THE DEPENDENT SAMPLES T AND WILCOXON SIGN RANK MAXIMUM TEST

by

SAVERPIERRE MAGGIO

DISSERTATION

Submitted to the Graduate School

of Wayne State University

Detroit, Michigan

in fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2012

MAJOR: EVALUATION AND RESEARCH

Approved by:

Advisor Date

DEDICATION

This dissertation is dedicated to (a) my wife Marisa for anxiously waiting for me to stop writing so I that I can give the kids a bath, (b) my children who anxiously waited for me to stop writing so that I can take them out of the bath, and (c) my parents who anxiously waited for me to just finish. They were behind all of my efforts and I am glad that were all able to see this day pass.

My wife allowed me the space and the time to complete what was needed to complete this project without any fuss. She in fact kept me in line. I wouldn't have completed this simple pleasure that life has afforded if it wasn't for her. Love you babe! My children will probably never forget me doing homework while they sat on my knees, hung on back, danced to loud music or asking me to proof read something that was due the next day et cetera and et cetera. My parents brought me into this world and I love them very much for doing so.

ACKNOWLEDGEMENTS

First and foremost I thank Dr. Sholmo Sawilowsky, my advisor. Dr. Sawilowsky attends to a great calling and that calling is being simply human. Dr. Sawilowsky is like no other and having the opportunity to work on a project with him I realized that there can't be any other. A kind man he is. I look forward to working together in the future with great interest.

I thank Dr. Gail Fahoome. Life is tough enough and Dr. Fahoome made life a little easier. She is kind, generous and patient. A special person indeed and I will never forget her.

I thank Dr. Stephen Brooks who has known me since undergraduate political science and at all times he was excited at the prospect of me completing my doctorate and made sure that I knew it. Looking back almost 26 years Dr. Brooks is optimistic as ever and if I had to do it all over he would certainly be called upon again.

I thank Dr. Tami Wright who joyfully participated on the committee. Her sense of humor at the beginning of my defense before it started set the tone suitably.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

Introduction

Ordinarily after an experiment is conducted a researcher will analyze the data that has been collected with a specific technique. To analyze obtained quantitative data a multitude of mathematical tools known as statistical tests of significance exist from which to choose. The purpose of a statistical test is to measure the consistency of data and to test the null hypothesis. There are two classifications of tests from which to select: parametric and nonparametric. Parametric tests are "statistical tests of hypotheses in which the null hypothesis includes a specific value for the population parameter and in which certain assumptions have been met" (Hinkle, Weirsman & Jurs, 1998, p. 620). Parametric tests are based on the following assumptions: (1) sample must be obtained from a population that is normally distributed; (2) variance is homogeneous among multiple groups; and (3) observations must be independent (Kerlinger & Lee, 2000). Nonparametric tests are not based on the assumption of normality. However they are based on the assumptions of random data selection, independence of observations and in some cases a continuous distribution.

To justify the use of a particular significance test factors such as study design, number of groups for comparison, type of data, easiness of computation, acceptability in the scientific community, and availability of tables of critical values can be used (MacDonald, 1999; Blair, 2002). However, misconceptions abound in the scientific community regarding certain statistical tests which if relied upon will lead researchers to wrongly choose one test over another (Sawilowsky, 2005). For example several misconceptions are known to exist in the scientific community surrounding the use of a particular t-test over a Wilcoxon test (Sawilowsky, 2005). Misconceptions must be sifted through and dispensed with prior to determining which test to use if the objective of a "sane" researcher is to avoid negative consequences for both the execution and the outcome of the research (Sawilowsky, 2005).

The null hypothesis, sometimes referred to as the hypothesis of no relationship, or difference, is the one that undergoes statistical testing (Howell, 1990; Isaac & Michael, 1997). The null hypothesis (H_o) is "a statement about some population parameter and the goal is to determine whether this statement is reasonable in light of actual data" (Wilcox, 1996, p.106). In the context of differences in means it is a statement that the means of two or more samples are drawn from the same population. Thus the results of testing provides the evidence a researcher requires in order to determine whether it is more appropriate to "fail to reject" or to "reject" the null hypothesis (Howell, 1990; Wilcox, 1996; Isaac & Michael, 1997).

Generally there are two types of errors that can be made when testing the null hypothesis: Type I and Type II. Both errors, but perhaps even more so Type I error, are considered embarrassing to the scientist if committed (Isaac & Michael, 1997). A Type I error, known as a false positive, occurs when two means are declared significantly different when in fact they are equal. In other words it occurs when a true null hypothesis is rejected. The probability of committing a Type I error is

$$P \text{ (Type I error)} = \text{significance level} = \alpha.$$

A Type II error, known as a false negative, occurs when a false null hypothesis is not rejected. Type I and Type II errors are inversely related meaning that as the probability of making a Type I error decreases the probability of making a Type II error increases and vice versa. Dudoit and van der Laan (2008) remarked that "ideally one would like to simultaneously minimize both the number of Type I errors and the number of Type II errors" however "this is not feasible as one seeks to trade- off between the two types of errors" (p. 18).

The probability of committing a Type II error is referred to as Beta (β). The probability of not committing a Type II error or rejecting the null hypothesis when it's false is referred to as the statistical power of a test (Cohen, 1992; Wilcox, 1996). Thus the statistical power of test is,

Power = 1 - probability of a Type II error (β).

Hence, statistical power is the probability that a statistical test will correctly reject the null hypothesis (Cohen, 1988). Insufficient statistical power hinders a researcher from making accurate conclusions regarding the null hypothesis by lowering the probability of detecting real effects and renders statistical outcomes to be insignificant (Dyba et al., 2005). Selecting a statistical test that is more powerful relative to another test not only reduces the probability of committing an error or Type I error but reduces the costs associated with experimentation as well (Good, 2005).

The most important factor to take into consideration when choosing a particular test, assuming it is robust with respect to Type I error, is comparative statistical power over its competitors (Blair, 2002; Gerke & Randles, 2010). Cox (1977) pointed out that in many circumstances the decision to use a particular statistic is handled by the "intuitive choice of a composite test statistics or by power considerations, compromising between the various kinds of alternatives" (p. 54). However it is unusual that one test be more powerful than its competitors under all plausible population models and usual that "one test or another may be more powerful than its primary competitor under a given set of circumstances" (Blair, 2002, p. 13). Gerke and Randles (2010) pointed out that,

> Unfortunately, exact power calculations are frequently unavailable, and when they are, they cannot be generalized due to the need for voluminous possibilities of sample sizes, magnitudes of alternative values, underlying distributions, and significance levels. One

> alternate approach is that tests be compared via the notion of asymptotic efficiency, where matters are simplified with the use of tests' asymptotic distributional properties. Such simplifications come at a price, however, as these asymptotic results may only approximate the finite sample size behavior of interest (p.1065).

Blair (2002) suggested that the interaction of certain factors in complex patterns is what precludes a researcher from concluding which test might be more powerful than another in certain situations. In the final analysis comparative statistical power as a determinant for the selection of a test may not offer enough information in order to make a conclusive choice between two or even more tests.

Consider the relative power of the parametric t-tests and the competing nonparametric Wilcoxon tests. The research has shown that neither the parametric t-tests nor the nonparametric Wilcoxon tests were more powerful than the other in all circumstances. At best the respective tests were found to be more powerful than the other in some circumstances. The parametric dependent samples t-test (t-test) and the nonparametric Wilcoxon signed ranks (WSR) are considered robust and powerful competitors of one another. Blair and Higgins (1985) compared the t-test to the WSR under various asymmetrical distributions and concluded that the often repeated and too frequently unqualified assertions made that the t- test was generally more powerful or efficient than the WSR test and that the WSR should be used instead of the t-test was clearly unjustified. In fact Blair and Higgins (1985) found no instance where "the t test held more than a modest power advantage over the Wilcoxon test with the largest advantage being only 0.105" (p. 126). Under normality the t-test was found to be more powerful than the WSR however the power advantage was too small to be of much practical importance (Blair & Higgins, 1985). The WSR was also found to be more often the more powerful test and the difference in favour of the WSR was often vast when the theoretical assumption of normality was relaxed (Blair & Higgins, 1985). For instance, under a chi square distribution the WSR held a clear power advantage over the t-test (Blair & Higgins, 1985). “Although the greatest advantage gained by the WSR test was 0.895, advantages greater than 0.1 were quite common” (Blair & Higgins, 1985, p.126).

Gerke and Randles (2010) in their simulation study compared the t-test to the WSR and found that the t-test offered a small advantage in efficiency over the WSR for detecting a shift under a uniform distribution. Gerke and Randles’ (2010) findings confirmed the results found in Blair and Higgins (1980) and Randles and Wolfe (1979). Gerke and Randles (2010) simulated 1,000,000 random samples of size 100 under local alternatives between 0.001 and 0.1. The results revealed a statistically significant power advantage of the t-test over the WSR test at shifts of 0.025 and above (Gerke & Randles, 2010). Wiederman and Alexandrowicz (2011) found that the t-test under a multitude of distributions was repeatedly less powerful than the WSR test (Arnold, 1965; Randles & Wolfe, 1979; Blair & Higgins, 1985). Wiederman and Alexandrowicz (2011) also found that the t- test proved to be robust under distributions such as Exponential, Weibul, Gumbel, Uniform, Laplace, Mixed Normal (symmetric unimodal), Mixed Normal (symmetric platykurtic), and Mixed Normal (asymmetric unimodal).

Also consider the relative power properties of the parametric independent t-test and its nonparametric competitor the Wilcoxon rank sum test. The power properties have been extensively and exhaustively studied (Hodges & Lehman, 1956; Chernoff & Savage, 1958; Boneau, 1960; Efron, 1969; Bradley, 1978; Blair & Higgins, 1980; Blair, Higgins & Smitely, 1980; Blair, 1981; Conover & Iman, 1981; Blair & Higgins, 1985; Rasmussen, 1985; Zimmerman, 1987; Zimmerman & Zumbo, 1989; Sawilowsky, 1990; Blair, 1991; Gibbons &

Chakraborti, 1991; Sawilowsky & Blair, 1992; Wilcox, 1996; Tanizaki, 1997; Bridge & Sawilowsky, 1999; Sawilowsky, 1999; Sawilowsky, 2005). It has been found that the Wilcoxon rank sum test was as powerful and in some cases more powerful than the independent t-test under a location shift model (Hodges & Lehman, 1956; Chernoff & Savage, 1958). Blair (1981) found that the independent t-test maintained its power when confronted with minimal violations to the assumption of normality. Sawilowsky and Blair (1992) concluded that unless certain conditions were met the independent t-test was non-robust to departures from population normality. Sawilowsky and Blair (1992) also noted that when certain conditions were met the independent t-test maintained only a small power advantage over the Wilcoxon-Mann-Whitney test (WMW). In comparison to the independent t-test, the WMW was found to be three or four times more powerful (Sawilowsky & Blair, 1992).

As it can be surmised the task of choosing a statistical test over another can be quite complicated, confusing and in some cases disappointing (Blair, 1985; MacDonald, 1999). Many textbook authors have provided practical advice to take into consideration when choosing between the noted tests in this study (Sawilowsky & Fahoome, 2003). However according to Sawilowsky and Fahoome (2003) the advice given is not without faults. They argued,

> In terms of practical application, many textbook authors suggest (1) the t test should be the preferred statistic, and either (a) data should be transformed to make them more "normal-like" prior to applying the t, or (b) preliminary tests of homogeneity and normality should be conducted first, followed by the t test, or (2) both the t and the Wilcoxon tests should be conducted, and the Wilcoxon test results should only be accepted if it is significant and the t test results are not significant (Sawilowsky & Fahoome, 2003, p. 227).

According to Sawilowsky and Fahoome (2003) the first suggestion is inherently weak because when the data is transformed the hypothesis to be tested must also be transformed leaving a researcher to guess which type of transformation is appropriate or the best. This is impossible because there are "approximately an infinite number of possibilities" (Sawilowsky & Fahoome, 2003, p. 288). With respect to the second suggestion Sawilowsky and Fahoome (2003) wrote,

> For every textbook author who suggests conducting both the t and the Wilcoxon, there is also a textbook author who points out that such a practice also leads to an inflation of experiment-wise Type I errors. In other words, if you have a data set, you get one shot at its analysis if the desire is to restrain the Type I error rate to nominal alpha (p.228).

Therefore according to Sawilowsky and Fahoome (2003) the second suggestion was also considered bad advice.

There is a solution to Type I error inflation however. To avoid inflating the Type I error rate and to avoid the daunting task of choosing one statistical test over another a researcher can conduct what is known as the "maximum test" (Cox, 1977; Algina, Blair & Coombs, 1995; Blair, 1991; Blair, 2002; Sawilowsky & Fahoome, 2003). Algina, Blair and Coombs (1995) defined the maximum test as a statistic "for a particular data set, two or more statistics and test the same hypothesis and selecting as the test statistic the one with the smallest *p* value" and in the event that "each statistic has the same critical value the maximum statistic is simply the most extreme of the calculated statistics" (p. 28).

The maximum test allows for the conducting of two or more tests without inflation the Type I error rate, in fact the test preserves Type I error to a nominal alpha without adding any new distributional assumptions (Sawilowsky & Fahoome, 2003). Cox (1977) referred to the maximum test as the most significant of the separate statistics. According to Cox (1977) the maximum test has useful diagnostic properties with extreme component statistics that is sensitive to any statistical departures from normal null hypothesis. This is the major advantage of the maximum test (Blair, 2002).

Several studies have been conducted analysing the statistical power of the maximum type tests (Tarone, 1981; Willan, 1988; Fleming & Harrington, 1991; Algina, Blair & Coombs, 1995; Lee, 1996; Ryan et al., 1999; Freidlin & Gastwirth, 2000; Blair, 2002; Freidlin et al., 2002; Weichert & Hothorn, 2002; Neuhauser et al., 2004; Yanga, Hsu & Zhaob, 2004; Salmaso & Solari, 2005; Yang et al., 2005; Kossler, 2010). Tarone (1981) examined the maximum of a log-rank statistic and modified Wilcoxon statistic. Lee (1996) considered a maximum or a linear combination of selected members of the family of weighted log-rank statistics. Freidlin et al., (1999) recommended that the maximum test should be used, especially when p <0.5. Gastwirth and Freidlin (2000) concluded that the maximum tests respective to their study were more powerful than the MERT and should be used to analyze extreme genetic models. Yanga, Hsu, and Zhaob (2004) considered the means of two standardized tests under contiguous alternatives and found that the maximum of the two standardized tests had an overall best performance relative to the other tests examined in the study. The results reported by to Algina, Blair and Coombs (1995) in general indicated that the maximum test had good Type I error rates and had much more power relative to the statistical power of a normal scores test and the WRS test.

Opdyke (2005) noted that the works of Fleming and Harrington (1991), Freidlin and Gastwirth (2000), Freidlin et al., (2002), Lee (1996), Ryan et al., (1999), Tarone (1981), Weichert and Hothorn (2002), Willan (1988), and Yang et al., (2005) all demonstrated the general purpose of the maximum test and that is "to trade-off minor power losses under ideal data conditions for a more robust statistic with large power gains across a wider range of possible and usually unknown data distributions" (pg. 37).

Statement of the problem

The relative statistical power of a test is the most important consideration a scientist could cite in order to justify the use of a particular test. The availability of tables of critical values is also an important reason to use in order to justify the use of a particular statistical test (Blair, 2002). Critical values are essential for inferential statistics and if critical values are not available a researcher will not be able to determine whether results obtained are statistically significant.

As of the date of this study a maximum test using the parametric dependent t-test and the non-parametric Wilcoxon sign rank test has not been created.

Purposes of the study

Therefore this study had three purposes; (1) to compute a maximum test using the parametric dependent t-test and the non-parametric Wilcoxon sign rank test through a FORTRAN program employing various subroutines of the International Mathematical and Statistical Libraries (IMSL, 1980); (2) to obtain critical values using sample deviates from a mixed normal distribution. Critical values obtained represented significance levels of 0.05, 0.025, 0.01 and 0.005 via sample sizes (n) 8 through 30, 45, 60, 90 and 120 for two- tailed test. The mixed normal distribution was chosen for this study for the same reasons given by Sawilowsky, Blair, and Higgins (1989) that is because it is familiar to many readers and because of its common use in robustness studies. The mixed normal distribution was formed by sampling with a probability of 0.95 from a normal distribution with mean of 0 and a standard deviation equal to 1 and with a probability of 0.05 from a normal distribution with a mean equal to 22 and a standard deviation of 10 as in Sawilowsky, Blair and Higgins (1989); and (3) to compare the critical values for sample sizes 8 and 120 to the Bonferroni adjustment.

Importance of the proposed study

It is incumbent upon a researcher to use the most powerful test when conducting an experiment. It is also incumbent upon a researcher to use a test that will not inflate Type I error. Therefore the present study contributes to the body of research in four ways. First, the maximum test using the dependent t-test and the Wilcoxon sign rank test is introduced which will assist researchers when choosing between using the dependent t-test or the wilcoxon. They will now be able to use both without concern for Type I error inflation. Second, a table of critical values for the maximum test is provided so that researchers will be able to refer to it in order to determine whether their study outcomes are statistically significant. Third, the body of research relating to the maximum test will be expanded. And, fourth, this study will add to the debate that exists in the scientific community as to the necessity of the Bonferroni method in the field of inferential statistics.

Limitations

This study had two limitations. The first limitation was due to the number of iterations conducted in order to compute the critical values of the maximum test. This study used Monte Carlo methods in order to obtain the results and therefore the accuracy of result improves as the number of pseudo-random numbers used increases (Upton & Cook, 2006s). Monte Carlo methods "refers to repeated sampling from a probability distribution to determine the long run average of some parameter or characteristic" (Sawilowsky & Fahoome, 2003, p. 46). Monte Carlo simulation "is the use of a computer program to simulate some aspect of reality, and making determinations of the nature of reality of change in reality through the repeated sampling via Monte Carlo methods" (Sawilowsky & Fahoome, 2003, p. 46). Therefore the accuracy of the critical values is dependent upon the number of iterations or pseudo-random numbers used. The study performed 200,000 iterations or pseudo-random numbers per sample size (*n*) and alpha (α). To increase the accuracy of the critical values more than 200,000 iterations would be required.

The second limitation is more of an acknowledgment of the WSR test's nature than a limitation of the study. Gibbons and Chakraborti (1991) stated that "the power functions of any two tests are not directly comparable unless the significance levels of both tests are exactly the same" (p. 259). In their comparison study of the relative power of the Student's t-test and the Mann-Whitney U (MWU) statistic, the statistical equivalent to the WRS, it was noted that the MWU is a discrete random variable and therefore the range of possible exact significance levels would be limited, especially for small sample sizes (Gibbons & Chakraborti, 1991). This means that "when each sample size is equal to 10, the only possible exact one tailed levels near .05 are .0446 and .0526" (Gibbons & Chakraborti, 1991, p. 259). Therefore it would not be fair to state that the t-test in this present study is more appropriate to use with smaller sample sizes over the WSR because the WSR cannot be used for sample sizes less than 8 for 0.05 levels as Gibbons and Chakraborti (1991) pointed out. For this reason the WSR can be used for levels less than 0.05 even if the norm has been to use it at both the 0.05 and 0.01. Thus for the purpose of this computational study critical values began at the sample size of 8 at 0.05, 0.025, 0.01 and 0.005 alpha levels. Since the maximum test under this study is a composite test made-up of the t-test and the WSR it was expected that the t-test and WSR would approximate each other as the sample sizes get larger. This expectation was met.

Definition of terms

Critical Value: A critical value is "an end point of a critical region. In a hypothesis test comparison of the value of a test statistic with the appropriate critical value determine the result of the test" (Upton & Cook, 2006, p. 108).

Experiment-wise Type I error: Experiment-wise Type I error is "the probability of making at least one Type I error among a group of tests conducted during some experiment" (Wilcox, 1996, pg. 257-256).

Monte Carlo Method: A method that uses "pseudo-random numbers in order to determine the properties of some function or set of functions" (Upton & Cook, 2006, p. 272).

Significance Level: Often noted as alpha which is the probability of making a Type I error (Upton & Cook, 2006, p. 200).

Type I error: To reject the null hypothesis of no differences when it is true (Isaac & Michael, 1997, p. 193).

Type II error: To accept the null hypothesis of no difference when it is false (Isaac & Michael, 1997, p. 193).

CHAPTER 2

Review of the literature

The maximum test

A review of the literature revealed that the dilemma of choosing one statistical test over another can be avoided through the use of a maximum test (Cox, 1977; Tarone, 1981; Willan, 1988; Fleming & Harrington, 1991; Algina, Blair & Coombs, 1995; Lee, 1996; Ryan et al., 1999; Freidlin & Gastwirth, 2000; Blair, 2002; Freidlin et al., 2002; Weichert & Hothorn, 2002; Sawilowsky & Fahoome, 2003; Neuhauser et al., 2004; Yanga, Hsu, & Zhaob, 2004; Salmaso & Solari, 2005; Yang et al., 2005; Kossler, 2010). Further the review of the literature revealed that no maximum test has been created via the parametric dependent sample t-test and the nonparametric Wilcoxon signed rank test and therefore no table of critical values exist. The following review of the literature canvassed studies that constructed maximum type test through other well-known statistical tests. The review of the literature revealed the following regarding the maximum test; (1) it is a conservative test, (2) it is a powerful test, (3) it maintains control of Type I error keeping it at the nominal level and (4) it does not add any new theoretical assumptions.

Kossler (2010) referred to the maximum test as a test whereby a scientist puts various score statistics together and takes the maximum of them. Algina, Blair and Coombs (1995) defined the maximum test as a statistic "for a particular data set, two or more statistics and test the same hypothesis and selecting as the test statistic the one with the smallest p value" and in the event that "each statistic has the same critical value the maximum statistic is simply the most extreme of the calculated statistics" (p. 28).

The maximum test was first attributed to Tippett (1931) by Cox (1977) because it was

Tippett who first proposed a test of statistical significance of combined results (Cox, 1977; Sen, 1985). However Cox (1976) presented the theoretical framework of the maximum test to the European Meeting of Statisticians in Grenoble. The presentation paper was published under the title "*The role of significance tests*" in 1977. Cox addressed Yates' condemnation of what Cox identified as the "overemphasis on tests of significance at the expense of interval estimation" (Cox, 1977, p. 49). Thus one aim of the presentation was to set out those "circumstances under which significance tests are likely to be valuable in the intermediate stages of an analysis or in the final statement of conclusions" (Cox, 1977, p. 50).

Cox (1977) described a significance test as a procedure for measuring the consistency of data with a null hypothesis and having the following form,

> Suppose that we have an observed vector, y, of response variables, sometimes written y_{obs}, and a null hypothesis H_o according to which y is the observed value of a random variable Y, which sample space S_y, and having probability density $f_Y(y)$ in some family H_o. The basis of a significance test is an ordering of the points in S_y in order of increasing inconsistency with H_o, in the respect under study. Equivalently there is a function $t = t(y)$ of the observations, called a test statistic, and such that the larger is $t(y)$, the stronger is the inconsistency of y with H_o, in the respect under study. The corresponding random variable is denoted by T (p. 50).

The computation of the observed level of significance was noted as

$$p(y_{\text{obs}}) - p_{\text{obs}} = \text{pr}\ (T \geq t_{\text{obs}} = t\ (y_{\text{obs}})\ ; H_o), \qquad (1)$$

where p_{obs} is the observed value of a random variable (Cox, 1977). The formal treatment of maximum test statistic is

$$q = \min\ (p_1 \ldots p_k), \qquad (2)$$

where *pj* is the significance level in the *jth* test and small values of q are evidence against H_0 (Cox, 1977). The required level of significance and the allowance for selection was noted as

$$\text{pr } (Q \leq q_{\text{obs}} ; H_0 = 1 - \text{pr } (P_j > q_{\text{obs}} ; j = 1, \ldots, k ; H_0). \qquad (3)$$

If component tests are independent and continuous (3) becomes $1 - (1 - q_{\text{obs}})^{\text{k}}$, and that an upper bound for (3) is in any case kq_{obs} (Cox, 1977).

According to Cox (1977) statistical tests may be derived from both simple and complicated circumstances. The maximum test statistic is derived out of the latter circumstance. In terms of simple situations Cox (1997) identified three main ways in which tests may be derived. The first way is by an absolute test which is applicable to simple null hypotheses involving discrete distributions. Under this first method the relationship between the probability of H_0 and the evidence against H_0 is inverse (the smaller is *pr* $(Y{=}y_{obs}; H_0)$, the stronger the evidence against H_0) (Cox, 1977). The second method is via choice of a test statistic on general grounds. The third method is via considerations of statistical power under alternative models chosen to represent the departures of interest. Both the second and third methods employ implicit and explicit probabilistic alternatives to the null hypothesis (Cox, 1977). The latter two methods are formally equivalent when a simple null hypothesis that has a density $f_o(y)$ with respect to an underlying measure $\mu(.)$ is considered (Cox, 1977).

In terms of complicated circumstances two situations were noted by Cox (1977) as being common occurrences. One common event is when the null hypothesis under investigation is composite. Composite null hypotheses are "those that specify the value of one component parameter and leave other parameters unspecified" (Cox, 1977, p. 54). Another common event occurs when the departure of several qualitatively different kinds from the null hypothesis is of interest.

The maximum test is derived out of the second most common occurrence. Cox (1977) presented four typical examples of departures that may be of interest. They are,

> (i) a distribution may depart from a null hypothesis of normality by skewness or kurtosis or both; (ii) the assumptions about error in the normal theory linear model may fail in numerous ways; (iii) the null hypothesis that random variables $Y_1 ...,Y_n$, are independently distributed in the standard normal distribution; (iv) the null hypotheses conventionally considered in univariate and multivariate analysis of variance can usually fail in multidimensional ways. Note that (iii) is not so academic as it might seem; the y_i's may be a similar test statistic calculated from independent sets of data (Cox, 1977, p. 54).

Cox (1977) concluded that the maximum test approximates an independent distribution under the null hypothesis and is able assess departures of qualitatively different kinds. Furthermore in terms of statistical power Cox (1977) concluded that the power of the maximum test in comparison to a quadratic statistic was found to be less powerful for some types of mixed departures. However it was found to be more powerful for "pure" departures along the component directions. When $k = 2$ Cox (1997) considered power differences to be unimportant. However when $k = 10$ and if departures are in several directions Cox (1977) advised that an appreciable loss of power may result. Solmaso and Salari (2005) explained why according to Cox (1977) findings that the power of the maximum test in comparison to a quadratic statistic was less powerful for some types of mixed departures from Ho and more powerful for "pure" departures along component directions as a "consequence of the fact that when applying the Bonferroni-adjustment, there is no chance of gathering information from the several applied tests since only the smallest p-value is of interest for the test decision rule" (p.332).

In symmetrical situations the implication is that at the level "α the affective power approaches α /k for alternatives very close to the null hypothesis" (Cox, 1977, p. 55). When one tests for univariate normality the max [| $h\frac{(1)}{n}\left(g\frac{1}{2}\right)$ | , | $h\frac{(2)}{n}(g_2)$], according to Cox (1977) may be taken "where *gl* and *g2* are the sample standardized third and fourth cumulants and the functions $h\frac{(1)}{n}(.)$and $h\frac{(2)}{n}(.)$ respectively transform *g1* and *g2* into standard normal variables" (p. 55). In a factorial experiment, let $MS^{(i)}$,...,$MS^{(k)}$ be the independent mean squares, and let MS_{res} be an independent estimate for error with f_{res} degrees of freedom, the maximum statistic is max $\{MS^{(i)}\}/MS_{res}$ for the case $f = 1$ (Cox, 1977).

Algina, Blair and Coombs (1995) examined type I error rates and the power of the a maximum test constructed through the use of the O'Brien's and the Brown-Forsyth test. The researchers estimated Type I error rates and power for the Brown-Forsthe test, O'Brien test and the maximum test. According to Algina, Blair and Coombs (1995) past studies pointed to these two tests as robust tests of scale (Brown & Forsythe, 1974; O'Brien, 1979; Conover, Johnson, & Johnson, 1981; Olejnik & Algina, 1987; Ramsey & Brailsford, 1989; Algina, Olejnik, & Ocanto, 1989). However power differences favored either the O'Brien or the Brown- Forsyth under various distributional conditions thereby complicating matters when choosing one of the tests over the other. For instance the researchers reported that O'Brien (1978, 1979) indicated "that with equal-sized samples power comparisons favor his test with short-tailed distributions and the Brown-Forsythe test with long tailed distributions" (p. 27). Algina, Blair and Coombs (1995) reported that in a two-sample case study where sample sizes were not equal power comparisons depended on the relationship between the sample sizes and population variances. Algina, Blair and Coombs (1995) noted that,

> When the relationship between variances and sample sizes were inverse, the power of the Brown- Forsythe test was more severely depressed than was the O'Brien's test. As a

> result, O'Brien's test was more powerful than the Brown-Forsythe when the data were sampled from either platykurtic distributions or mesokurtic distributions; the two tests had more similar power when applied to data sampled from leptokurtic distributions. When the relationship was direct, the power of O'Brien's test was more severely depressed. Consequently, the Brown-Forsythe test was more powerful with data sampled from either leptokurtic distributions or mesokurtic distributions; when the data were sampled from platykurtic distributions, the procedures had more similar power" (p. 27).

Thus researchers remarked that choosing between the O'Brien's and the Brown-Forsyth can be complicated since "sample kurtosis may be an unreliable basis for selecting between the two tests" (p. 28). Therefore an alternative the researchers used an alternative the maximum test. In Algina, Blair and Coombs (1995) Type I error rates for the maximum test were near nominal level. The Brown-Forsythe was defined as an "analysis of variance using $|x_{ij} - \hat{\Theta}_j|$ where $\hat{\Theta}_j$ the median for the *j*th sample" (p. 27). The O'Brien test was defined as "an analysis of variance using $[(n_j - 1.5)\, n_j\, (\overline{x}_{ij}\ - x_j)^2 - .5S_j^2(n_j - 1)]/[(n_j - 1)(n_j - 2)]$" (p. 27).

The purpose of the Algina, Blair and Coombs (1995) study "was to evaluate a maximum statistic computed using O'Brien's test statistic and the Brown- Forsythe test statistic applied to two samples" (p.28). The problem was defined as follows,

> Let x_1 and x_2 be independent random variables and let σ_j and θ_j denote the standard deviation and median for the *j*th variable. Assume that the cumulative distribution functions of x_1, and x_2 are F_{x1}, (t) = F_{x2} ((t - μ*)/σ*) where μ* and σ* are arbitrary location and scale parameters. The null hypothesis tested by the maximum test is H_0: $\sigma^* = 1$ which is equivalent to H_0: $\sigma_1 = \sigma_2$ and to $H_0 : \mathrm{E}|x_1 - \theta_1| = \mathrm{E}|x_2 - \theta_2|$ (Algina, Blair and Coombs, 1995, p. 28).

Algina, Blair and Coombs (1995) found that under a normal distribution using alpha 0.05, equal *n*, the power results for two-tailed tests were very similar. "In conditions in which the maximum test had more power, both the O'Brien and Brown- Forsythe tests tended to be conservative. Similar results were observed for α = .01 and .10 and for directional tests" (Algina, Blair and Coombs, 1995, p. 32). When *n* was unequal and data sampled were from a normal distribution power results indicated the advantage of the maximum test. The power of the maximum test mimicked that of the Brown-Forsythe test when there was a direct relationship between n_j and σ_j, and mimics that of O'Brien's test when the relationship was indirect (Algina, Blair and Coombs, 1995). When data were sampled from platykurtic and leptokurtic distributions the maximum test mimicked the O'Brien's test when data were short-tailed and mimicked the Brown- Forsythe when the data were long-tailed. "Similar results occurred for α = .01 and .10 and for directional tests" (Algina, Blair and Coombs, 1995, p. 33). According to the resulted reported by Algina, Blair and Coombs (1995) the maximum test under some conditions was the most powerful test and in others it was equal or just below. The results reported by to Algina, Blair and Coombs (1995) indicated that the maximum test had good Type I error rates and had "much more power than the more powerful of the Brown-Forsythe and O'Brien's tests" (p. 37).

Blair (2002) presented a maximum test statistic that was used in lieu of the Wilcoxon Rank-Sum test and the Terry-Hoeffding tests. According to Blair (2002) these two tests are considered in the research as robust and powerful competitors of one another. The Terry-Hoeffding test was noted as the "normal" scores counterpart to the Wilcoxon Rank-Sum test. Blair (2002) explained,

> Both procedures are used to test the null hypotheses that are from a common population. Asymptotic results suggest that these two tests may manifest substantial power

> differences, with the magnitudes and advantages of such difference depending on the shape of the population. Asymptotic Relative Efficiencies (ARE) indicate that, in general when alternatives are expressed as simple shifts in location, the normal scores test is more efficient than the rank test when sampling is from a light tailed distribution. However, the normal scores were at a disadvantage when the populations are heavy-tailed (p. 13).

The new maximum statistic was defined as,

$t_{max} = \ \{t_W ; |t_W| \geq |t_{NS}\}$

$\{t_{NS}; |t_{NS}| \geq |t_W\}.$

In the event scores tied, $|t_W| = |t_{NS}|$, then either Wilcoxon Rank-Sum test or Terry-Hoeffding test was used. In terms of Type I error Blair (2002) concluded that,

> (1) The t-distribution provides a good approximation for the distribution of t_W and t_{WS} is reaffirmed. (2) Critical values...produce Type I error rates for t_{MAX} near nominal levels both in the case of $n_1 - n_2$. (3) Referencing t_{MAX} to a t- distribution with $n_1 + n_2$ - 2 degrees of freedom results in only minor Type I error inflations (p. 18).

In terms of power under different distributions such as normal, uniform, and Cauchy distributions, Blair (2002) found that under a normal distribution there was "little difference in the power of the three tests. However Blair (2002) also found that the

> [d]ifferences that did occur favoured t_{NS} and t_{MAX}. In the case of the uniform distribution t_{NS} was the most powerful test, with t_{MAX} showing power similar to but slightly less than, that of t_{NS}. Under the heavy-tailed Cauchy distribution, t_W was the most powerful statistic, with t_{MAX} once again demonstrating power similar to, but slightly less than, that

> of the most powerful test (p. 17).

Blair (2002) concluded that the inflation of the Type I error under t_{MAX} was only minor. In terms of power the t_{MAX} displayed a major advantage. Blair (2002) noted, that major advantage of the maximum test was "the fact that the test is automatically adaptive to the weight in the tail of the population from which the data were sampled" (p. 17).

Sawilowsky and Fahoome (2003) referred to the work of Blair and Higgins (1992) wherein the researchers formulated a maximum test statistic using both the t-test and the Wilcoxon rank-sum test. The maximum test was computed as follows,

> Random variates were obtained from a normal distribution and randomly assigned to two groups. Both the t and the Wilcoxon were computed. The Wilcoxon test was computed in the same metric as the t by replacing the original scores with their ranks, and calculating the t statistic on those ranks. The obtained t on the original data (ordinary t statistic) and the obtained t on the ranks (Wilcoxon equivalent statistic) were compared, and whichever obtained statistic was higher, or "maximum", was recorded. This process was repeated on million times. The vector of one million maximum values was then sorted from low to high (p. 229).

Critical values obtained by Blair and Higgins (1992) were presented by Sawilowsky and Fahoome (2003). Sawilowsky and Fahoome (2003) reported that the maximum test preserved the Type I error rate to nominal alpha because it assumes that only one test is being computed. Furthermore, the maximum test did not require any new assumptions (Sawilowsky & Fahoome, 2003).

Neuhauser et al., (2004) pointed out that the maximum test was useful for all sample sizes. The researchers compared adaptive tests based on a selector statistic with a maximum test and a sum test for a nonparametric two sample problem. The problem investigated is repeated here,

> Let X_1,. . . ,X_n and Y_1, . . ., Y_m denote two random samples. The observations within each sample are independent and identically distributed, and we assume independence between the two samples. Let F_1 and F_2 be the continuous distribution functions corresponding to populations 1 and 2, respectively. In the location-shift model considered here the distribution functions are the same except perhaps for a change in their locations; that is, $F_1(t) = F_2(t - \theta)$ for every t. The null hypothesis is H_0: $\theta = 0$, whereas the alternative states $\theta \neq 0$ (Neuhauser et al., 2004, p. 215).

Neuhauser et al., (2004) found that for unrestricted continuous distribution the sum test was not robust. However the adaptive tests and the maximum test were robust (Neuhauser et al., 2004). The researchers also found that when the Maximum Efficiency Robust Test and the maximum test were compared the maximum test was found to be superior. Neuhauser et al., (2004) reported,

> The Maximum Efficiency Robust Test (MERT) maximizes the minimum asymptotic efficiency over the possible tests. However, for different test problems the MERT was not superior to a maximum test (Freidlin et al.,, 1999, 2002; Neuhauser & Hothorn, 1999; Gastwirth & Freidlin, 2000; Zheng et al., 2002), especially when the family of models is broad (p. 225).

The researchers also compared the maximum test against a family of t distributions and found p^* <0.7 and the MERT being not as power as the maximum test.

Solmaso and Salari (2005) referred to the maximum test as a conservative approach to the main issue in multiple analysis or multiple hypotheses testing that being the "problem of interpreting several statistical tests within a single research effort" (p. 331). Solmaso and Salari (2005) referred to Fisher's (1935), The Design of Experiments, to theorize that "different tests of significance are appropriate to test different features of the same null hypothesis" it may happen that several tests could be applied to the same data, each test corresponding to a distinct kind of deviation from H_o" (p. 331). The maximum test by design calculates the "overall *p*-value, $q = k \cdot \min (\lambda_1, \ldots, \lambda_k)$, where q is an upper bound for the significance level, and λi , $i - l, \ldots, k$, is the *p*-value associated with the *i*-th test" (p. 332). Furthermore in terms of statistical power Cox (1977) concluded that the power of the maximum test in comparison to a quadratic statistic was found to be less powerful for some types of mixed departures. However it was found to be more powerful for "pure" departures along the component directions. When $k = 2$ Cox (1997) considered power differences to be unimportant. However when $k = 10$ and if departures are in several directions Cox (1977) advised that an appreciable loss of power may result. This fact is a consequence of the application the Bonferroni-adjustment as "there is no chance of gathering information from the several applied tests since only the smallest p-value is of interest for the test decision rule" (Solmaso & Salari, 2005, p.332). Solmaso and Salari (2005) further maintained Cox (1977) in that the maximum test can be used to give a very direct measure of the significance probability where just one of the λi is associated with a real significant departure from H_o (Cox, 1977).

Interestingly Solmaso and Salari (2005) commented on Spjotvoll criticism of Cox (1977). Spjotvoll stated that "the imaginative statistician can think of many relevant test statistics, while the one with less imagination may produce one or two. Seemingly the imaginative one is penalized since he has to consider a given significance probability as less significant than his colleague just because he started out with a large number of statistics". Solmaso and Salari

(2005) response to Spjotvoll statement was that “Bonferroni-type adjustments require that each single test produces *p*-values as small as α/k occurring at a frequency reasonably close to α/k when the null hypothesis is true” (p. 332).

Opdyke (2005) pointed to studies conducted by Tarone (1981), Willan (1988), Fleming and Harrington (1991), Lee (1996), Ryan et al., (1999), Freidlin and Gastwirth (2000), Freidlin et al., (2002), Weichert and Hothorn (2002), and Yang et al., (2005) as research that demonstrated the purpose of the maximum tests that being to "trade-off minor power losses under ideal data conditions for a more robust statistic with larger power gains across a wider range of possible and usually unknown data distributions” (pg. 275-376). Opdyke (2005) analyzed distributions of continuous-data and argued that maximum test would be useful than the tests recommended in Opdyke (2004) for the following reasons: 1) using only one statistical test unarguably would be more straightforward to implement than (potentially) relying on the four statistics; 2) possible power losses may be small or negligible; and 3) depending on conditions and tests used, the maximum statistic may be even more powerful. Opdyke (2005) found that maximum test employed maintained reasonable Type I error control and was always either nearly as powerful and almost more often even more powerful than constituent tests. Furthermore the maximum test displayed dramatic power gains over the modified t and superiority in detecting disparity (Opdyke, 2005).

Kossler (2010) obtained maximum type tests based on U-statistics. Kossler (2010) examined a two-sample location problem where two types of adaptive tests were considered. One type was based on linear rank tests with various scores and the other was based on U-statistics. The asymptotic and finite power properties of the adaptive and the maximum tests were investigated. Kossler (2005) derived the maximum tests from Neuhäuser et al. (2004).

Kossler (2005) defined the maximum test as an idea wherein the maximum test is obtained by putting "various scores statistics together, and to take the maximum of them" (p. 2053). Kossler (2010) concluded,

> [T]hat not only the adaptive tests are the best in most cases, but also that a simple max-type test based on linear rank statistics or on U-statistics may be useful. The tests MAX'$_4$, MAX$_5$ and MAXU_3 are the best among all considered max-type tests, where the test MAX'$_4$ requires an initial estimate of skewness. Max-type tests based on linear rank statistics and such based on U-statistics are of the same value. Summarizing, for the case of an unknown density, all of the four variants, adaptive tests and max-type tests based on linear rank tests or on U-statistics, are justified. The adaptive tests are asymptotically the best, and they are proposed for larger sample sizes (about $n_i \geq 40$). For small sample sizes (about $n_i \leq 20$) the max-type tests are to prefer. Linear rank statistics are simpler, but U-statistics may better smooth the effect of extreme densities (p. 2065).

In other words the adaptive tests had a larger asymptotic power than the maximum tests. However maximum tests where samples were small they were found to be preferable.

CHAPTER 3

Methodology

For the purposes of the present study Monte Carlo methods were used. Monte Carlo methods "refers to repeated sampling from a probability distribution to determine the long run average of some parameter or characteristic" (Sawilowsky & Fahoome, 2003, p. 46). Monte Carlo simulation "is the use of a computer program to simulate some aspect of reality, and making determinations of the nature of reality of change in reality through the repeated sampling via Monte Carlo methods" (Sawilowsky & Fahoome, 2003, p. 46).

Computation of maximum test

Using a Compaq Fortran 6.6c program employing various subroutines of the International Mathematical and Statistical Libraries (IMSL, 1980) sample deviates were obtained from a mixed normal distribution for the purposes of computing and obtaining critical values. The mixed normal distribution was formed by sampling with the probability 0.95 from a normal distribution with mean of 0 and standard deviation equal to 1, and with the probability 0.05 from a normal distribution with mean equal to 22 and standard deviation of 10 as in Sawilowsky, Blair and Higgins (1989).

Critical values were obtained as follows; Random variates were assigned to two groups. Both the t-test and the WSR test were computed on the obtained scores. The Wilcoxon test was then computed in the same metric as the t-test then the t-test was calculated. The obtained t on the original data and the obtained t were compared, and whichever was higher was recorded. This process was repeated 200,000 times. The vector of 200,000 was then sorted from low to high. Critical values obtained represented values at the 0.05, 0.025, 0.01 and 0.005 significance levels via sample sizes (*n*) 8 through 30, 45, 60, 90 and 120 (See Table C).

Illustrated use of the maximum test

The following illustration demonstrates the usage of the maximum test. The illustration was directly adapted from Sawilowsky and Fahoome (2003). Consider testing the difference in average performance of a treatment versus a control group where $n1 = n2 = 20$, for a two-sided test with an α level at 0.05. The first step is to conduct both the dependent t-test and the Wilcoxon sign rank test. The second step is to select whichever obtained statistic is higher in magnitude (i.e., select the statistic whose absolute value is greater). The third step is to enter Table C with n=20, α at 0.025 retrieve the critical value ± 2.189228. If the obtained maximum statistic is either greater than 2.189228, or if it is less than -2.189228, reject the null hypothesis in favour of the alternative.

CHAPTER 4

Results

Critical values for the maximum test for sample size 8 and 120 are contained in both Table A and B respectively and were computed alongside critical values of the theoretical t and the Bonferroni method. The Bonferroni test was calculated through the formula α/n. The Bonferroni method is based on probability inequality and attempts to ensure "that the probability of rejecting at least one hypothesis when all are true is no greater than α" (Simes, 1986, p. 751).

Table A below contains the critical values for the theoretical t-test, the maximum test and the respective Bonferroni adjustment for α levels 0.05, 0.025, 0.01 and 0.005 where n= 8. These critical values are based on a Monte Carlo study with 200,000 replicates per n and α level combination for a two tailed test.

Table A- Critical values for the theoretical t, maximum test and bonferroni (n= 8, 2 tailed test).

Alpha	Theoretical *t*	Maximum test	Bonferroni
0.05	2.364625	2.403852	2.841245
0.025	2.841244	2.886371	3.335292
0.01	3.499484	3.542152	4.029338
0.005	4.029337	4.076645	4.594619

Table B below contains the critical values for the theoretical t-test, the maximum test and the respective Bonferroni adjustment for α levels 0.05, 0.025, 0.01 and 0.005 where n= 120. These critical values are based on a Monte Carlo study with 200,000 replicates per n and α level combination for a two tailed test.

Table B. Critical values for the theoretical t, maximum test and bonferroni (*n*= 120, 2 tailed test).

Alpha	Theoretical t	Maximum test	Bonferroni
0.05	1.98010	2.151836	2.270117
0.025	2.270117	2.427165	2.536239
0.01	2.617776	2.726720	2.860317
0.005	2.860317	2.920707	3.089022

Table C below contains the critical values for the maximum test for *α* levels 0.05, 0.025, 0.01 and 0.005 for sample sizes 8 through 30, 45, 60, 90 and 120. The critical values are based on a Monte Carlo study with 200,000 replicates per *n* and α level combination for a two tailed test.

Table C. Critical Values for the dependent samples t and wilcoxon sign rank maximum test (two tailed test).

Sample size (n)	0.05	0.025	0.01	0.005
8	2.403852	2.886371	3.542152	4.076645
9	2.035884	2.421404	2.967075	3.244028
10	2.026771	2.345177	2.863018	3.214813
11	1.997151	2.336451	2.768128	3.128150
12	1.992236	2.322638	2.738997	3.106802
13	2.117835	2.282598	2.674553	2.977316
14	2.159492	2.234117	2.695154	2.990559
15	1.952256	2.255887	2.648058	2.939312
16	1.925327	2.233314	2.616452	2.894716
17	2.264775	2.264775	2.589332	2.864603
18	2.294804	2.294804	2.584305	2.837560
19	1.971807	2.191765	2.551924	2.834909
20	2.003801	2.189228	2.539654	2.798223
21	2.192236	2.373854	2.499491	2.794317
22	2.163950	2.397194	2.524580	2.752670
23	2.088616	2.163827	2.480621	2.765330
24	2.113799	2.171688	2.494098	2.759386
25	2.109428	2.460219	2.484072	2.731530

26	2.099438	2.479246	2.497775	2.753092
27	2.096300	2.182043	2.497411	2.752280
28	2.090605	2.202707	2.475614	2.696693
29	2.079096	2.531521	2.531521	2.726458
30	2.074818	2.520912	2.547550	2.711261
45	2.146091	2.349346	2.732897	2.732897
60	2.155180	2.351812	2.600462	2.745563
90	2.215773	2.454552	2.718683	3.023477
120	2.151836	2.427165	2.726720	2.920707

CHAPTER 5

Discussion

The review of the literature regarding the maximum test revealed the following about the test; (1) it is a conservative test, (2) it is a powerful test, (3) it maintains control of Type I error by keeping it at the nominal alpha, and (4) it does not add any new theoretical assumptions.

The maximum test computed in this study was computed through the use of the dependent t-test and the Wilcoxon sign rank test under a mixed normal distribution. Mixed normal distributions are considered important population models across a variety of disciplines. Blair and Higgins (1980) noted that it is to find forms of mixed normal distributions. The Pitman efficiencies of these distributions, as noted by Blair and Higgins (1980), have large power advantages for the Wilcoxon statistic relative to the t statistic. The mixed normal distribution used was formed by sampling with the probability 0.95 from a normal distribution with the mean 0 and the standard deviation equal to 1, and with the probability 0.05 from a normal distribution with the mean equal to 22 and the standard deviation of 10 (Sawilowsky, Blair & Higgins, 1989). The maximum test could be used for any model of mixed normal populations, or for that matter, any situation when the population is not normally distributed.

Table A presented critical values for the sample size of 8. Table B presented critical values for the sample size of 120. Table C presented all of the obtained critical values for this study. Tables A and B contained the critical values for the theoretical t-test, the maximum test and the Bonferroni adjustment for α levels 0.05, 0.025, 0.01 and 0.005. Table C contained critical values for the maximum test for sample sizes 8 through 30, 45, 60, 90 and 120 for the α levels 0.05, 0.025, 0.01 and 0.005.

Table A contains the critical values for the theoretical t-test, the maximum test and the

Bonferroni adjustment for α levels 0.05, 0.025, 0.01 and 0.005 where n=8. The maximum test produced values almost equivalent to the theoretical t-test with the exception of the sample size 120 at the 0.05 α level. The critical value for the theoretical t-test was 1.9801 and the comparative value for the maximum test was 2.270117. For all α levels 0.05, 0.025, 0.01 and 0.005 for the sample size of 8 the corresponding theoretical t-test values obtained were 2.364625, 2.841244, 3.499484, and 4.029337 respectively. The maximum test provided 2.403852, 2.886371, 3.54212 and 4.076645 respectively for the same α levels. The Bonferroni values were 2.841245, 3.335292, 4.029338, and 4.594619 for the same respective α levels.

Table B contains the critical values for the theoretical t-test, the maximum test and the Bonferroni adjustment for α levels 0.05, 0.025, 0.01 and 0.005 where n=120. For all α levels 0.05, 0.025, 0.01 and 0.005 for the sample size of 120 the corresponding theoretical t-test values obtained were 1.9801, 2.270117, 2.617776 and 2.860317 respectively. The maximum test provided 2.151836, 2.427165, 2.72672 and 2.920707 respectively for the same α levels. The Bonferroni values obtained were 2.270117, 2.536239, 2.860317 and 3.089022 for the same respective α levels.

Table C contained critical values for the maximum test for sample sizes 8 through 30, 45, 60, 90 and 120 at the 0.05, 0.025, 0.01 and 0.005 alpha levels. Critical values were based 200,000 replicates per n and α level combination for a two tailed test. Critical values presented in Table C moved inversely with table sample sizes meaning that is as sample size increased the table values decreased. However in some instances critical values increased momentarily or "pop up" and then continued to descend in value as n increased and other table values "repeated" at different α levels.

With respect to the "pop up" in values observe the behavior of the values through n= 15 to n=22 at the 0.05 α level. The critical value obtained at n=15 was 1.952256. The value continues to descend at n=16 to 1.925327. Then at n=17 a “pop up” occurred to 2.264775 and then slightly again at n=18 to 2.294804. Then at n= 19 the value descends to 1.971807 a value is higher than the value obtained at n=15 (1.952256) but lower than the values obtained at n 8 through 14. Another instance where behavior of the maximum test is similar occurred between n at 19 through 26 at α 0.025. The critical value obtained at n 19 was 2.191765. The value continues to descend at n=20 (2.189228). Then at n= 21 a “pop up” to 2.373854 occurred. Then again slightly at n=22 to 2.397194. At n= 23 the value descends to 2.163827 to only “pop up” again at n=24 (2.171688), again at n=25 (2.460219), and again at n=26 (2.479246). In two instances critical values “repeated” across different levels of α. The first instance is at n= 29 at both the 0.025 and 0.01 α levels. The value obtained 2.531521. The second instance where this occurred was at n= 45 at the 0.01 and 0.005 α levels. The critical value obtained was 2.732897 for both levels.

The “pop up” and the “repeated” values are attributed to the computational nature of the maximum test. The maximum test reports the highest values of the two or more significance tests computed. The maximum test is based on the assumption that only one test is being computed. The maximum test and naturally the values in Table C are based on the assumption that only one test is being computed. Thus the values are highest or the “maximum” obtained of either the dependent samples t-test or the Wilcoxon sign rank test. Therefore the “pop ups” or the “repeated values” are the result of either the dependent samples t-test or the Wilcoxon sign rank test and not because of any new theoretical or distributional assumption (Blair, 2002; Sawilowsky & Fahoome, 2003). To clarify the nature of the maximum test the definitions of the maximum test provided by Algina, Blair and Coombs (1995) and Kossler (2010) are repeated here. The maximum test is a statistic “for a particular data set, two or more statistics and test the same hypothesis and selecting as the test statistic the one with the smallest p value” and in the event that “each statistic has the same critical value the maximum statistic is simply the most extreme of the calculated statistics” (Algina, Blair & Coombs, 1995, p. 28). Kossler (2010) described the process of the maximum as one whereby a scientist puts “various score statistics together and takes the maximum of them” (p. 2). Also it is important to note that because of the computational nature of the maximum test Type I error is also restrained to nominal alpha.

The present study found that the maximum test as computed provided critical values that were much lower than the values that were obtained through the use of the Bonferroni method. In multiple testing the Bonferroni adjustment is widely recommended and generally applied. The Bonferroni attempts to maintain the experimental-wise error rate at the nominal level by adjusting the point-wise error rate. In other words the Bonferroni attempts to control the probability of rejecting at least one true hypothesis at some specified level a by testing each of the hypotheses of interest at level of significance α and it and it is often used “when conducting multiple tests of significance to set an upper bound on the overall significance level α (Simes, 1986). Simes (1986) explained that “If $T_1, . . ., T_n$" is a set of n statistics with corresponding *p*-values $P_1,.., P_n$, for testing hypotheses $H_1, . . ., H_n$, the classical Bonferroni multiple test procedure is usually performed by rejecting $H_o= \{H_1, ... , H_n\}$ if any p-value is less than α/n. Furthermore the specific hypothesis H_l is rejected for each $P_i \leq S\ \alpha/n$ $(i = 1,\ldots,n)$” (p. 751). However, this is generally not a powerful data analysis method to use when there is a lack of clarity on which of multiple tests to use, because the Bonferroni procedure greatly reduces the statistical power as a by-product of controlling experiment-wise error through reducing alpha.

The critical values as reported in Tables A and B for the maximum test were significantly lower compared to the Bonferroni adjustment. Being that the maximum test is a test of significance and not a correction method based on probability inequality as is the Bonferroni method the maximum test will reject fewer hypotheses than the Bonferroni procedure since the smallest overall significance level at which the individual H_0 would be rejected. The maximum test unlike Bonferroni will not inflate the Type I error rate.

Indeed there exists a debate in the literature regarding whether it is warranted to employ the Bonferroni method of adjustment. At the center of the debate is the accusation that the Bonferroni method is too conservative and lacks requisite power to reject an individual hypothesis as the number of tests increases thereby having the effect of "missing real differences" (Hochberg, 1988, p. 800). Perneger (1998) trenchantly argued that Bonferroni adjustments are unnecessary and at worst deleterious to sound statistical inference (Rothman, 1990; Bland & Altman, 1995). The problem with the method according to Perneger (1998) can be traced all the way back to the misuse of the statistical test theory of Neyman and Pearson (1928), originally intended to aid decisions in repetitive situations so that Type I and Type II would be minimised. The problem is that Type I errors cannot be reduced without inflating Type II errors. Perneger (1998) put forward three reasons why the Bonferroni correction is deleterious to inferential statistics. The first reason is that it provides a correct answer to a largely irrelevant null hypothesis or question and therefore it is concerned with the wrong hypothesis (Siemiatycki et al., 1985; Rothman, 1990; Savitz & Olshan, 1995). The second reason is that inferences made from the method defy common sense. Considerations to the notion that a given comparison will be interpreted differently according to how many other tests were performed are irrelevant (Perneger, 1998). And the third reason is that the Bonferroni adjustments increase Type II errors. Perneger (1998) pointed to the inverse relationship between Type I and Type II errors and noted

Type I errors cannot decrease without inflating Type II errors, which decreasing the Type I error rate which is the whole purpose of the Bonferroni method (Bland & Altman, 1995).

The maximum test resolves the most complicated, confusing, and sometimes disappointing task that researchers face that being choosing a particular statistical test over another. The present study will assist researchers in the process of selecting a statistical test over another particularly when the choice is between the dependent t-test and the Wilcoxon sign rank test. Researchers will now be able to use the maximum test as created in this study in lieu of choice.

An important finding of this study was that the maximum test provided critical values that were smaller than those obtained by the Bonferroni method. This is tremendously significant for the field of inferential statistics. The comparison between the values of the maximum test and the Bonferroni revealed that the Bonferroni method is an unnecessary procedure. The maximum test is easy to calculate and unlike the Bonferroni as alleged by Perneger (1998) does not create more problems. Results obtained through the use of the maximum test need to be adjusted. Cox (1977) referred to the maximum test as the most significant of the separate statistics and accordingly has useful diagnostic properties with extreme component statistics that is sensitive to any statistical departures from normal null hypothesis. This is the major advantage of the maximum test (Blair, 2002). Whereas the Bonferroni according to Perneger (1998) is "concerned with the general null hypothesis (that all null hypotheses are true simultaneously), which is rarely of interest or use to researchers" (p, 2036). Consequently Bonferroni increases the probability of making a Type II error. The maximum test preserves the Type I error to a nominal alpha and by virtue of inverse relations the maximum test does not increase the probability of committing Type II errors. In the final analysis the maximum test as computed in this study not only eliminates the necessity of choosing the t-test over the Wilcoxon signed rank test and vice versa but it also allows a researcher to obtain easily and interpret results more freely without concern for Type I error inflation.

REFERENCES

Algina, J., Blair, R.C., and Coombs, W.T. (1995). A maximum test for scale: Type I error rates and power. *Journal of Educational and Behavioral Statistics*, 20, 27-39.

Algina, J., Olejnik, S. F., and Ocanto, R. (1989). Error rates and power estimates for selected two-sample tests of scale. *Journal of Educational Statistics*, 14, 373- 384.

Arnold, H. J. (1965). Small sample power of the one sample wilcoxon test for non-normal shift alternatives. *The Annals of Mathematical Statistics*, 36, 1767–1778.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of American Statistical Association*, 37, 325-35.

Blair, R.C. (1981). A reaction to consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research*, 51(4), 499-507.

Blair, R.C. (1991). New critical values for the generalized t and generalized rank- sum procedures. *Communications in Statistics*, 20, 981-994.

Blair, C. R. (2002). Combining two nonparametric tests of location. *Journal of Modern Applied Statistical Methods*, 1, 13–18.

Blair, R, C., and Higgins, J.J. (1980). The power of t and wilcoxon statistics: A comparison. *Evaluation Review*, 4, 645–656.

Blair, R. C., and Higgins, J.J. (1980a). A comparison of the t test and the Wilcoxon statistics when samples are drawn from a certain mixed normal distribution. *Evaluation Review, 4*, 645-656.

Blair, R. C., and 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*(4), 309-335

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

Blair, R. C. and 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 and Higgins (unpublished, 1992) as referred to in Sawilowsky, S. S., and Fahoome, G. F. (2003). Statistics through monte carlo simulation with fortran. Oak Park, Michigan: JMASM, Inc.

Bland, J.M., Altman, D.G. (1995). Multiple significance tests: The bonferroni method. *British Medical Journal*, 310, 170.

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

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152.

Brainerd, W. S., Goldberg, C. H., and Adams. (1996). *Programmer's Guide to Fortran 90 (3rd Edition).* New York, NY: Springer- Verlag New York Inc.

Bridge, P. D., and Sawilowsky, S. S. (1999). Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the t-test and Wilcoxon Rank-Sum test in small samples applied research. *Journal of Clinical Epidemiology, 52*, 229-235.

Brown, M. B., and Forsythe, A. B. (1974). Robust tests for the equality of variances.

Journal of the American Statistical Association, 69, 364-367.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Laurence Erlbaum, Hillsdale.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159.

Conover, W. J., and Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician, 35*(3), 124-129.

Conover, W. J., Johnson, M. E., and Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, 23, 351-361.

Cox, D. R. (1977). The role of significance tests. *Scandinavian Journal of Statistics, 4*, 49-70.

Chernoff, H., and Savage, I. R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistic. *Annals of Mathematical Statistics, 29*, 972-994.

Dudoit, S. and van der Laan, M. (2008). *Multiple testing procedures with application to genomics.* New York, NY: Springer Science and Business Media, Inc.

Dyba˚ , T., Kitchenham, B.A., and Jørgensen, M. (2005). Evidence-based software engineering for practitioners. *IEEE Software,* 22 (1), 58–65.

Efron, B. (1969). Student's t-test under symmetry conditions. *Journal of the American Statistical Association*, 64, 1278- 1302.

Feise, R. J. (2002). Do multiple outcome measures require p-value adjustment? *British Medical Centre for Medical Research and Methodology.* 2, 8.

Fleming, T. R. and Harrington, D. P. (1991). *Counting processes and survival analysis*. New York, NY: Wiley.

Freidlin, B., and Gastwirth, J. (2000a). Change point tests designed for the analysis of hiring data arising in employment discrimination cases, *Journal of Business and Economic Statistics, 18* (3), 315-322.

Freidlin, B., and Gastwirth, J. (2000b). On power and efficiency robust linkage tests for affected sibs. *Annals of Human Genetics, 64*, 443-453.

Freidlin, B., Zheng, G., Zhaohai, L., and Gastwirth, J. (2002). Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Human Heredity,* 53(3), 146- 152.

Gerke, T.A., and Randles, H. (2010). A method for resolving ties in asymptotic relative efficiency. *Statistics and Probability Letters*, 80 (13-14), 1065- 1069.

Gibbons, J.D., and Chakraborti, S. (1991). Comparisons of the mann-whitney, students t, and Alternate t tests for means of normal distributions. *Journal of Experimental Education*, 59, 258-267.

Gibbons, J.D., and Chakraborti, S. (2003). *Nonparametric Statistical Inference (4th edition.).* New York: Marcel Dekker.

Good, P. (2005). *Permutations, Parametric, and Bootstrap Tests of Hypotheses (3rd edition).* New York, NY: Springer Science and Business Media, Inc.

Goodman, S.N., and Royall, R. (1988). Evidence and scientific research. *American Journal of Public Health*, 78, 1568-74.

Greenhalgh, T. (1997). Statistics for the non-statistician: Different types of data need different statistical tests. *British Medical Journal*, 315, 364-6.

Heeren, T., and D'Agostino, R. (1987). Robustness of the two independent samples t-test when applied to ordinal scaled data. *Statistical Medicine,* 6, 79-90.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1998). *Applied Statistics for the Behavioral Sciences*, 4th ed., Boston, MA: Houghton Mifflin Company.

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

Holland, B. S., and Copenhaver, M. D. (1987). An improved sequentially rejective bonferroni test procedure. *Biometrics*, 43, 417-423.

Howell, D. C. (2011). *Fundamental Statistics for the Behavioral Sciences*. Belmont, CA: Wadsworth, Cengage Learning.

IMSL. (1980). *International Mathematical and Statistical Libraries*. Houston, Texas.

Isaac, S., and Michael, W. (1997). *Handbook in Research and Evaluation: For Education and the Behavioral Sciences*. San Diego, CA: Educational and Industrial Testing Services.

Jones, D.R., and Rushton, L. (1982). Simultaneous inference in epidemiologic studies. *International Journal of Epidemiology*, 11, 276-82

Keppel, G. (1991). *Design and Analysis: A Researcher's Handbook*. Englewood Cliffs (NJ): Prentice-Hall.

Kossler, W. (2010). Max-type rank tests, U-tests, and adaptive tests for the two-sample location problem - an asymptotic power study. *Journal of Computational Statistics & Data Analysis*, 54 (9), 2053-2065.

Lee, W. J. (1996). Some versatile tests based on the simultaneous use of weighted log rank statistics. *Biometrics, 52*, 721-725.

Metcalf, M. (1990). *Fortran 90 Explained*. New York, NY: Oxford University Press, Inc.

Miller, H. L., and Lower, J. S. (1967). Much ado about the null hypothesis. *Psychological Bulletin, 67*, 188-196.

Morten W., Fagerland, and Sandvik, L. (2009). The wilcoxon-mann-whitney test under scrutiny. *Statistical Medicine, 28*, 1487-1497.

Neuhäuser, M. Büning, H., and Hothorn, L. (2004). Maximum test versus adaptive tests for the two-sample location problem. *Journal of Applied Statistics*, 31 (2), 215-227.

Neyman J., and Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* , 20A, 175-240.

O'Brien, R. G. (1978). Robust techniques for testing heterogeneity of variance effects in factorial designs. *Psychometrika*, 43, 327-342.

O'Brien, R. G. (1979). A general anova method for robust tests of additive models for variances. *Journal of the American Statistical Association*, 74, 877-880.

Oakes, M. (1990). *Statistical Inference*. Boston: Epidemiologic Resources.

Olejnik, S., and Algina, J. (1987). Type I error rates and power estimates of selected parametric and nonparametric tests of scale. *Journal of Educational Statistics*, 12, 45-61.

Opdyke, J. D. (2005). A single, powerful, nonparametric statistic for continuous- data telecommunications parity testing. *Journal of Modern Applied statistical methods*. 4 (2), 372 - 393.

Perneger, T. V. (1998). What's wrong with bonferroni adjustments? *British Medical Journal*, 316(7139), 1236-1238.

Perneger, T. V. (1999). Multiple testing. *British Medical Journal* . 322, 226- 231.

Potvin, C. and Roff, D. (1993). Distribution-free and robust statistical methods: Viable alternative to parametric statistics. *Ecology*, 74(6), 1617-1628.

Ramsey, P. H., and Brailsford, E. A. (1989). Robustness and power of tests of variability on two independent groups. *British Journal of Mathematical and Statistical Psychology*, 43, 113-130.

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

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, 1, 43–46.

Royall, R.M. (1997). *Statistical Inference: A Likelihood Paradigm*. New York: Chapman and Hall.

Ryan, L. M., Freidlin, B., Podgor, M. J., and Gastwirth, J. L. (1999). Efficiency robust tests for survival or ordered categorical data. *Biometrics, 55* (3), 883-886.

Salmaso, L., and Solari, A. (2005). Multiple aspects testing for case-control designs. *Metrika,* 62, 331-340.

Savitz, D.A, and Olshan, A. F. (1995). Multiple comparisons and related issues in the interpretation of epidemiologic data. *American Journal of Epidemiology*, 142, 904–908.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design parametric versus non-parametric statistics in the analysis for randomized trials and non-normally distributed data. *Review of Educational Research Spring*, 60 (1).

Sawilowsky, S. S. (1993). Comments on using alternative to normal theory statistics in social and behavioural science. *Canadian Psychology,* 34, 432-439.

Sawilowsky, S. S. (1999). Nonparametric tests of interaction in experimental designs. *Review of Educational Research*, 60 (91), 91- 126.

Sawilowsky, S. S. (2005). Misconceptions leading to choosing the t test over the wilcoxon-mann- whitney test for shift in location parameter. *Journal of Modern Applied Statistical Method,* November, 4(2), 598-600.

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*(2), 352-360.

Sawilowsky, S., S., Blair, R. C., and Higgins, J. J. (1989). An investigation of the type I error and power properties of the rank transform in factorial ANOVA. *Communications in Statistics*, 14, 25-267.

Sawilowsky, S. S., and Fahoome, G. F. (2003). Statistics through monte carlo simulation with fortran. Oak Park, Michigan: JMASM, Inc.

Seigel, S., Castellan, N. J. (1988). *Non-parametric Statistics for the Behavioral Sciences, 2nd Edition*. New York, NY: McGraw-Hill.

Skovlund, E., and Fenstad, G. U. (2001). Should we always choose a nonparametric test when comparing two apparently non-normal distributions? *Journal of Clinical Epidemiology*, 54, 86-92.

Stonehouse, J. M., and Forrester, G. J. (1998). Robustness of the *t* and *u* tests under combined assumption violations. *Journal of Applied Statistics, 25*(1), 63-74.

Tanizaki, H. (1997). Power comparisons of non-parametric tests: Small- sample properties from monte carlo experiments. *Journal of Applied Statistics*, 24, 603- 632.

Tarone, R. E. (1981). On the distribution of the maximum of the log-rank statistic and the modified wilcoxon statistic. *Biometrics*, 37, 79-85.

Thomas, D.C., Siemiatycki, J., Dewar R., Robins J., Goldberg M., and Armstrong R.G. (1985). The problem of multiple inference in studies designed to generate hypotheses. *American Journal of Epidemiology*, 122, 1080–1095.

Tippett, L. H. C. (1934). *The Methods of Statistics.* London, England: Williams and Norgate.

Tukey, J.W. (1977). Some thoughts on clinical trials, special problems of multiplicity. *Science*, 198, 679-84.

Upton, G., and Cook, I. (2006). *Oxford Dictionary of Statistics*. New York, NY: Oxford University Press Inc.

Weichert, M. and Hothorn, L.A. (2002). Robust hybrid tests for the two-sample location problem. *Communications in Statistics – Simulation and Computation,* 31, 175-187.

Wiederman, W. T., and Alexandrowicz, R. W. (2011). A modified normal scores test for paired data. *European Journal of Research Methods for the Behavioral and Social Sciences*, 7 (1), 25-38.

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

Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypothesis Testing (2nd edition)*. San Diego, CA: Academic Press.

Willan, A. R. (1988). Using the maximum test statistic in the two-period crossover clinical trial. *Biometrics,* 44(1), 211-218.

Yang, S., Hsu, L., and Zhao, L. (2005). Combining asymptotically normal tests: case studies in comparison to two groups. *Journal of Statistical Planning and Inference,* 133(1), 139-158.

Zimmerman, D.W. (1987). Comparative power of student t test and mann whitney u test for Unequal Sample Sizes and Variances. *Journal of Experimental Education*, 55, 171-174.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education*, 67, 55-68.

Zimmerman, D. W., and Zumbo, B. D. (1989). A note on rank transformations and comparative power of the student t-test and wilcoxon- mann- whitney test. *Perceptual and Motor Skills*, 68 (2), 1139- 1146.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education*, 67, 55-68.

Zimmerman, D. W. (2000). Statistical significance level of nonparametric tests biased by heterogeneous variances of treatment groups. *Journal of General Psychology,* 127, 354-364.

Zimmerman, D. W., Williams, R. H., and Zumbo, B. D. (1993). Effect of nonindependence of sample observations on some parametric and nonparametric statistical tests. *Communications in Statistics – Computation & Simulation*, 22, 779–789.

Zimmerman, D. W., & Zumbo, B. D. (1993a). The relative power of parametric and nonparametric statistical methods. In G. Kernen & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues*, (pp. 481–517). New Jersey: Erlbaum.

Zimmerman, D. W., and Zumbo, B. D. (1993b). Relative power of the wilcoxon test, the Friedman test, and the repeated measures ANOVA on ranks. *Journal of Experimental Education*, 62, 75–86.

Zumbo, B. D., and Jennings, M. J. (2002). The robustness of validity and efficiency of the related samples t-test in the presence of outliers. *Psicologica*, 23, 415–450.

ABSTRACT

THE DEPENDENT SAMPLES T AND WILCOXON SIGN RANK MAXIMUM TEST

by

SAVERPIERRE MAGGIO

December 2012

Advisor: Dr. Sholmo Sawilowsky

Major: Theoretical Evaluation and Research (Quantitative)

Degree: Doctor of Philosophy

The task of choosing a statistical test over another can be quite complicated, confusing and in some cases disappointing (Blair, 1985; MacDonald, 1999). The main purposes of the study was to compute the maximum test using the parametric dependent t-test and the non-parametric Wilcoxon sign rank test through the using a FORTRAN program using various subroutines of the International Mathematical and Statistical Libraries (IMSL, 1980) and to obtain critical using sample deviates from a mixed normal distribution. Critical values obtained were at the 0.05, 0.025, 0.01 and 0.005 alpha levels via sample sizes (n) 8 through 30, 45, 60, 90 and 120 for a two- tailed test. Critical values 8 and 120 were for the maximum tests were compared to the Bonferroni correction values. This is tremendously important for the field inferential statistics. Using the maximum test in lieu of choice renders the Bonferroni method unnecessary. The maximum test restrains Type I error to nominal alpha. In the end maximum test not only rids the problem of choice but also allows a researcher to obtain and interpret results more freely without the concern and worry for Type I error inflation.

AUTOBIOGRAPHICAL STATEMENT

Saverpierre Maggio

EDUCATION

2007- 2012	Doctor of Philosophy (Ph.D.) Wayne State University Major: Theoretical Evaluation and Research Cognate: Educational Administration
1993- 1996	Bachelor of Laws (LL.B.) University of Windsor
1992- 1994	Master of Education (M.Ed.) University of Windsor Major: Educational Administration
1990- 1992	Master of Arts (M.A.) University of Windsor Major: Political Science
1986- 1990	Bachelor of Arts (B.A.) University of Windsor Major: Political Science

FACULTY APPOINTMENTS

1993-1998	Adjunct Faculty- Multicultural Education Wayne State University

PROFESSIONAL EXPERIENCE

1998 - PRESENT	Lawyer Sole Practitioner Windsor, Ontario, Canada