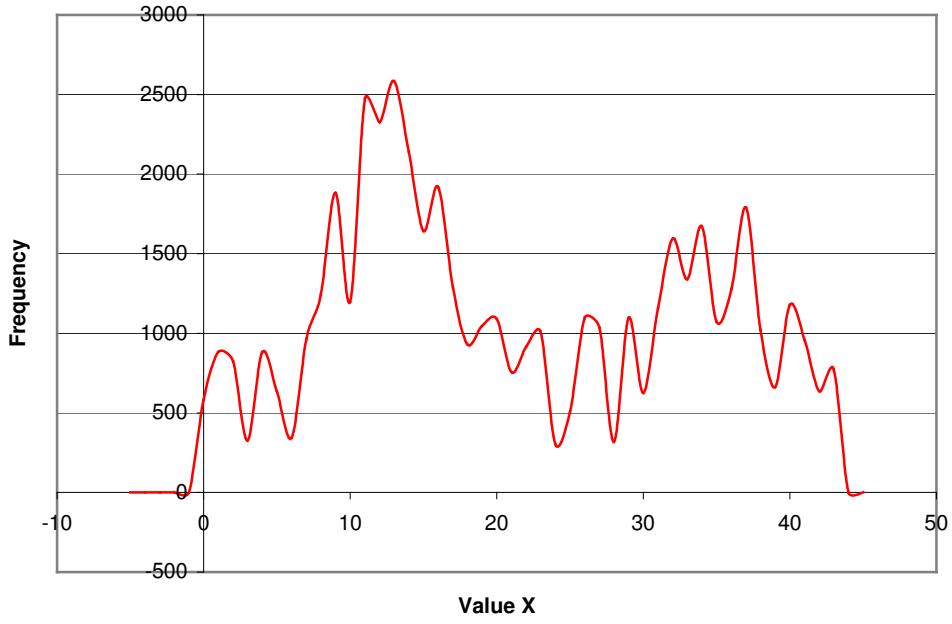

Figure 4: Multimodal-Lumpy Data Set Distribution from Pseudo-Random Data.

These distributions and the real data set were chosen to examine the Type I error of the t statistic under normality and non-normality situations.

The next simulation involves the same steps described earlier but with an unequal sample size $n_1 = 5$ and $n_2 = 15$. The third part of the simulations focuses on larger samples. The same procedures are repeated, for sample sizes of $n_1 = n_2 = 20$, and $n_1 = 10$ and $n_2 = 30$ of randomly selected numbers.

Finally for each of the previously described 1,500 samples per distribution, an initial t value is calculated and now compared, not with the critical t-value, but with the 95^{th} percentile t-value obtained by permutation of the respective samples. The percentage of t-values larger than or equal to the respective 95^{th} percentile t-value is the p-value for nominal $\alpha = .05$. Then, the Wilcoxon test's Type I error rate is also calculated with the permutation technique for all four samples sizes, three distributions and the real data set. The Wilcoxon test is performed by ranking the random numbers from the highest to the lowest and calculating the Wilcoxon rank-sum statistic using the sum of the ranks of the numbers in one sample and comparing the U-value with the tabled critical value for the U-statistic. As the Wilcoxon test is done by ranking the data, the Type I error of the Wilcoxon test does not depend on the distribution type, but only on the sample size and the used critical value, that is in the absence of ties.

In the second section of the Monte Carlo study, the power of the t-test, permutation test and Wilcoxon test is calculated and compared for the same four distributions. The samples are again randomly generated from two identical distributions but this time with different distribution means. The means are shifted by .2σ, .5σ, .8σ, 1.2σ of the respective distribution to

generate different power levels. The first simulation part is performed with the sample size of $n_1 = n_2 = 10$, and the size sample of $n_1 = 5$ and $n_2 = 15$. The second set of simulations is done with larger sample sizes of $n_1 = n_2 = 20$, and $n_1 = 10$ and $n_2 = 30$. The power of the permutation test is obtained by recording the percentage of t-values of the original samples before permutation that is higher than or equal to the 95^{th} percentile t-value obtained by permutation. This procedure is repeated 1,500 times for the all sample sizes, all distributions and data set. The respective t-test power is calculated as the percentage of samples with an original t-value larger than or equal to the critical t-value. It is expected that the power of the permutations and the t-test will be approximately the same, with the permutation's power being slightly higher.

As a comparison, the Wilcoxon test will be computed for each iteration discussed above. The power is then calculated as the U-values smaller than or equal to the critical value divided by all U-values. In the same way as the t-test and permutation t-test, graphs for each sample size are plotted showing power versus distribution shift for the Wilcoxon test.

In summary, through Monte Carlo simulations, p-values are obtained with critical t-test value for $\alpha = .05$ and the 95^{th} percentile values generated by permutation for a normal distribution and two non-normal distributions, i.e. exponential ($\mu = \sigma = 1$), Chi-square ($df = 6$) and the Multimodal-Lumpy data set, which is a real data set. The second section is to compare the power of the permutation t-test, the t-test and the Wilcoxon test for samples from shifted

distributions for normal and non-normal situations, equal and unequal sample sizes.

CHAPTER IV

RESULTS

The Monte Carlo simulations were performed to investigate two main concepts. The first focused on the comparison of the t-test and permutation t-test with regard to the Type I error rate in normal, non-normal distributions and a real data set with small equal and unequal samples, and larger equal and unequal samples. The second concept compared the power of the t-test and permutation t-test compared to the Wilcoxon test both under normality and non-normality for the different types of samples.

Type I Error Rate

Equal Sample Sizes

The first comparison focused on the normal distribution and sample sizes $n_1 = n_2 = 10$. After generating 1,000,000 random sets of two sample size of $n_1 = n_2 = 10$, the Type I error rate was approximately 5%, in that 50,010 of the calculated t-values were above or equal to the critical value of 1.734. The permutation t-test was performed on 1,500 of these random sets to show the rehabilitation of the Type I error rate to .05. The Wilcoxon test yielded a Type I error rate at .045.

The next step was to proceed with the same sample sizes ($n_1 = n_2 = 10$) drawn from the exponential distribution ($\mu = \sigma = 1$). The 1,000,000 repetitions yielded a Type I error rate of 4.81%. The permutation t-test, again restored the Type I error rate to .05 and the Wilcoxon test gave a Type I error

rate of .045. With respect to the Chi-square distribution (df = 6), there was a Type I error rate of 4.94% for the million repetitions. The permutation t-test rehabilitated its Type I error rate to .05. The Wilcoxon test was also .045. The Multimodal Lumpy data set gave a Type I error rate of 4.97% after the million repetitions. The permutation t-test again rehabilitated the Type I error rate and the Wilcoxon test yielded a Type I error at .045.

The following simulations dealt with sample sizes $n_1 = n_2 = 20$ random numbers. After 1,000,000 repetitions, the t-test in the normal distribution gave a Type I error rate of 5.00%. The Type I error rate on the exponential distribution ($\mu = \sigma = 1$) resulted in 4.94%. The Chi-square distribution (df = 6) generated a rate of 4.98% with regard to the Type I error on the original 1,000,000 t-tests. In the Multimodal Lumpy data set, the Type I error rate was 5.43% for the t-tests. After 1,500 repetitions, the permutation t-tests rehabilitated the Type I error rate to .05 for all distributions and data set. The Wilcoxon test's Type I error rate was .048.

Unequal Sample Sizes

The focus, here, is on the sample sizes $n_1 = 5$ and $n_2 = 15$. For the normal distribution, the simulation for the original t-test was performed 1,000,000 times and gave a Type I error rate of 4.99%. The exponential distribution ($\mu = \sigma = 1$) generated a Type I error rate of 2.76% for the million repetitions of the original t-tests. The Chi-square distribution (df = 6) simulations produced a rate of 3.84% in regards to the Type I error after the million repetitions. In the Multimodal Lumpy data set, the Type I error rate was

4.84%. For all distributions and data set, the permutation t-tests again restored the Type I error rate at .05 and the Wilcoxon tests gave a Type I error rate of .049.

The following simulations dealt with sample sizes $n_1 = 10$ and $n_2 = 30$. After 1,000,000 repetitions, the t-test in the normal distribution gave a Type I error rate of 4.99%. The Type I error rate on the exponential distribution ($\mu = \sigma = 1$) resulted in 3.56%. The Chi-square distribution (df = 6) reflected a rate of 4.25% with regard to the Type I error on the original 1,000,000 t-tests. The Multimodal Lumpy data set gave a Type I error rate of 4.90% on the regular t-tests. As previously, the permutation t-test restores the Type I error rate to the significance level for the three distributions and the real data set. The Wilcoxon test yielded to the same Type I error rate of .048 for all the distributions and data set.

Table 1 summarizes all the Type I error rate findings for the three distributions and data set along with the different sample sizes.

	Type I Error Rate		
	t-test	Permutation t-test	Wilcoxon test
Normal Distribution			
$n_1 = n_2 = 10$	0.050	0.050	0.045
$n_1 = 5, n_2 = 15$	0.050	0.050	0.049
$n_1 = n_2 = 20$	0.050	0.050	0.048
$n_1 = 10, n_2 = 30$	0.050	0.050	0.048
Exponential Distribution			
$n_1 = n_2 = 10$	0.048	0.050	0.045
$n_1 = 5, n_2 = 15$	0.028	0.050	0.049
$n_1 = n_2 = 20$	0.049	0.050	0.048
$n_1 = 10, n_2 = 30$	0.036	0.050	0.048
Chi-square Distribution			
$n_1 = n_2 = 10$	0.049	0.050	0.045
$n_1 = 5, n_2 = 15$	0.038	0.050	0.049
$n_1 = n_2 = 20$	0.050	0.050	0.048
$n_1 = 10, n_2 = 30$	0.043	0.050	0.048
Multimodal Lumpy Data set			
$n_1 = n_2 = 10$	0.05	0.050	0.045
$n_1 = 5, n_2 = 15$	0.048	0.050	0.049
$n_1 = n_2 = 20$	0.054	0.050	0.048
$n_1 = 10, n_2 = 30$	0.049	0.050	0.048

Table 1: Type I error Rate for all Distributions & Data Set with all Sample Sizes.

The next step was to use permutations to calculate the 95^{tile} t-values for the 1500 samples of each distribution and they were compared to the t-test critical t-value. The unequal sample size $n_1 = 5$ and $n_2 = 15$ was used because its Type I error rate was extensively different from the .05 level of significance.05. In the normal distribution, 23% of the permutation t-values were below the critical t-value of 1.734. These twenty three percent were divided into four parts. The lower quartile, which contains the first 25% of the permutation t-values below 1.734, was 1.718, .9% smaller than 1.734. The upper quartile, which contains 75% of the permutation t-values below 1.734, was 3.5% smaller than 1.734. Forty six percent of the permutation t-values were above the critical value. For this case, the lower and upper quartiles were 1% and 3.7% larger than the critical t-value, respectively.

In the exponential distribution ($\mu = \sigma = 1$), the lower quartile below the critical value was 5.4% and the upper quartile was 16.2%. For the permutation t-values above 1.734, the lower quartile was .2% and the upper quartile was 2.1%.

Seventy eight percent of the permuted t-values were below 1.734 in the Chi-square distribution (df = 6). The lower and upper quartiles were 2.7% and 9.4% respectively. The lower and upper quartiles of the values above the critical value were .5% and 1.6 % respectively.

In the Multimodal Lumpy data set, 58% of the permutation t-values were below the critical value. The lower quartile was .9% and the upper

quartile was 3.0%. Thirty four percent were above 1.734. The lower and upper quartiles were respectively .5% and 2.1%.

Table 2 summarizes the findings of the different quartiles in the different distributions and data set.

	Normal Dist.	Exponential Dist.	Chi-square Dist.	Multimodal Lumpy
Below 1.734				
Lower Quartile	0.9%	5.4%	2.7%	0.9%
Upper Quartile	3.5%	16.2%	9.4%	3.0%
Above 1.734				
Lower Quartile	1.0%	0.2%	0.5%	0.5%
Upper Quartile	3.7%	2.1%	1.6%	2.1%

Table 2: Quartiles Findings for all Distributions & Data Set for $n_1 = 5$ & $n_2 = 15$

Analysis of Power of t-test, Permutation t-test and Wilcoxon test

Small Equal Sample sizes

The first section focuses on the comparison of power between the t-test, permutation t-test and the Wilcoxon test for sample sizes $n_1 = n_2 = 10$. The random numbers were selected from the same distribution with shift = .2σ, the resulting power was .108 for the t-test, .109 for the permutation t-test, and .096 for the Wilcoxon test. In the .5σ shift, the powers increased to .276 for both t and permutation t-tests and .245 for the Wilcoxon test. The next shift was at .8σ, where the powers were .526 for the t-test, .0525 for the permutation t-test and .484 for the Wilcoxon test. The last shift was 1.2σ. The normal distribution resulted in powers of .817 for both the t and permutation t-tests and .773 for the Wilcoxon test.

The following figure gives a clearer picture of the power in the normal distribution with the different magnitudes of shift. For the sample size of $n_1 = n_2 = 10$, the powers of the t-test and the permutation t-test are equal but the Wilcoxon test is less powerful than the two other tests.

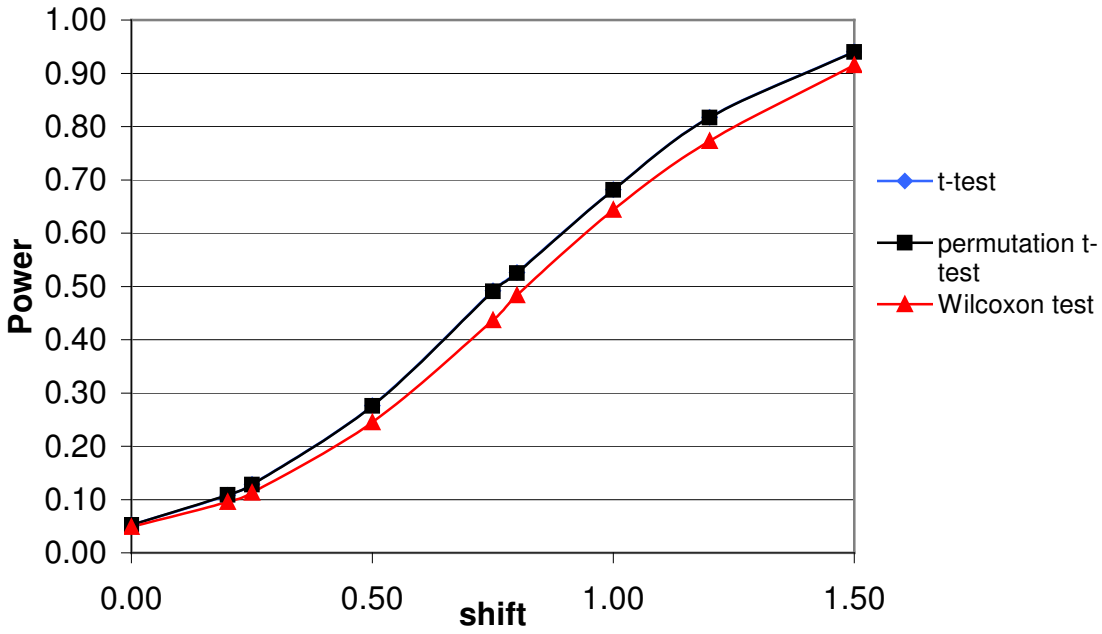

Figure 5: Shift vs. Power in the Normal Distribution for $n_1 = n_2 = 10$

The results for the exponential distribution ($\mu = \sigma = 1$) with shift = .2σ gave powers of .130, .132, and .156 for the t, permutation t and the Wilcoxon tests, respectively. The next set of simulations was at shift .5σ. This distribution gave powers of .328 for the t, .334 for the permutation t-test, and .424 for the Wilcoxon test. At 1.2σ shift, the powers were .829 for the t-test, .831 for the permutation t-test and .893 for the Wilcoxon test.

Figure 6 shows that the Wilcoxon test is more powerful than both the t and permutation t-tests, which have essentially the same power. The ratio

between the t-test and the Wilcoxon test increased from .77 to .83 at .5σ and .8σ shifts respectively, then decreased as the shifts increased.

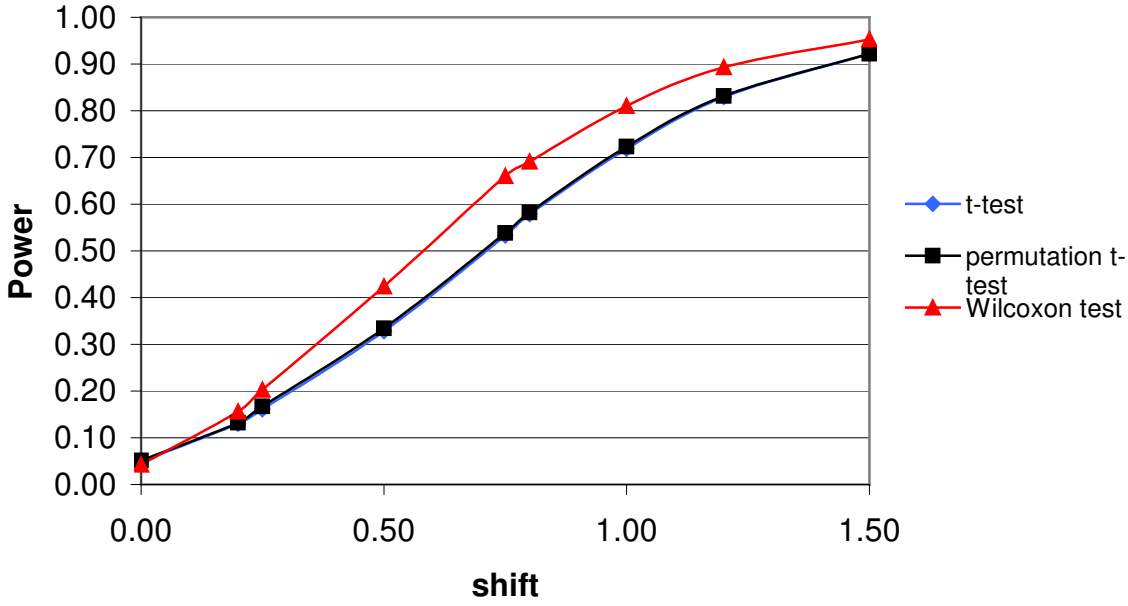

Figure 6: Shift vs. Power in the Exponential Distribution for Sample Size $n_1 = n_2 = 10$.

The Chi-square distribution (df = 6) with shift = .2σ yielded powers of .110 for the t and the permutation t, and .113 for the Wilcoxon tests. The shift was changed to .5σ for the next set of simulations. The Chi-square distribution produced .301 for the t and the permutation t-tests, and a slightly higher power of .303 for the Wilcoxon test. At .8σ shift, the powers were .552 for the t-test, .553 for the permutation t-test, and .579 for the Wilcoxon test. At the last shift considered 1.2σ, the powers were .824, .825, and .833 for the t-test permutation t-test and Wilcoxon test, respectively.

In the Chi-square distribution (df = 6), the following figure shows a slight difference between the tests, with the Wilcoxon test being somewhat more powerful.

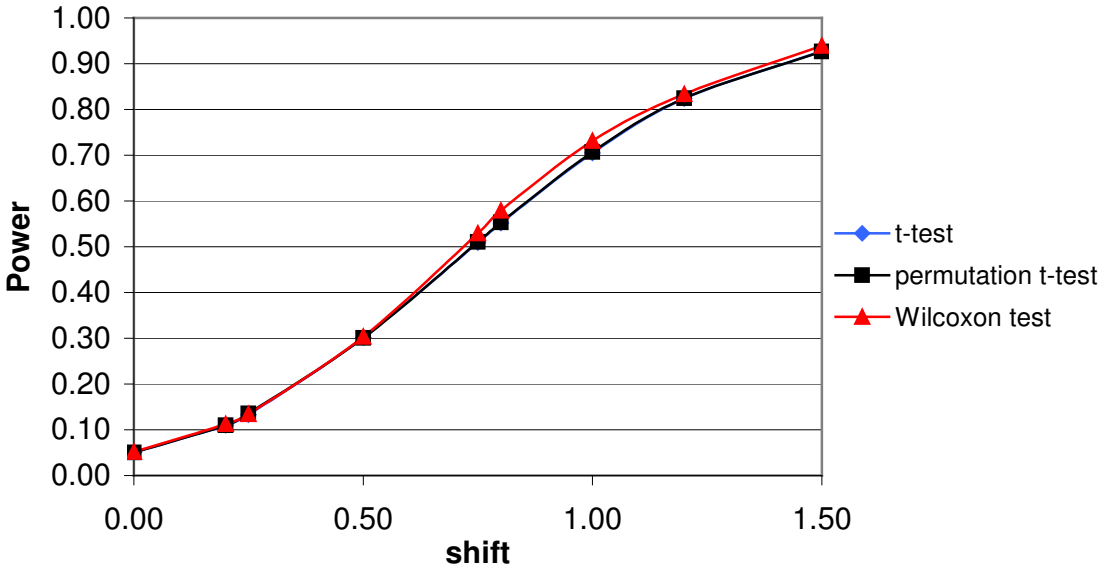

Figure 7: Shift vs. Power in the Chi-Square Distribution for Sample Size $n_1 = n_2 = 10$.

For the Multimodal Lumpy data set with a shift = .2σ shift, the obtained powers were .096 for the t-test and the Wilcoxon test, and .0.95 for the permutation t-test. The next shift was .5σ and gave powers of .259 for the t-test, .257 for the permutation, and .241 for the Wilcoxon test. The next shift was at .8σ. The powers were .505, .506, and .478 for the three tests respectively. The last shift was 1.2σ. The powers were .824 for the t and permutation t-tests and .738 for the Wilcoxon test.

Figure 8 shows that the Wilcoxon test is fairly less powerful than the t and permutation t-tests in the Multimodal Lumpy data set.

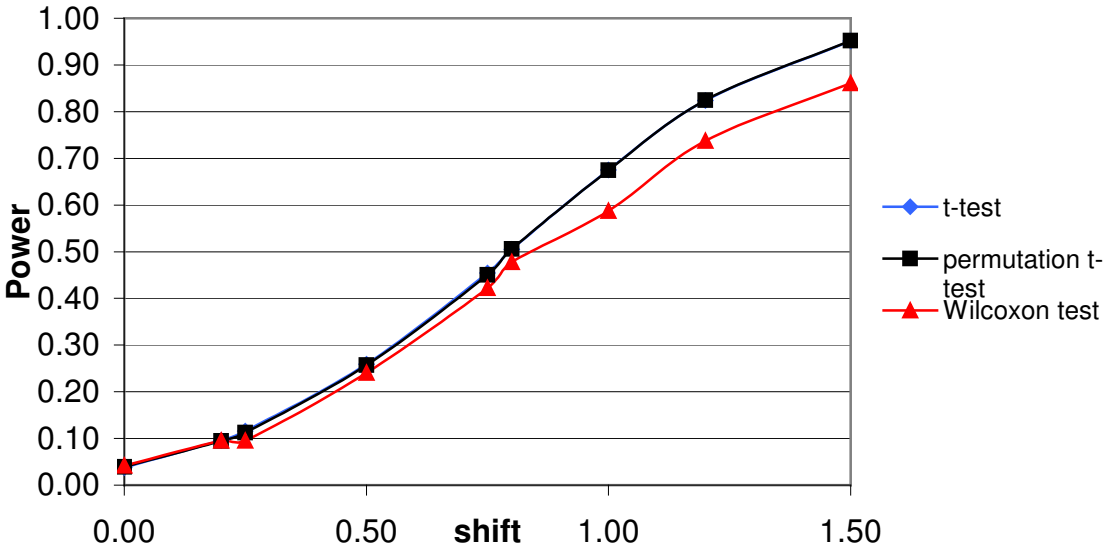

Figure 8: Shift vs. Power in the Multimodal Lumpy Data Set for Sample Size $n_1 = n_2 = 10$.

Table 3, below, summarizes the results of the four distributions with the four shifts in the sample sizes $n_1 = n_2 = 10$.

	Normal Distribution			Exponential Distribution			Chi-Square Distribution			Multimodal Lumpy Dist.		
Shift	t-test	Permut. t-test	Wil-coxon	t-test	Permut. t-test	Wil-coxon	t-test	Permut. t-test	Wil-coxon	t-test	Permut. t-test	Wil-coxon
0	0.052	0.052	0.048	0.049	0.051	0.048	0.051	0.051	0.052	0.038	0.039	0.042
.2σ	0.108	0.109	0.096	0.130	0.132	0.156	0.110	0.110	0.113	0.096	0.095	0.096
.5σ	0.276	0.276	0.245	0.328	0.334	0.424	0.301	0.301	0.303	0.259	0.257	0.241
.8σ	0.526	0.525	0.484	0.579	0.582	0.691	0.552	0.553	0.579	0.505	0.506	0.479
1.2σ	0.817	0.817	0.773	0.829	0.832	0.893	0.824	0.825	0.833	0.824	0.824	0.738

Table 3: Obtained Power for Different Shifts for Sample Size $n_1 = n_2 = 10$.

Small Unequal Sample sizes

This section deals with unequal sample sizes $n_1 = 5$ and $n_2 = 15$. As previously, the four shifts of the identical distributions are 0, .2σ, .5σ, .8σ and 1.2σ.

Under normality and at .2σ shift, the t-test and permutation t-test performed somewhat better, .102 than the Wilcoxon test with a power of .097. The next shift was .5σ in the distributions. The powers were the same for the t and permutation t-test, .222, and the power of Wilcoxon test was slightly smaller at .215. At .8σ shift, the powers obtained were as follows: .420 for t-test, .421 for the permutation t-test and .405 for the Wilcoxon test. The normal distribution with 1.2σ shift reflected powers of .718 for the t-test, .720 for the permutation t-test and .682 for the Wilcoxon test.

Figure 9 below shows that the t and permutation t-tests have the same power. The Wilcoxon test becomes less powerful than the two others tests after a shift of .5σ.

A shift of .2σ, for the exponential distribution ($\mu = \sigma = 1$), showed increasing powers of .090 for the t-test, .144 for the permutation t-test and .158 for the Wilcoxon test. The next shift was .5σ and showed an increase in power with .278 for the t-test, .353 for the permutation t-test and .376 for the Wilcoxon test. At .8σ shift in the distributions, the powers increased from .512 for the t-test, .574 for the permutation t-test to .597 for the Wilcoxon test. At the last shift of 1.2σ, the power of the permutation t-test was higher, .787 compared to .755 for the t-test and .773 for the Wilcoxon test.

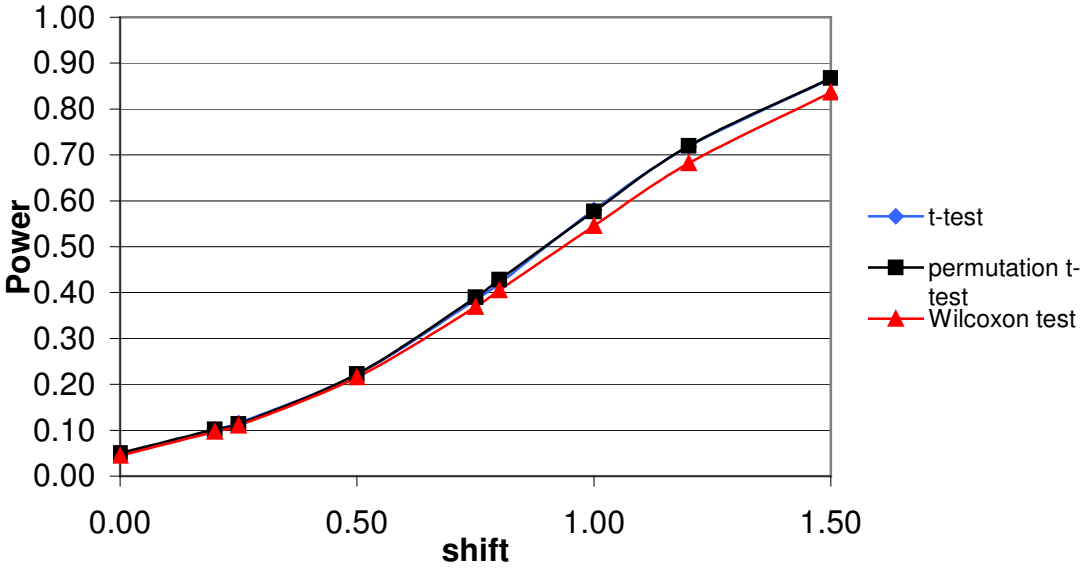

Figure 9: Shift vs. Power in the Normal Distribution for Sample Size $n_1 = 5$ & $n_2 = 15$.

The following figure shows that the Wilcoxon test is slightly more powerful than the permutation t-test at small shifts. Both tests are more powerful than the t-test. As they reach approximately 1.2σ, they have similar powers.

With shift = .2σ in the Chi-square distribution (df = 6), the powers were .090 for the t-test, .110 for the permutation t-test and .121 for the Wilcoxon test. The next shift was .5σ in the distributions. The power of the t-test was smaller at .248 than both the permutation t and Wilcoxon tests, which was .276 for both. At .8σ shift, the power of the permutation t-test was higher, .505 than both the t-test and Wilcoxon test with .472 and .491 respectively. For the last shift of 1.2σ, the power of the permutation t-test was somewhat higher with .761 than .739 for the t-test and .731 for the Wilcoxon test.

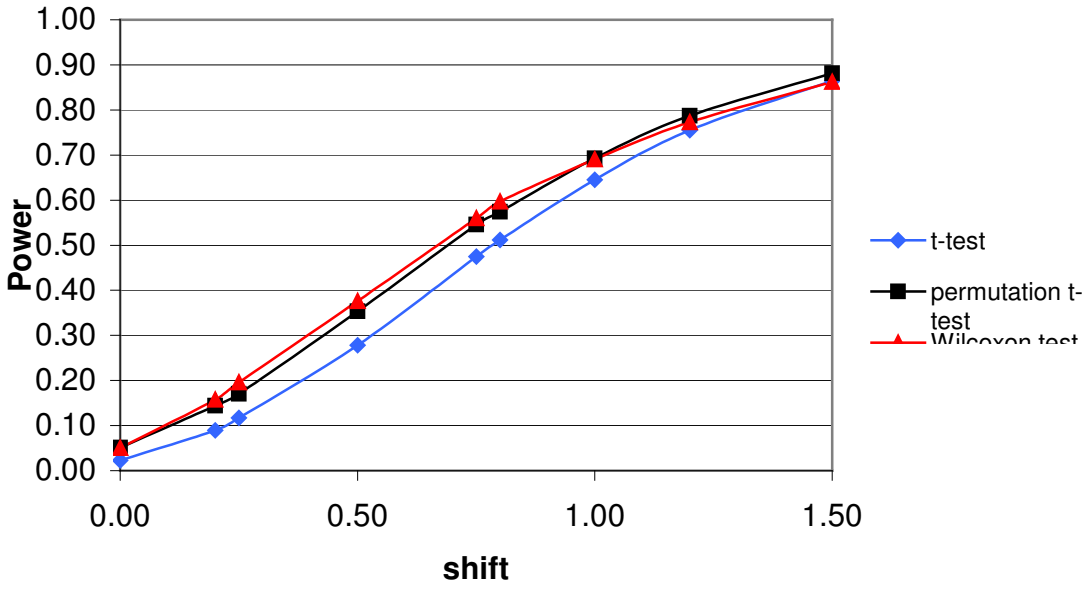

Figure 10: Shift vs. Power in the Exponential Distribution for Sample Size $n_1 = 5$ & $n_2 = 15$.

The figure below reflects the same power increase for the permutation t-test and the Wilcoxon test over the t-test until the .75σ shift. Then, the power of the t-test increases over the Wilcoxon test while the permutation t-test remains more powerful than both.

At .2σ shift in the Multimodal Lumpy data set, the power was .093 for the t-test, to .098 for the permutation t-test and .106 for the Wilcoxon test. The next shift was .5σ and gave decreasing powers with .236 for the t-test, .234 for the permutation t-test and .225 for the Wilcoxon test. At .8σ shift in the distributions, the power of the permutation t-test was higher than the two other tests with .436, .427, and .399 respectively. At 1.2σ shift, this distribution showed a decrease in power from .870 for both the t and the permutation t-test to .764.

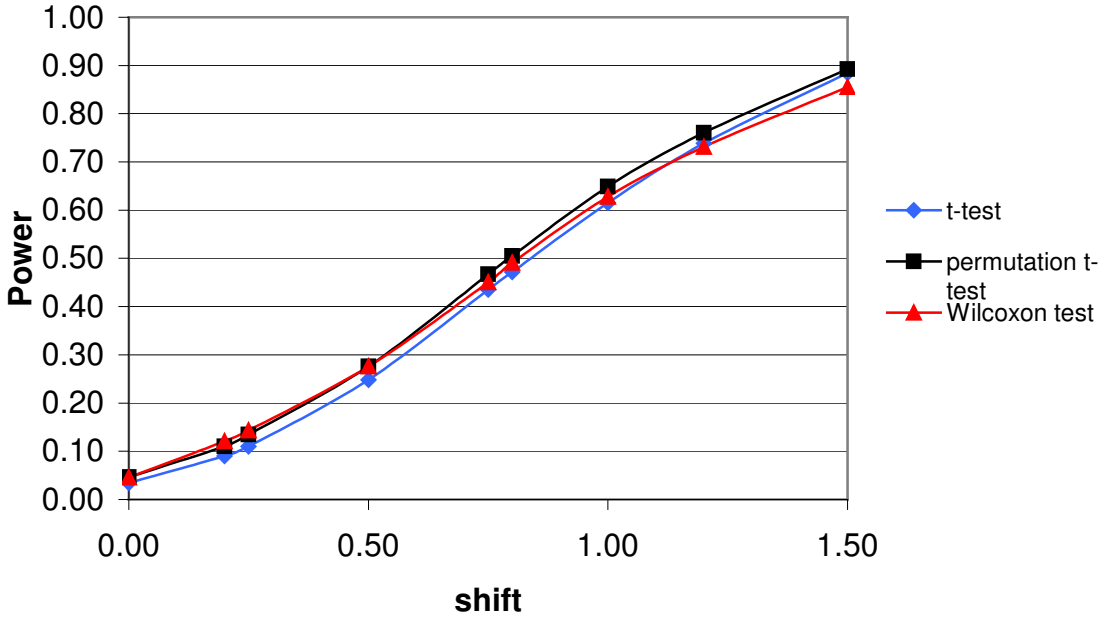

Figure 11: Shift vs. Power in the Chi-Square Distribution for Sample Sizes $n_1 = 5$ & $n_2 = 15$.

In the Multimodal Lumpy data set figure below, the Wilcoxon test starts off to be more powerful than the two other tests, which have the same power. By .2σ shift, both t and permutation t-tests become more powerful than the Wilcoxon test.

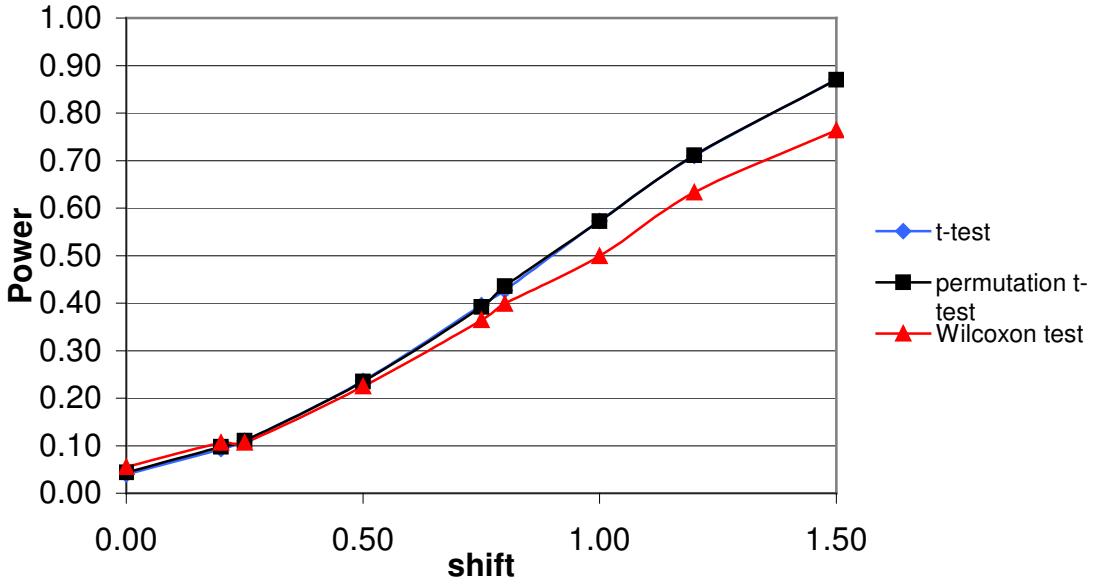

Figure 12: Shift vs. Power in the Multimodal Lumpy Data Set for Sample Sizes $n_1 = 5$ & $n_2 = 15$.

Table 4, below, summarizes the results of the four distributions with the four shifts in the sample sizes $n_1 = 5$ and $n_2 = 15$.

	Normal Distribution			Exponential Distribution			Chi-Square Distribution			Multimodal Lumpy Dist.		
Shift	t-test	Permut. t-test	Wil-coxon	t-test	Permut. t-test	Wil-coxon	t-test	Permut. t-test	Wil-coxon	t-test	Permut. t-test	Wil-coxon
0	0.046	0.050	0.044	0.023	0.051	0.050	0.035	0.047	0.046	0.040	0.044	0.055
.2σ	0.102	0.102	0.097	0.090	0.144	0.158	0.090	0.110	0.158	0.093	0.097	0.106
.5σ	0.222	0.222	0.215	0.278	0.353	0.376	0.248	0.276	0.276	0.236	0.235	0.225
.8σ	0.42	0.428	0.406	0.512	0.574	0.597	0.472	0.505	0.491	0.427	0.436	0.399
1.2σ	0.718	0.722	0.682	0.755	0.787	0.773	0.739	0.761	0.731	0.710	0.711	0.633

Table 4: Obtained Power for Different Shifts for Sample Sizes $n_1 = 5$ & $n_2 = 15$.

Large Equal Sample sizes

The same analysis was applied for this section, i.e. the power was calculated for the four distributions with the shifts of 0, .2σ, .5σ, .8σ and 1.2σ, but with sample sizes $n_1 = n_2 = 20$.

In the normal distribution with a shift of .2σ, the powers were .149 for both the t and the permutation t-test, and .156 for the Wilcoxon test. With a .5σ shift, the power of the t and the permutation t-test was .454, higher than the Wilcoxon test's power of .433. At .8σ shift in the location parameter, the powers were .801 for the t-test, .808 for the permutation t-test and .785 for the Wilcoxon test. At 1.2σ shift, the power of the permutation t-test was slightly higher, .974, than the t-test's power of .969 and the Wilcoxon test's with .973.

In the pattern for the normal distribution in the figure below, the overall power of the t and the permutation t-test is the same, and is slightly higher than the power of the Wilcoxon test. However, from 0 to .2σ shift and from 1.2σ to 1.5σ shift, the Wilcoxon test is as powerful as the two other tests.

The exponential distribution ($\mu = \sigma = 1$) with a shift of .2σ, the powers of the three tests were .160, .162 and .255 respectively. At a shift of .5σ, the results for the exponential distribution showed a slight increase of the power of the t-test, which was .496, compared to that of the permutation t-test of .497. However, there was a considerable increase in the power of the Wilcoxon test to .712. With the .8σ shift in the distributions, the powers increased from .798, .806 to .937 for the respective tests. The last shift was 1.2σ. The powers of

these three tests were increasing from .961, .965 to .995 respectively in the exponential distribution.

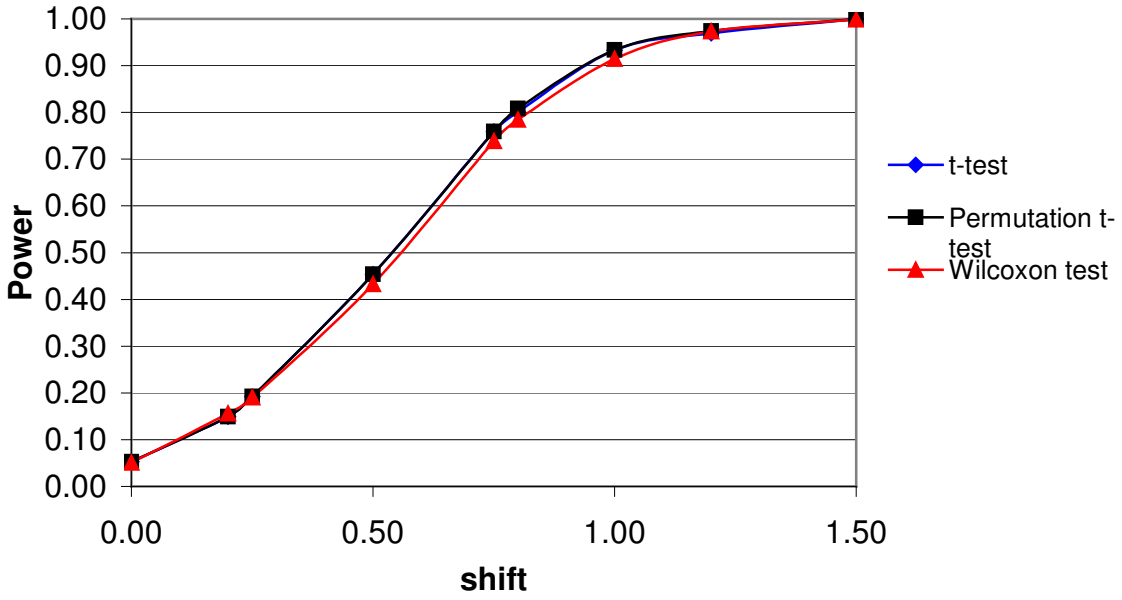

Figure 13: Shift vs. Power in the Normal Distribution for Sample Sizes $n_1 = n_2 = 20$.

In the exponential distribution ($\mu = \sigma = 1$) figure, while the t and the permutation t-tests are as powerful as one another, the power of the Wilcoxon test is much higher. Indeed, the ration between the Wilcoxon test and the t-test increased rapidly to a maximum of .85 at shift = .8σ then decreased until all three tests' power level off at 1.00.

In the Chi-square distribution (df = 6) with shift = .2σ, the powers were slightly with .151 for the t-test, .152 for the permutation t-test, and .153 for the Wilcoxon test. At .5σ shift, the power of the t and the permutation t-tests was the same .469, but higher for the Wilcoxon test at .523. At a .8σ shift in the distributions, the powers obtained slightly increased for the three tests and

were .804, .812, and .858 respectively. At 1.2σ shift, the increasing powers were .967 for the t-test, .972 for the permutation t-test, and .987 for the Wilcoxon test.

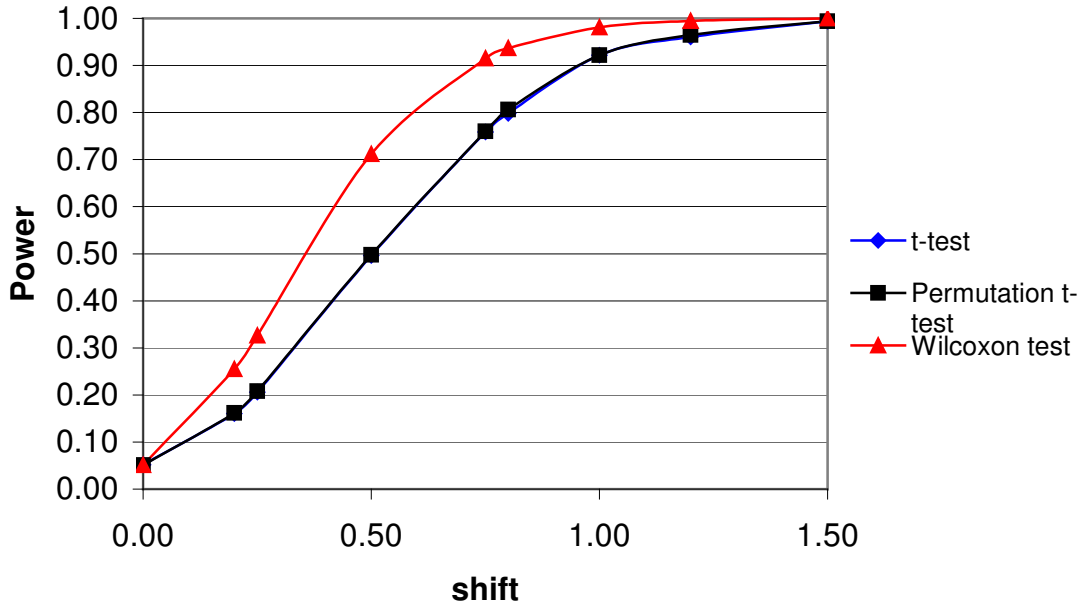

Figure 14: Shift vs. Power in the Exponential Distribution for Sample Sizes $n_1 = n_2 = 20$.

In the following figure, all three tests start with the same power until the .25σ shift. After that shift, the Wilcoxon test becomes somewhat more powerful than the t and the permutation t-test, which have the same power.

In the Multimodal Lumpy data set with a shift of .2σ, the powers were .151 for the t-test, .150 for the permutation t-test, and .180 for the Wilcoxon test. At a shift of .5σ, the power was .461 for both the t and permutation t-tests, but somewhat lower, .439 for the Wilcoxon test. The next simulation run dealt with a .8σ shift in the distributions. The Multimodal Lumpy data set gave

powers of .811 for the t-test, .819 for the permutation t-test and .760 for the Wilcoxon test. The last shift was 1.2σ. The power of the Wilcoxon test was lower – .957 – than the power of the t-test with .971 and the permutation t-test with .975.

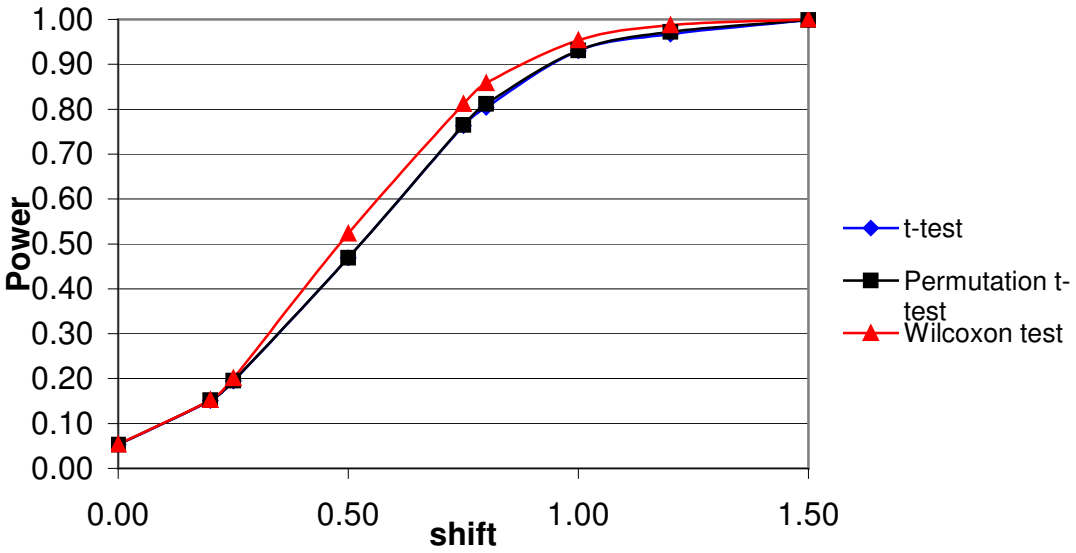

Figure 15: Shift vs. Power in the Chi-Square Distribution for Sample Sizes $n_1 = n_2 = 20$.

Figure 16 shows that the Wilcoxon test starts off being slightly more powerful than the t and the permutation t-test. Then these two tests become somewhat more powerful.

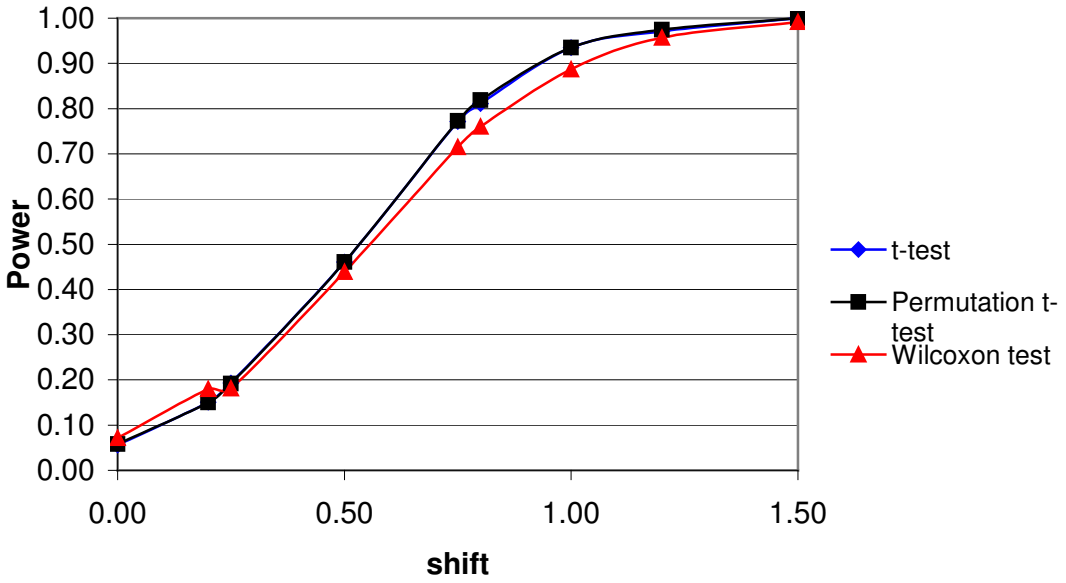

Figure 16: Shift vs. Power in the Multimodal Lumpy Data Set for Sample Sizes $n_1 = n_2 = 20$.

Table 5 summarizes the power results for sample sizes $n_1 = n_2 = 20$.

	Normal Distribution			Exponential Distribution			Chi-Square Distribution			Multimodal Lumpy Dist.		
Shift	t-test	Permut. t-test	Wil-coxon	t-test	Permut. t-test	Wil-coxon	t-test	Permut. t-test	Wil-coxon	t-test	Permut. t-test	Wil-coxon
0	0.053	0.053	0.051	0.052	0.051	0.051	0.053	0.053	0.053	0.056	0.058	0.071
.2σ	0.149	0.149	0.156	0.160	0.162	0.255	0.151	0.152	0.153	0.151	0.150	0.180
.5σ	0.454	0.454	0.433	0.496	0.497	0.712	0.469	0.469	0.523	0.461	0.461	0.439
.8σ	0.801	0.808	0.785	0.798	0.806	0.937	0.804	0.812	0.858	0.811	0.819	0.76
1.2σ	0.969	0.974	0.973	0.961	0.965	0.995	0.967	0.972	0.987	0.971	0.975	0.957

Table 5: Obtained Power for Different Shifts for Sample Size $n_1 = n_2 = 20$.

Large Unequal Sample sizes

This last simulation focused on sample sizes $n_1 = 10$ and $n_2 = 30$. Again, the same shifts were performed on the four distributions.

At the .2σ shift, this distribution gave powers of .141 for the t-test, .142 for the permutation t-test, and .137 for the Wilcoxon test. The next shift was .5σ. The powers obtained were .395 for both the t and permutation t-tests and .367 for the Wilcoxon test. The next section focused on the .8σ shift in the distributions that gave powers of .711 for the t-test, .709 for the permutation t-test, and .678 for the Wilcoxon test. At the 1.2σ shift, the normal distribution gave increasing powers of .922 for the t-test, .923 for the permutation t-test, and .934 for the Wilcoxon test.

The results for the normal distribution in the figure below show that all three tests begin with the same power. However, after the .2σ shift, the t and the permutation t-tests are slightly more powerful than the Wilcoxon test until they reach the shift of 1.2 σ where, again they have comparable power.

In the exponential distribution ($\mu = \sigma = 1$) with shift = .2σ, the powers increased: .148 for the t-test, .175 for the permutation t-test, and .223 for the Wilcoxon test. The next shift was .5σ. The powers increased for the three tests: .414, .456 and .565, respectively. At .8σ shift, the powers increased to .710 for the t-test, .743 for the permutation t-test, and .827 for the Wilcoxon test. At the 1.2σ shift, the exponential distribution resulted in a power of .914 for the t-test, .927 for the permutation t-test, and .957 for the Wilcoxon test.

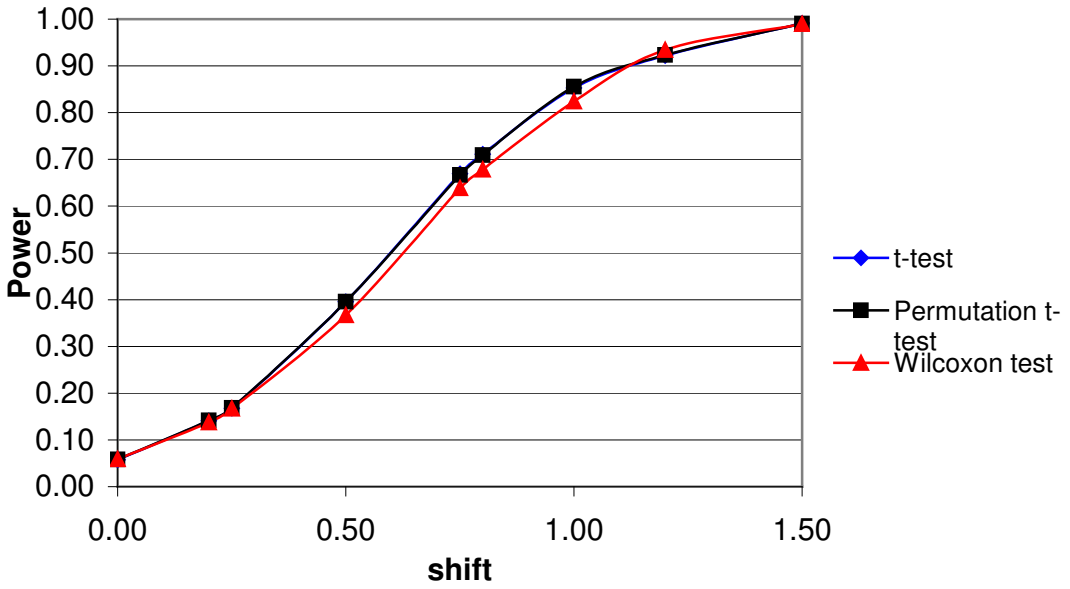

Figure 17: Shift vs. Power in the Normal Distribution for Sample Sizes $n_1 = 10$ & $n_2 = 30$.

Figure 18 reflects the superiority of the Wilcoxon test over the t and permutation t-tests. The permutation t-test is slightly more powerful than the t-test.

At shift = .2σ, the Chi-square distribution (df = 6) gave powers of .165, which was higher than .156 for the Wilcoxon test and .146 for the t-test in the Chi-square distribution. At .5σ shift, the powers slightly increased: .405 for the t-test, .432 for the permutation t-test, and .439 for the Wilcoxon test. At the .8σ shift of the distributions, the increasing powers were .699, .720, and .728 respectively. At the 1.2σ shift, the Chi-square distribution generated powers of .914, .921 and .927 respectively.

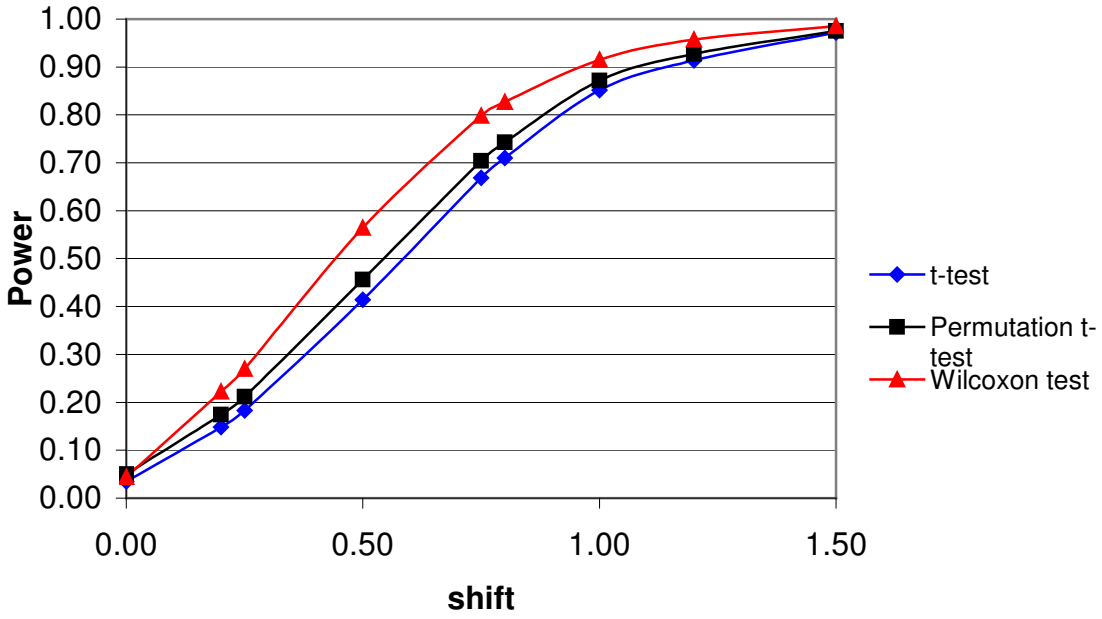

Figure 18: Shift vs. Power in the Exponential Distribution for Sample Sizes $n_1 = 10$ & $n_2 = 30$.

In the Chi-square distribution (df = 6) figure, the permutation t-test and the Wilcoxon test have comparable power. The t-test is slightly less powerful than the other two tests.

In the Multimodal Lumpy data set with the .2σ shift, the powers of the t and of the permutation t-tests were the same .121, which were lower than .149 for the Wilcoxon test. The next shift was .5σ. The power of the Wilcoxon was .350, lower than .379 for the t-test and .385 for the permutation t-test in the Multimodal Lumpy data set. At the .8σ shift in the distributions, the powers were .691 for the t-test, .693 for the permutation t-test, and .645 for the Wilcoxon test. For the last shift of 1.2σ, the powers were .923 for the t-test, .926 for the permutation t-test, and .896 for the Wilcoxon test.

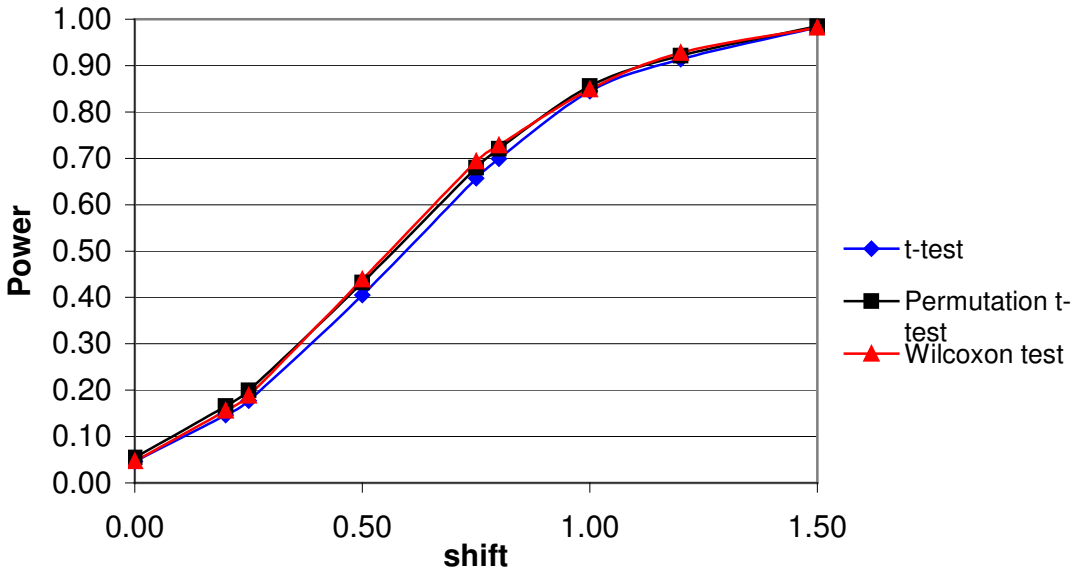

Figure 19: Shift vs. Power in the Chi-square Distribution for Sample Sizes $n_1 = 10$ & $n_2 = 30$.

Figure 20 is similar to the Multimodal Lumpy data set graph for sample sizes $n_1 = n_2 = 20$. It shows that the Wilcoxon test starts off being slightly more powerful than the t and the permutation t-test. Then it becomes less powerful.

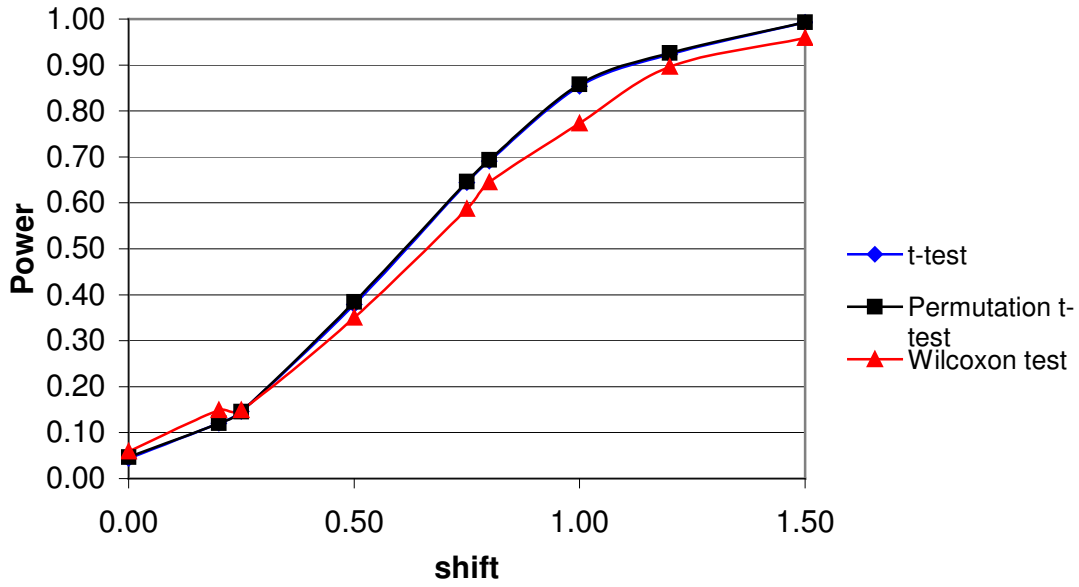

Figure 20: Shift vs. Power in the Multimodal Lumpy Data Set for Sample Sizes $n_1 = 10$ & $n_2 = 30$.

Table 6 summarizes the power results of the four distributions with the four shifts in the sample sizes with $n_1 = 10$ and $n_2 = 30$.

	Normal Distribution			Exponential Distribution			Chi-Square Distribution			Multimodal Lumpy Dist.		
Shift	t-test	Permut. t-test	Wil-coxon	t-test	Permut. t-test	Wil-coxon	t-test	Permut. t-test	Wil-coxon	t-test	Permut. t-test	Wil-coxon
0	0.057	0.058	0.059	0.035	0.050	0.059	0.046	0.055	0.047	0.044	0.047	0.059
.2σ	0.141	0.142	0.137	0.148	0.175	0.223	0.146	0.165	0.156	0.121	0.121	0.149
.5σ	0.395	0.395	0.367	0.414	0.456	0.565	0.405	0.432	0.439	0.379	0.385	0.350
.8σ	0.711	0.709	0.678	0.71	0.743	0.827	0.699	0.720	0.728	0.691	0.693	0.645
1.2σ	0.922	0.923	0.934	0.914	0.927	0.957	0.914	0.921	0.927	0.923	0.926	0.896

Table 6: Obtained Power for Different Shifts for Sample Sizes $n_1 = 10$ & $n_2 = 30$.

CHAPTER V

DISCUSSION

The interpretations of the results are divided into two major sections. The first part focuses on the explanations of the Type I error rate and the second section deals with the power comparison for the different distributions studied and the various treatment models.

Type I Error Rate

The million repetitions of the t-test for the Type I error yielded a rate of 5% for the normal distribution for the different sample sizes. The exponential, Chi-square distributions and the Multimodal Lumpy data set produced a Type I error below .05, which means that for these distributions, the t-test is conservative. However, the magnitude of the Type I error inflations was minor. These findings were expected as authors such as Sawilowsky and Blair (1992) previously demonstrated that the t-test is robust with regard to Type I error even under departure of normality for balanced layouts, large alpha levels, and for two tailed tests.

The only exception where the Type I error rate was inflated above 5% (.054) was found for the Multimodal Lumpy data set for the sample size of $n_1 = n_2 = 20$, although it is within Bradley's (1968) liberal definition of robustness (between .045 and .055). The Multimodal Lumpy data set has more than one peak and a few minima, which are likely factors leading to this result.

Permutation methods were then applied to the t-test in order to rehabilitate its Type I error properties. As the permutation t-test uses the 95^{tile} t-value as its critical value, it, by definition, restores the Type I error to .05. These findings reinforce the fact that the permutation t-test maintains the Type I error to the nominal α, regardless of the sample size and the distribution, in agreement with the literature from Edgington (1995) and Mielke and Berry (2001).

The Wilcoxon test gave a Type I error rate of .048 or .049 for all distributions and data set for all sample sizes except for $n_1 = n_2 = 10$, which was .045. This finding is in agreement with Fahoome and Sawilowsky (2000) where they found that the Wilcoxon test is robust with regard to Type I error regardless of the distributions or data set when samples sizes are greater than 14.

With regard to the generated permutation critical t-values compared to the tabled critical t-value for the sample size of $n_1 = 5$, $n_2 = 15$, the majority of the lower and upper quartiles above 1.734 for all the distributions and data set gave the Type I error that remains within Bradley's (1968) liberal definition of robustness. The exception lies within the upper quartile of the normal distribution at 3.7%. Below 1.734, the findings were mixed, i.e. some fit in the liberal robustness definition while others did not even adhere to this liberal robustness. However, none of the values within these quartiles fit the conservative definition of robustness.

Power Comparison

Small Equal Sample Sizes

For data sampled from the normal distribution, the results found in Figure 1 reflect that the t and permutation t-tests have approximately the same power, which is somewhat higher than the power of the Wilcoxon test, as expected (Blair & Higgins, 1980). However, Edgington (1995) and Good (1994) presume that the permutation t-test is considerably more powerful than nonparametric tests such as the Wilcoxon test. In the normal distribution with small equal sample sizes simulations, this assumption was not verified, as the permutation test only had the same modest increase over the Wilcoxon test as predicted by the Asymptotic Relative Efficiencies (ARE), which is in fact typically between .01 and .003.

The results for data sampled from the exponential distribution ($\mu = \sigma = 1$) indicated that the Wilcoxon test is much more powerful than the t and permutation t-tests, which was in agreement with theoretical speculation based on AREs by Hodges and Lehmann (1956), and small samples studies by Blair (1980), Blair and Higgins (1980b), Blair, Higgins and Smitley (1980), and Sawilowsky and Blair (1992). This finding confirms the fact that nonparametric methods are better suited for the detection of a treatment modeled as a shift in location parameter than the t-test or its permutation analog. This result replicates the findings of Blair and Higgins (1985)'s study, and contradicts statements made to the contrary by many other authors.

For data sampled from the Chi-square distribution (df = 6), the power of the t, permutation t and Wilcoxon tests are very similar. Between approximately .7σ and 1.2σ shifts, the power of the Wilcoxon test is slightly higher. This fact was observed because the chosen degree of freedom 6 for the Chi-square makes this particular distribution fairly symmetric. Sawilowsky and Blair (1992) showed spectacular gains of power for the Wilcoxon test in the Chi-square distribution with degree of freedom 1, 2 and 3. However, by time the degree of freedom is 6, the visual inspection of Figure 3 in Chapter III demonstrates that the distribution becomes more symmetric and thus, the t-test has recovered making the power of the Wilcoxon only slightly higher.

For data sampled from the Multimodal Lumpy data set, the power of the t and permutation t-tests is approximately the same. The power of the Wilcoxon tests begins at being as high as the two other tests' power, but at shift = .7σ, a drop is observed in the power. At 1.5σ shift, there is a ratio of .89 between the power of the t-test and that of the Wilcoxon. This finding is predictable, however, the Multimodal Lumpy data set, albeit with two modes, is relatively symmetric with light tails. This does suggest that skewness plays more of a role than kurtosis does with regard to the t-test's properties, as suggested in many places in the literature.

Small Unequal Sample Sizes

For data sampled from the normal distribution, the results from Figure 5 are in agreement with the literature where the t and permutation t-tests are most powerful even when the samples are unequal.

For data sampled from the exponential distribution ($\mu = \sigma = 1$), the results compiled in Figure 6 also agreed with the literature, in that the t-test is the least powerful test of the three. Therefore the statement from Edgington (1995) and Lu et al. (2001) is confirmed by the findings for the unequal sample size that the permutation t-test and the Wilcoxon test are more powerful than the t-test. However, these findings show a difference between the equal and unequal sample sizes. In this particular situation where the samples are not equal, the Wilcoxon test is only slightly more powerful than the permutation t-test.

The results for data sampled from the Chi-square distribution (df = 6) reflects the findings that were previously obtained from Edgington (1995) and Lu et al. (2001), i.e. the Wilcoxon test is slightly more powerful than the t-test, but slightly less powerful than the permutation t-test at least for smaller shifts. It is possible that these authors had chosen higher than 6 degree of freedom for the Chi-square. However, these results did not strongly agree with Sawilowsky and Blair (1992)'s study that found that the Wilcoxon test was much more powerful than the t and permutation t-tests. Again, this finding was obtained because the Chi-square distribution (df = 6) seems more symmetric than the one with 1, 2 or 3 degree of freedom.

Figure 8 depicts the results for data sampled from the Multimodal Lumpy data set. It reflects the same results as Figure 4 where the layout was equal. Therefore, the same logic used earlier is again applied here, i.e. the

Multimodal Lumpy data set is nevertheless essentially symmetrically with light tails.

Large Equal Sample Sizes

For data sampled from the normal distribution, Figure 9 reflects the information obtained from Bradley (1968), Blair and Higgins (1980), Kerlinger and Lee (2000), Noreen (1989), and others showing that the t and permutation t-tests have the same power and are the most powerful tests under normality. The graph reflects an ideal power curve, reaching the maximum power of 1.00.

Figure 10 depicts the results for data sampled from the exponential distribution ($\mu = \sigma = 1$) strengthens the findings reflected in Figure 2. However, the power ratio between the Wilcoxon and the two other tests is even higher at .83 compared to .81 for the smaller equal samples, at their highest points. These results replicate and confirm the findings from Sawilowsky and Blair (1992), which found that the Wilcoxon test is much more powerful than the t-test under non-normality, and extend it to show the same for the permutation version of the t-test. These results affirm the importance of the problem that states that nonparametric tests under non-normality are more likely to detect even small differences in treatments than parametric and permutation tests if the treatment changes the means of the two independent samples.

For the large samples of $n_1 = n_2 = 20$ from the Chi-square distribution (df = 6), Figure 11 reflects a slight difference than in the findings of Figure 3 for the small $n_1 = n_2 = 10$ samples. The power of the Wilcoxon test remains

somewhat superior to the two other tests. This fact is only partly in agreement with the findings from Sawilowsky and Blair (1992), and Zimmerman and Zumbo (1989), where the Wilcoxon test is expected to be much more powerful than the t-test, and extends that to the permutation t-test. Because of the partial symmetry of the Chi-square (df = 6), there is a fairly large power difference in the Wilcoxon test between the exponential and the Chi-square distributions.

For data sampled from the Multimodal Lumpy data set, Figure 12 reflected good power for the t and permutation t-tests. Although the larger sample sizes improved the power of the Wilcoxon test compared to that of the smaller samples, its power is still lower than the one from the two other tests. Again, this result is predicted because this data set is essentially symmetric with light tails, albeit with two modes.

Large Unequal Sample Sizes

For data sampled from the normal distribution, Figure 13 reflected the same results as for the large equal, small unequal and equal sample sizes. Again, the results are in agreement with Blair and Higgins (1980), Edgington (1995), Good (1994), Kerlinger and Lee (2000), Noreen (1989), and others, stating that the t and permutation t-tests are more powerful than its nonparametric counterpart when the referent distribution is normally distributed.

For the exponential distribution ($\mu = \sigma = 1$), however, Figure 14 showed well the difference in power between the three tests. The Wilcoxon test is far

superior to both the t and permutation t-tests under non-normality, as Blair and Higgins (1985), and Sawilowsky and Fahoome (2003) stated in their study. However, although its power is not as high as in the samples $n_1 = n_2 = 20$. The power of the permutation t-test is higher than the t-test when the data are skewed as Edgington (1995) and Lu et al. (2001) stated.

For data sampled from the Chi-square distribution (df = 6), Figure 15 reflected similar power for the three tests, with the t-test's power being slightly lower than Wilcoxon test's power. Yet again, this result was obtained because this distribution becomes relatively symmetric when the degree of freedom is 6.

Figure 16 depicts results for data sampled from the Multimodal Lumpy data set. It showed the same results as obtained from the large equal, small equal and unequal sample sizes where the t and permutation t-tests are somewhat more powerful than the Wilcoxon test. This is due to the fact that this data set is symmetric with light tails, albeit with multiple modes.

It is interesting to note that the power of the three tests, for the small equal sample sizes, does not exceed .92 in average. The highest power obtained is .95 in the Multimodal Lumpy Distribution for the t and permutation t-test when the shift = 1.5σ. In the case of the unequal small sample sizes, the highest power obtained is .89 at shift = 1.5σ in the Chi-square distribution for the permutation t-test. In the simulation where $n_1 = n_2 = 20$, the highest power obtained was 1.00, which is the maximum power that can be reached. All three tests' powers reached that point in all distributions. When the samples

are large but unequal, the power of the three tests reached .99. The results confirm that the power of any statistical test is related to the size of the samples, along with the significance level, the effect size, and the directional nature of the hypothesis (Hinkle et al., 2003).

Conclusion

The Type I error rate simulations may be summarized as follows: The t-test is robust with regard to the Type I error under normality and non-normality for larger and equal samples sizes, but not so when the sample sizes are small and unequal. The permutation t-test always maintains the Type I error to a nominal α. The Wilcoxon test is robust under normality and non-normality.

With regard to statistical power, the overall results of the simulations in the normal distribution show that, regardless of the size and evenness of the samples, the t-test is the most powerful test under normality, as expected, and that it has the same power as its permutation counterpart.

For the exponential distribution ($\mu = \sigma = 1$), the Wilcoxon test is, without a doubt, much more powerful regardless of the size and equality of the samples than the t-test or permutation t-test. Thus, when the data are skewed, the best test is certainly the Wilcoxon test for detecting even small effects in treatments as its power can be quite high.

By degree of freedom 6, the Chi-square distribution becomes more or less symmetric and rehabilitates the power of the t-test. Thus the Chi-square distribution simulations reflected a slightly more power for the Wilcoxon test when compared to the t and permutation t-tests for small sample sizes.

However, when the samples are large and equal where the Wilcoxon test becomes noticeable more powerful.

The interesting point was noticed for the Multimodal Lumpy distribution, a real data set. Indeed, the simulations for all four types of samples showed that the t and permutation t-tests are both equally powerful, and their power is higher than that of the Wilcoxon test. When determining possible effects in treatments, the test statistic recommended would be the permutation t-test. Not only is it as powerful as the t-test, it is also robust and maintains the Type I error to a nominal α, regardless of the size and equality of the sample treatment groups.

In summary, if the nature of the treatment changes the mean of two independent samples, the Student t-test is a suitable statistical test for the detection of effects under normality and when the samples are even. However, under non-normality, the Wilcoxon test is a better-suited statistical test, compared to the t and permutation t-tests with regard to robustness and power.

REFERENCES

Adams, D. C. & Anthony, C. D. (1996). Using randomization techniques to analyse behavioural data. *Animal Behaviour*, 54, 4, 733-738.

Albert, A., Bickel, P. J. & van Zwet (1976). Asymptotic expansions for the power of distribution-free tests in the one sample problem. *Anal Stat*, 4, 108-156.

Berger, V. W. (2001). The p-value interval as an inferential tool. *The Statistician*, 50, 1, 79-85.

Berger, V. W.; Lunneborg, C. E.; Ernst, M. D. & Levine, J. G. (2002). Parametric analyses in randomized clinical trials. *Journal of Modern Statistical Methods*, 1,1, 74-82.

Blair, R.C. (1980). *A comparison of the power of the two independent means t testto that of the Wilcoxon's rank-sum test for samples of various populations*. Unpublished doctoral dissertation. University of South Florida, Tampa, FL.

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

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

Blair, R. C., Higgins, J.J. & Smitley, W.D. (1980). On the relative power of the U and t tests. *British Journal of Mathematical and Statistical Psychology*, 33, 114-120.

Boik, R. J. (1987). The Fisher-Pitman permutation test: A non-robust alternative to the normal theory F test when variances are heterogeneous. *British Journal of Mathematical and Statistical Psychology*, 40, 26-42.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin*, 57, 49-64.

Borg, W. R. (1987). *Applying Educational Research: A Guide for Teachers*. White Plains, NY: Longman.

Bradley, J. V. (1968). *Distribution-Free Statistical Tests*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Campbell, D. T. & Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Boston, MA: Houghton Mifflin Company.

Chase, C. (1976). *Elementary Statistical Procedures* (2nd ed.). New York, NY: McGraw-Hill.

Commenges, D. (2003). Transformations which preserve exchangeability and application to permutation tests. *Journal of Nonparametric Statistics*, 15, 2, 171-185.

Edgington, E. S. (1980). *Randomization Tests*. New York, NY: Marcel Dekker, Inc.

Edgington, E. S. (1995). *Randomization Tests* (3rd ed). New York, NY: Marcel Dekker, Inc.

Feinstein, A. R. (1985). *Clinical Epidemiology: The Architecture of Clinical Research.* Philadelphia, PA: W. B. Saunders Company.

Fisher, R. A. (1935). *Design of Experiments*. Edinburgh, UK: Oliver and Boyd.

Fisher, R. A. (1958). *Statistical Methods for Research Workers* (13th ed.). New York, NY: Hafner Publishing Company, Inc.

Fisher, R. A. (1960). *Design of Experiments*. New York, NY: Hafner Pub. Company

Garrett, H. (1966). *Statistical Methods in Psychology and Education* (6th ed.). New York, NY: David McKay.

Glass, G., Peckham, P. & Sanders, J. (1972). Consequences of failure to meet assumptions in the fixed effects analysis of variance and covariance. *Review of Educational Research*, 42, 237-288.

Good, P. (1994). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New York, NY: Springer-Verlag New York Inc.

Good, P. (2002). Extentions of the concept of exchangeability and their applications. *Journal of Modern Applied Statistical Methods*, 1, 2, 243-247.

Hair, J. F.; Anderson, R. E.; Tatham, R. L. & Black, W. C. (1998). *Multivariate Data Analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Hinkle, D. E., Wiersma, W. & Jurs, S. G. (2003). *Applied Statistics for the Behavioral Sciences* (5th ed.). Boston, MA: Houghton Mifflin Company.

Heckel, D.; Arndt, S.; Cizadlo, T. & Andreasen, N. C. (1998). An efficient procedure for permutation tests in imaging research. *Computers and Biomedical Research*, 31, 164-171.

Hodges, J. & Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t test. *Annals of Mathematical Statistics*, 27, 324-335.

Hogg, R. V. & Craig, A.T. (1995). *Introduction to Mathematical Statistics.* Englewood Cliffs, NJ: Prentice-Hall.

Huber, P. J. (1977). *Robust Statistical Procedures*. Philadelphia, PA: SIAM.

Huber, P. J. (2003). *Robust Statistics*. New York, NY: Wiley, John & Sons, Inc.

Huberty, C. J. (1987). On statistical testing. *Educational Researcher*, 16(8), 4-9.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *The Journal of Experimental Education*, 61(4), 317-333.

Hunter, M. A. & May, R. B. (1993). Some myths concerning parametric and nonparametric tests. *Canadian Psychology*, 34, 4, 384-407.

Kempthorne, O. (1966). Some aspects of experimental inference. *Journal of American Statistical Association*, 61, 11-34.

Langbehn, D. R.; Berger, V. W.; Higgins, J. J.; Blair, R. C.; Mallows, C.L. (2000, February). Letters to the Editor. *The American Statistician*, 54, 85-88.

Lehman, E. L. (1975). *Nonparametrics*. San Francisco, CA: Holden-Day.

Lehman, E.L. & D'Abrera, H.J. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. New York, NY: McGraw-Hill.

Lehman, E. L. & Stein, C. (1959). *Testing Statistical Hypotheses*. New York, NY: John Wiley.

Lu, M., Chase, G. & Li, S. (2001). Permutation tests and other tests statistics for ill-behaved data: Experience of the NINDS t-PA stroke trial. *Communications in Statistics- Theory and Methods*, 30, 7, 1481- 1496.

Ludbrook, J. & Dudley, H. (1998). Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician*, 52, 2, 127-133.

Man, M. Z, Wang, X. & Wang, Y. (2000). POWER_SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics*, 16, 11, 953-959.

Manly, B. F. (1995). Randomization tests to compare means with unequal variation. *Sankhyā: The Indian Journal of Statistics*, 57, B, 200-222.

Marascuilo, L. A. & McSweeney, M. (1977). *Nonparametric and Distribution-Free Methods for the Social Sciences*. New York, NY: Brooks-Cole.

Maritz, J. S. (1981). *Distribution Free Methods*. London, England: Chapman and Hall.

McArdle, B. & Anderson, M. (2004). Variance heterogeneity, transformations, and models of species abundance: a cautionary tale. *Canadian Journal of Fisheries and Aquatic Sciences*, 61, 1294-1302.

Micceri, T. (1989). The unicorn, the normal curve and other improbable creatures. *Psychological Bulletin*, 105, 156-166.

Mielke, P. W. & Berry, K. J. (2001). *Permutation Methods: A Distance Function Approach*. New York, NY: Springer.

Neave, H. R. & Worthington, P. L. (1988). *Distribution-Free Tests*. London, England: Unwin Hyman, Ltd.

Neyman, J. & Pearson, E. S. (1928). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical inference. Part I. *Biometrika*, 20A, 175-240.

Nix, T. W. & Barrette, J. J. (1998). A review of Hypothesis testing Revisited: Rejoinder to Thompson, Knapp, and Levin. *Research in the Schools*, 5, 2, 55-57.

Noreen, E. W. (1989). *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. New York, NY: Wiley.

Nunally, J. (1975). *Introduction to Statistics for Psychology and Education*. New York, NY: McGraw-Hill.

Nunally, J. (1978). *Psychometric Theory* (2nd ed.). New York, NY: McGraw-Hill.

Opdyke, J. D.(2003). Fast Permutation tests that Maximize power under Conventional Monte Carlo sapling for pairwise and multiple comparisons. *Journal of Modern Statistical Methods*, 2, 27-49.

Rao, C. R. & Sen, P. K. (2002). Permutation scores tests for homogeneity of angular and compositional gaussian distributions. *Journal of Nomparametric Statistics*, 14, 4, 421-433.

Rasmussen, J. L. (1985). The power of Student's t and Wilcoxon W statistics: A comparison. *Evaluation Review*, 9, 4, 505-510.

Rogan, J. C. & Keselma, H. J. (1977). Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal?: An investigation via a coefficient of variation. *American Educational Research Journal*, 14, 493-498.

Rosenberger, W. F. & Rukhin, A. L. (2003). Bias propertiesand nonparametric inference for truncated binomial randomization. *Nonparametric Statistics*, 15, 445-465.

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

Sawilowsky, S. S. (2002). Fermat, Schubert, Einstein, and Behrens-Fisher: The probable difference between two means when $\sigma_1^2 \neq \sigma_2^2$. J*ournal of Modern Applied Statistical Methods*, 2, 461-472.

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, 353-360.

Sawilowsky, S.S., Blair, R.C. & Micceri, T. (1990). A PC Fortran subroutine library of psychological and education data sets. *Psychometrika*, 55, p.729.

Sawilowsky, S. S. & Fahoome, G. F. (2003). *Statistics Through Monte Carlo Simulation With Fortran.* Oak Park, MI: JMASM, Inc.

Walsh, E. O'F. (1968). *An Introduction to Biochemistry*. London, England: English Universities.

Weinbach, R. W. & Grinnell, R. M. (1997). *Statistics for Social Workers* (4th ed.). New York, NY: Addison-Wesley Educational Publishers, Inc.

Wilcox, R. R. (1996). *Statistics For The Social Sciences*. New York, NY: Academic Press, Inc.

Wolfowitz, J. (1949). Non-parametric statistical inference. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability* (pp. 93-113). Berkeley, CA: University of California Press.

Zaremba, S. K. (1965). Note on the Wilcoxon-Mann-Whitney statistic. *Annals of Mathematical Statistics*, 36, 1058-1060.

Zimmerman, D. W. (1996). A note on homogeneity of variance of scores and ranks. *The Journal of Experimental Education*, 64, 351-362.

Zimmerman, D. W. (1998). Invalidation of Parametric and Nonparametric Statistical Tests by Concurrent Violation of Two Assumptions. *The Journal of Experimental Education*, 67, 55-68.

Zimmerman, D.W. & Zumbo, B. D. (1993) Rank transformations and the power of the Student t test and Welch t test for non-normal populations with unequal variance. *Canadian Journal of Experimental Psychology*, 47, 523-539.

ABSTRACT

ROBUSTNESS AND POWER OF THE T, PERMUTATION T AND WILCOXON TESTS

by

MICHELE WEBER

December 2006

Advisor: Dr. Shlomo Sawilowsky

Major: Educational Evaluation and Research

Degree: Doctor of Philosophy

Data analysis is accomplished via parametric or nonparametric methods, depending on the data at hand. Authors have stated that parametric techniques are more robust with regard to Type I error and more powerful than nonparametric techniques. This statement is correct assuming that the numerous parametric requirements are verified, such as normality, homogeneity of variance and random selection and assignment among others. Nonparametric methods, however, are good alternatives to parametric methods as they are robust and powerful under non-normality. Although they have fewer assumptions compared to their parametric counterparts, they still have some requirements that need to be met, such as random selection and assignment. Permutation tests offer advantages compared to parametric tests as well, as they require fewer assumptions. They are distribution-free and exact statistics. It was found that they are robust with regard to Type I error and powerful. The problem resides in the fact that permutation tests maintain

the Type I error to the nominal α, however, there is no evidence that they are more powerful than nonparametric tests.

Monte Carlo simulations were used to investigate the Type I error and power of the t-test, permutation t-test and the Wilcoxon test for the normal, exponential (μ = σ = 1), Chi-square (df = 6) distributions and the Multimodal Lumpy, a real data set obtained from educational and psychological studies. It was found that, under normality, the t and permutation t-tests were robust with regard to Type I error compared to the Wilcoxon test. They were also slightly more powerful than the Wilcoxon test. However, under non-normality (especially as the departure from normality increases), the Wilcoxon test was, of course, robust with regard to Type I error and much more powerful than the t and permutation t-tests.

Michèle Weber LMSW, LMFT

37610 Bristol Court
Livonia, MI 48154
Email: rfweber@worldnet.att.net

Education:

2001 - **Master's of Social Work,** Clark Atlanta University, Atlanta, GA
1996 - **Bachelor of Science: Biology,** University of Massachusetts, Boston, MA

Professional Experience:

2006 – Present **Adjunct Professor -** Wayne State University, Detroit, MI

2004 –2006 **Co – Evaluator:** Team Member on Program Evaluation of The University Preparatory Academy and University Preparatory High School Mathematics and Science Evaluation Project. Gail Fahoome: Principal Investigator. Wayne State University, Detroit, MI

2004 – 2005 **Utilization Review Specialist** - Cruz Clinic PC, Livonia, MI

2001 – 2003 **Child and Family Therapist** -Southwest Counseling & Development Services, Detroit, MI

Presentations

2004 Michèle Weber. "How About Power?" – Examination of the importance of power in statistical analysis. AEA Annual Conference, Nov 04.

2002 Michèle Weber. Evaluating the Impact of Increased Caseloads on Social Workers: How can the System be Reformed. AEA Annual Conference, Nov. 02.

2001 Michèle Weber. Mainstreaming Evaluation. AEA Annual Conference, Nov.01.

Publications

2003 Michèle Weber – An Exploratory Study on the Effects of Perceived Gender Inequities on Financial Stress in Black marriages. *Race, Gender & Class Journal, Vol.10, #2.*

Professional Memberships

- 2001 – Present – American Evaluation Association
- 2001 – Present – National Association of Social Workers
- 2004 – Present – American Association of Marriage & Family Therapists
- 2006 – Present – Society for Social Work and Research
- 2006 – Present – Council on Social Work Education

Languages

- Fluent in French
- Fluent in spoken Haitian Creole
- Understanding of Written and Spoken Spanish
- Familiar with American Sign Language and Deaf Culture