Investigating tests for equal variances

by

David W. Nordstokke

B.A., The University of British Columbia, 1995
M.Sc., The University of Northern British Columbia, 2005

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Measurement, Evaluation, and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

March 2009

ABSTRACT

One of the central messages of this dissertation is that (a) unequal variances may be more prevalent than typically recognized in educational and policy research, and (b) when considering tests of equal variances, one needs to be cautious about what is being referred to as "Levene's test" because Levene's test is actually a family of techniques. Depending on which of the Levene tests that are being implemented, and particularly the Levene's test based on means which is found in widely used software like SPSS, one may be using a statistical technique that is as bad (if not worse) than the F test which the Levene test was intended to replace.

The primary goals of this dissertation are to (a) demonstrate that the current statistical practice of testing for equality of variances in hypothesis testing (as prescribed by textbooks and statistical software programs) is insufficient, (b) introduce a new non-parametric statistical test for homogeneity of variances, and (c) investigate the Type I error rate and power of the non-parametric Levene test with that of the median version of the Levene test. Under all conditions investigated, both tests maintained their nominal Type I error rates. As population distributions become more skewed, the non-parametric Levene test becomes more powerful than the median version of the Levene test. These results promise to impact applied statistical practice by informing researchers about the relative efficiencies of the two tests.

This dissertation concludes with remarks about the implications of the findings, and the future work that has arisen from the results.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank all of those who made this dissertation project possible. First of all, I would like to thank Dr. Tim Oberlander for providing me with employment over the course of my PhD work. I would also like to thank the members of the Edgeworth lab at UBC for supporting me and providing a forum for the sharing of ideas.

I would like to thank my committee members, Dr. Anita Hubley and Dr. Kimberly Schonert-Reichl, for their support and feedback throughout the process of my doctoral studies. Their contributions have made this experience rich and productive.

Last, but certainly not least, I would like to thank my supervisor Dr. Bruno Zumbo for his guidance throughout the entire process. Being a student of Dr. Zumbo has been a very pleasant learning experience in a very supportive learning environment. In all, I feel that my growth as a scholar is directly related to my time working with Dr. Zumbo, and I greatly appreciate the time that he has spent mentoring me.

DEDICATION

I would like to dedicate this dissertation to my father, who passed away earlier in 2008. He taught me about the value of hard work and persistence. He also instilled in me the internal fortitude to take the road less travelled, and to pursue my own dreams. His guidance taught me to be responsible for myself and that if I want something, I should make an effort to get it on my own. He taught me to be pragmatic and independent which helped me make this journey. I thank him.

Co-Authorship Statement

Chapter Two has been published. The citation is:
Nordstokke, D. W., & Zumbo B. D. (2007). A Cautionary Tale About Levene's Tests for Equal Variances. *Journal of Educational Research and Policy Studies, 7*, 1-14.
As the first author, I was the in charge of all aspects of this research project including identification of the research questions, literature reviews, syntheses, critiques, and conclusions.

Chapter Three will be revised into a manuscript co-authored with Dr. Bruno D. Zumbo at the University of British Columbia. As the first author, I was in charge of all aspects of the project including formulating research questions, literature review, research design, data analyses. I will also be in charge of the writing of the manuscript. Both co-authors contributed to the identification and design of the research project and will assist in the preparation and revision of this manuscript.

1 INTRODUCTION

1.1 Test of Equal Variances

A common practice when conducting statistical analyses for social, educational, and health research is to test for the equality of variances as a preliminary test for examining mean differences between two or more groups. The assumption of homogeneity (equality) of variances states that the population variance for each group is equal on the dependent variable, and is assessed prior to conducting tests of mean differences. It is important to avoid violating the assumption of equal variances because the results from the statistical test being conducted will be biased. As a result a researcher may be making incorrect decisions in their daily statistical practices, thereby providing misleading information in their area of research thus impeding scientific progress.

There are at least three possible occasions where testing for equality of variances are a concern. The first is when one wants to make inferences about population variances because they are of scientific interest on their own. For example, a health researcher may be interested in studying the effects of a new drug that helps prevent mood swings on some members of a mood management program. The researcher hypothesizes that the drug will decrease the severity of mood swings in patients. In this case, the researcher is interested in the overall increase or decrease in the severity of mood swings (operationalized as the change in the range of mood scores from high extremes to low extremes to more moderate shifts in mood) in which case a test for equal variances would be conducted to test for differences because those in the group that received the program would be hypothesized to have less severity in the range of their scores. The second is

when there is suspected heterogeneity of variances in a t-test or an analysis of variance (ANOVA) in which not all of the factors have fixed effects. For example, a researcher is interested in spatial ability and uses a categorical variable, such as gender, as a grouping variable in a t-test. It cannot be assumed that males and females vary equally on spatial ability so, prior to the t-test, a test for equal variances must be carried out. A third occasion when one might suspect heterogeneity of variances is in a fixed-effects analysis of variance in which the numbers of observations in the groups are widely disparate (Glass, 1966). When there is reasonable evidence suggesting that the variances of two or more groups are unequal, a preliminary test of equal variances is conducted prior to conducting the t-test or ANOVA.

This dissertation will focus on several tests of equal variances. Commonly, tests of equal variances are used as a preliminary test for checking the assumption of equal variances when using a t-test or ANOVA. When planning to conduct a t-test, the result of the test for homogeneity of variances makes the choice of t-test conditional upon its result. If the conditional test is being biased by some external factor, then distorted error levels (i.e., elevated Type I error rates) could lead to improper inferences being made because of the violation of the assumptions of the t-test and, as a result, incorrect decisions may be made regarding the selection of statistical test. For use as a conditional test, the importance of testing for equality of variances lies in the choice of which version of the t-test to select, the pooled variances t-test (Student's test) or the Welch's version of the t-test. The Welch's version of the t-test adjusts the degrees of freedom to account for unequal variances, unequal sample sizes, and also small sample sizes. Its calculation of the standard error does not include pooling the variances to estimate a common

population variance. See Gravetter and Wallnau (2005) for a detailed description of the calculation of each test.

To provide a more concrete example, a researcher is interested in determining whether or not there is a statistically significant difference between boys and girls when it comes to empathic concern. The researcher must decide whether to use a Student's t-test, which uses a pooled variance in the estimation of the standard error, or use the Welch's t-test, which uses adjusted degrees of freedom and is usually recommended when variances and/or sample sizes are unequal. At this point a test for equal variances is carried out. Unknown to the researcher, there was a television show on that is extremely popular amongst a number of the girls that had a storyline that revolved around empathy, which resulted in a number of them scoring much higher than the rest of their classmates. None of the boys watched this program. When the scores of the test were calculated and a test for homogeneity of variances was conducted, it was statistically significant. Based on this the researcher opted to use the Welch's version of the t-test, which tends to be more conservative hence reducing the chance of detecting an actual difference if one is present, but does protect against the inflation of Type I errors when variances are unequal. Whether boys and girls differ in empathic concern is irrelevant for the purpose of this dissertation, however, this does illustrate the steps that a researcher must take. The decision of whether to use the Student's t-test or the Welch's t-test is based on the test for equal variances and if that test is not performing adequately, then it may lead to the improper choice of t-test, which may lead to the selection of a less powerful test and may result in researchers missing actual differences between groups.

When testing for equal variances between groups, a problem arises when samples are collected from populations that result in skewed data. Data can become skewed because there are extreme scores in one end of the distribution resulting in an asymmetrically shaped distribution. In fact, it can be argued that, in many cases, data commonly collected in educational, behavioral and health research do not meet the assumption of normality or symmetry (Bradley, 1977; Micceri, 1989).

Typically much of data analysis is guided by information provided from textbooks and the actual statistical tests that are provided by statistical software companies who produce the statistical packages that are used daily by researchers and widely taught to students in undergraduate and graduate programs. In the past 15 years, there has been a great improvement in the interactive properties of statistical software. Many statistical tests can be requested and implemented with several clicks of a button via drop-down windows. This makes statistical analyses more accessible to greater numbers of researchers than in the past and, who for the most part, are not trained as statisticians. Instead, they are researchers who purchase statistical software to analyze their data. This should put a great deal of pressure on the authors of textbooks and the producers of statistical packages to provide the most up-to-date and valid information to ensure that the statistical tests that are recommended in statistical texts and offered in statistical software packages are the most appropriate test for the application required.

It has been shown that as data becomes more skewed, the Type I error rates of the F test for equal variances become elevated and/or power is reduced, resulting in invalid inferences (Box, 1953). The Levene test for equal variances is recommended in textbooks and used in commonly used statistical software packages (e.g., SPSS).

Part of the main focus of this dissertation will be on three versions of the Levene's test for equal variances. This includes, the mean version that is supplied as a default test in SPSS, the median version, and, particularly, a newly developed rank-based version of the Levene test for equal variances. There have been numerous attempts to create tests for equal variances for use with skewed data. For a review of most of the tests for equal variances that have been formulated, readers should consult Conover, Johnson, and Johnson (1984). Conover et al. conducted a simulation study that investigated the Type I error rates and average statistical power of a wide number of tests for equal variances.

Conover et al. (1984) found that, for skewed distributions, the median version of the Levene test for equal variances was superior to all of the tests investigated for Type I error rates and statistical power. One limitation of this work, however, is that to summarize the vast number of results, the average power values were given. Average power (as with any average) masks point by point power values of a statistical test when analyzing its power across several conditions. Average power is useful for summarizing and comparing the bias of a number of tests across a wide variety of conditions and it is useful when mathematicians compare a large number of tests at asymptotic levels, but the applied researcher is generally interested in how a test performs at a certain point under a specific set of conditions, so they require information on a point to point basis to help them assess a test's performance. For instance, Figure 1.1, on page 23, illustrates why looking at the point by point values for power is important and average power may not be representing power results in a manner that is useful to researchers who are often interested in the power of a test at certain points along the power continuum. The x-axis

represents the effect sizes that are being investigated that range from zero to infinity. The y-axis represents the power of the statistical test under investigation. The point where the x value is on the line would represent approximately the average power of the test.

Providing average power results are troubling because there are points on the line where the power of the test is quite poor, perhaps misleading researchers into believing that a test performs better than it actually does. It is problematic for researchers to practically use the results of the average power of a statistical test because the information provided by average power is not easily accessible for everyday researchers. For this reason, this dissertation will be providing point by point power values and not average power to describe the results of the simulations that were conducted, as is the standard for reporting in many simulation studies (e.g. Blair & Higgins, 1980; Zimmerman & Zumbo, 1993). In addition, statistical power and Type I errors, which are typically identified as a probability (e.g., .05) will be described interchangeably as percentages (e.g. 5.0%).

Based on the Conover et al. (1984) results, which showed that, comparatively, the Levene median test performs well concerning its Type I error rate and statistical power, this dissertation will investigate four tests for equal variances, the mean, median, and nonparametric versions of the Levene test, as well as the traditional F test. This will be achieved by assessing their performance when the assumption of normality has been violated due to skewed population distributions when sample sizes are unequal across three sample sizes, and three ratios of sample sizes. The mean version represents what is being used in current statistical practice (if you use, for example, SPSS), the median version represents the gold standard and needs to be investigated under a wider range of

conditions than have been investigated to date, and the rank version represents a new approach to the problem of testing for equality of variances on skewed data.

The impact of skewed population distributions on the comparative performance of tests for equal variances (Levene mean, F-test, Levene median, and the non-parametric (rank) Levene) will be investigated by using Monte Carlo simulations to test the performance of the three versions of the Levene's test for equality of variances under question to more fully understand their utility and implications for their use in day to day research. Simulation provides a method for investigating the utility of a statistical test under a wide variety of test conditions. For example, a researcher is interested in the impact of unequal group sizes on the performance of the Student's t-test. The parameters of the population distribution would be specified and simulations would be conducted under various conditions where the ratio of numbers in each group changes between conditions. The researcher then tracks the performance of the test under each of the varying conditions. This highlights the usefulness of simulation as a method of statistical inquiry providing empirical evidence of how a statistical test behaves that can be applied to daily research practice.

Another theme present in this dissertation is what is occurring in daily research practice and how this affects inferences made from the results of statistical tests. As will become evident in this dissertation, in the context of the test for equal variances, many popular statistical software packages (e.g., SPSS) provide an algorithm that is inappropriate in many circumstances. Also, the publishers of many statistical textbooks give erroneous recommendations for testing for equal variances. Together, this has laid

the foundation for a widespread misuse and misrepresentation of tests for equality of variances.

The format of this dissertation follows the guidelines of a "manuscript-based dissertation" of the Faculty of Graduate Studies at the University of British Columbia. This dissertation is centrally comprised of two manuscripts for publication (Chapters 2 and 3 of the dissertation). The first manuscript (Chapter 2) in this dissertation investigates the performance of the commonly used version of Levene's test for equal variances and has been published in a peer reviewed journal. The findings of this manuscript motivated paper two of this dissertation. Particularly, Chapter 2 will focus on what is referred to as Levene's test for equal variances and how this test has been misrepresented in software packages, textbooks, and in the literature. The second manuscript (Chapter 3) introduces and investigates the performance of a new statistical approach for testing for equality of variances. The final chapter (Chapter 4) will provide an overall discussion of the implications of this research and further directions for future work.

Before turning to the two manuscripts, however, I will provide a review of the relevant research literature and background information to facilitate a reading of the two manuscripts. The intended audience for this dissertation is not necessarily the mathematical statistician who can read and understand the mathematical literature. For this reason, mathematical notation is used sparingly throughout this dissertation. The target audience is educational, health, and social science researchers who may not have strong mathematical backgrounds. For this reason, there is a need to introduce and explain material that will give everyday researchers not fluent in mathematics the ability

to read and understand the subsequent chapters of this dissertation; however a basic understanding of statistical concepts will facilitate the reading of this dissertation.

1.2 Background Literature

The purpose of the introductory chapter is to educate readers on the importance of assumptions when we are comparing two or more groups, and provide the appropriate information that is required to read the subsequent chapters of this dissertation. Each of the primary assumptions of parametric statistical methods related to the t-test and homogeneity of variances tests will be briefly reviewed. In addition, a description of Type I errors, statistical power, rank transformations, and simulation methods will be included and followed by a declaration of the primary objectives of this dissertation.

1.2.1 Assumptions Related To Statistical Tests

Central to the use of statistics are the assumptions that are required in order to utilize these methods in an appropriate manner. The ideology of assumptions is embedded in statistical thought and practice. Fisher (1973) notes that, when referring to the prior knowledge of a variable being measured, "Most frequently, however, and especially when the probabilities of contrasted scientific theories are in question, a candid examination of the data at the disposal of the scientist shows that nothing of the kind can be claimed" (p.17). This is central to the necessity of assumptions due to the fact that, in many cases, prior knowledge of probabilities related to the population of interest are unavailable, especially when there are multiple, possibly contradictory, theories related to the existence of a phenomenon, and assumptions must be made to compensate for their unavailability. This highlights the necessity of assumptions in statistical testing because, in many cases, we do not have a full understanding of the variables being investigated; in

fact, if that knowledge was available, there would probably not be any need to study the phenomena. Based on this logic, it is clear that assumptions are a necessity for statistical inference; this highlights the importance of meeting assumptions when conducting a statistical analysis because the validity of the results depends on the strength of the assumptions.

Central to this dissertation are the assumptions that are related to the t-test and tests for equal variances because, in daily practice, the t-test is widely used to compare differences in group means, and a test for equal variances is often used to accompany the t-test. The Students t-test and the mean version of the Levene test are considered a parametric statistical technique; parametric statistics refer to a statistical approach that is outlined by specific statistical assumptions that, if not met, can lead to the inflation of Type I error rates and reduced power, thus leading to inappropriate inferences based on the results of the statistical technique that is employed. For a brief discussion related to the use of parametric and non-parametric statistics, see Gravetter and Wallnau (2005). For the purpose of this dissertation, three major assumptions will be discussed: (1) independent observations, (2) the underlying population is normally distributed, and (3) equality of variances. It should be noted that two of these assumptions are requirements of tests for equal variances: (1) independent observations and (2) that they are normally distributed across the groups.

Independence is achieved when observations for any individual's score is not dependent on any other individual's score (i.e., conditional probability). This assumption is related to the sampling procedure used to build probability statements about variables. Gaito (1959) reminds us that, "The errors must be independent, i.e., the selection of any

one case must not affect the chances of any other case for inclusion in the sample" (p. 115). This means that, if there are conditional probabilities present, the probabilities that are calculated are not true representations of the phenomenon in question and inferences based on this information will likely be aberrant, thus the necessity for the assumption of independence. Violation of this assumption will distort the estimate of the standard error of the mean differences (making it too small) and hence affecting the Type I error rate.

The assumption of normally distributed population distributions is important for parametric statistics (e.g., t-test, test of equal variances based on the mean) and is based on a normal theory approach for conceptualizing variables that are being studied. The normal (Gaussian) distribution approach is justified through the central limit theorem which basically states, in nonmathematical terms, that as the size of a sample of scores increases, the sampling distribution about the mean approaches the normal distribution (Reber & Reber, 2001). When distributions are non-normal (e.g., highly skewed, multimodal, or heavily-tailed), the ability to identify a viable probability distribution using the normal theory approach is reduced. This is because, under normal theory statistics, the probability distributions are symmetric and more or less bell-shaped and, when distributions are non-normal, especially in the ways listed above, the probability distribution is distorted which effects the estimation of means and variances leading to erroneous statistical results.

There has been much debate in the literature over the years as to whether the assumption of a Gaussian distribution is a true requirement of parametric methods. The discussion revolves around the issue of the robustness of parametric methods, which can be generally defined as the tendency of a statistic or statistical procedure to be relatively

unaffected by the presence of a small number of unusual or incorrect data values (Upton & Cook, 2004). Zumbo and Jennings (2002) remind us that researchers generally refer to two types of robustness: (1) robustness of validity which, is said to occur if the accuracy of the statement made from a statistical test is invariant to the violations of the assumptions which, is interpreted as a test is valid if it maintains its nominal Type I error rates, and (2) robustness of efficiency which, refers to the power of a statistical procedure to find significant differences when underlying assumptions are violated. This suggests that, when considering the performance of a statistical test, one must not only be aware of the Type I error rate and power of a test, but also the validity of the statements that are made based on the results of a statistical test. In this dissertation, robustness will be defined by the robustness of validity and statistical efficiency (i.e., Type I error rates and statistical power) of each test while varying several population parameters that incrementally change across a wide variety of conditions.

Early robustness studies conducted have concluded that parametric tests are generally robust to departures from normality (e.g., Boneau, 1960, Glass et al., 1972), which allowed researchers to continue to use parametric methods without much concern. In fact, much of the departure from rank based methods can be attributed to the inordinate attention paid to Glass et al. in relation to the demonstrated robustness of the fixed effects analysis of variance (ANOVA) and covariance (ANCOVA) (Blair, 1981). This early work focused on the Type I error rates of the statistical tests that they were investigating. Bradley (1978) points out in response to the early robustness literature that "The actual behavior of the probability of a Type I error under the assumption violation is quite complex, depending upon a wide variety of interacting factors. Yet allegations of

robustness tend to ignore its highly particularistic nature and neglect to mention important qualifying conditions" (p. 144). This indicates that the early robustness studies did not take all of the factors into consideration. The issue of robustness is much too important to dismiss as being of theoretic interest, but not of any practical use, and caution should be exercised when making general statements about the robustness of parametric techniques.

The above position was supported by Micceri (1989) who studied the distributional properties of 440 large-sample achievement and psychometric measures and found that none of the tests were normally distributed. Moreover, he pointed out that many of the distributions had not been examined in the literature. Based on these claims, Sawilowsky and Blair (1992) investigated the properties of eight real distribution shapes that were identified by Micceri as representative of those found in educational and psychological research. They found that the *t* test was reasonably robust to Type I errors when (a) sample sizes were equal, (b) sample sizes were fairly large (25 to 30), and (c) tests were two-tailed rather than one-tailed. This finding supports the notion that parametric techniques are generally robust to non-normality; however, the robustness of the *t* test is dependant on the above conditions (a, b, and c), and, in many cases, non-parametric tests are more powerful.

The assumption of equality of variances is based on the premise that the population variances on the variable being analyzed for each group are equal. Formally, let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$, where $m + n = N$, be two continuous distribution functions. $F(x)$ and $G(y) = F(\theta_y)$ where $\theta > 0$ is a scale parameter. To assume the equality of variances translates to, in essence, assuming that $\theta = 1$. There are, at least, two

situations in which one cannot assume equality of variances: (a) when the groups of participants (i.e., subjects or experimental units) are formed by domain differences such as age groups, gender, or educational level, and/or, (b) when the participants (knowingly or unknowingly to the researcher) differ on some important, possibly unmeasured variable (Zumbo & Coulombe, 1997). This suggests that one cannot necessarily assume that the participants are homogeneous or exchangeable and so there is no basis to assume equality of variances when testing the null hypothesis of no difference between two or more groups.

Even though equal sample sizes is not an explicit assumption of the t-test or ANOVA, it is important to note that, when sample sizes and variances are unequal, it leads to Type I error rates becoming conservative or liberal in comparison to the nominal alpha. When larger sample sizes are associated with larger variances, the Type I error rates tend to become very conservative and when the smaller sample size is associated with the larger variance, Type I error rates become elevated (Tomarken & Serlin, 1986). This is important because it highlights the importance of investigating bias under a wide variety of conditions.

1.2.2 Type I Errors and Statistical Power

The performance of a statistical test can be determined through the investigation of its Type I error rate and its statistical power. When conducting a simple hypothesis test for investigating differences between two groups, there are two competing statements that are used to test for differences. The first is the *null hypothesis* (H0) that states that there is no difference between the group means on the dependent variable. The second is

alternative hypothesis (H1) that states that there is a difference between the group means on the dependent variable.

In hypothesis testing, there are two ways to come to a correct decision and two ways to come to an incorrect decision. These possibilities are listed as illustrated in Table 1.1, which is a 2X2 table that represents all of the possible outcomes that could occur during a hypothesis test. The probability of correctly retaining H0 when it is true is denoted as 1-α, this is represented in the above diagram in the upper left box in the decision quadrant. In the box in the upper right of the decision quadrant is β, the probability of making a Type II error (not rejecting H0 when it is actually false). For the purpose of this dissertation, these first two possibilities are not of particular interest. The two choices of interest are in the lower two cells of the quadrant, when H0 is rejected. When H0 is rejected when it is in fact true, a Type I error has occurred (represented in the table as α). If H0 is rejected when it is actually false, it is represented as 1-β, and this is statistical *power*. The Type I error rate is generally selected by the researcher and represents the probability that the statistical difference is due to chance. This is most often set nominally by the researcher at .05 or .01 (Hayes, 1988).

The goal of mathematicians and statisticians is to develop tests that maintain their Type I error rates under a wide variety of conditions. That is, the probability of incorrectly rejecting the null hypothesis when it is true is ideally held at a maximal 5 percent (when using a nominal alpha of .05). In addition, when tests maintain their nominal alpha (i.e., Type I error rate) it is important that the test possess sufficient statistical power. For example, under some conditions, a test for equal variances may maintain its Type I error rate; however, it may not be powerful enough to detect true

differences often enough. As mentioned earlier, Type I error rates and power are often represented as percentages. For example, if the nominal alpha is .05, that means that 5 percent of the time the test will reject the null hypothesis when it should not be rejected; and the power of the test may be .20, meaning that 20 percent of the time the test will be powerful enough to detect real differences between groups.

1.2.3 Rank Transformations

Often data do not meet the assumption of normality and data analysts are required to transform data into a form that is appropriate for analysis. This is the basis for many of the non-parametric tests that are available, and in this dissertation the terms "non-parametric" and "rank" will be used interchangeably. The rank transformation is useful for removing extreme skew from a distribution. The rank transformation takes a column of data and simply ranks the scores with the lowest score getting a rank of 1 (the lowest rank), and the highest score getting the highest rank. The rank transformation re-scales the scores from data that is interval or ratio into its relative placement in the distribution resulting in a uniform distribution with ordinal level information.

An essential feature of rank based techniques is that they are very robust to outliers (Potvin & Roff, 1993; Zumbo & Coulombe, 1997). Consider a normal distribution. If one or several scores are added to the extreme end of the scale, the distribution will be greatly affected with an increase in its skew. When a rank transformation is applied, the distance between the scale points essentially disappears and the distribution becomes uniform and the influence of the extreme scores are reduced. This is the basic principle of rank transformations. When the data are extremely non-normal, perhaps caused by several outliers or intervening variables, the transformation

removes their effect on the distribution. As outlined by Conover (1999), rank tests are "valid for all types of populations, whether continuous, discrete, or mixtures of the two" (p. 270). This demonstrates the flexibility of these techniques across many situations. Some of these tests include the Mann-Whitney-U and Kruskal-Wallis tests on ranks.

Another possibility that has been proposed is to perform parametric procedures on the ranked data. As stated by Conover and Iman (1981), "Many of the more useful and powerful non-parametric procedures may be presented in a unified manner by treating them as rank transformation procedures. Rank transformation procedures are ones in which the usual parametric procedure applied to ranks of the data instead of the data themselves" (p.124). This demonstrates the possibilities that could be applied using these techniques and how they may be utilized given the large number of distributional possibilities encountered by researchers.

The use of parametric statistics on ordinal data was considered by Stevens (1946) and others as out of bounds because ordinal data does not carry enough information (i.e., equal intervals between scale points) and are, therefore, unsuitable because the interpretation would hold no meaning. Zumbo and Zimmerman (1993) conducted a series of computer simulations where they performed a two-sample Student *t*-test on data whose underlying structure was ordinal. Results showed that parametric techniques can be used when the data are ordinal, at least under the conditions they investigated. Their results show that the statistical properties of the t-test hold when data are ordinal and the results are interpretable.

The idea that parametric statistics can be computed on ordinal data allows for tests of equal variances to be tested for using rank transformed data; however it should be

noted that the rank transformation does not provide a solution to extreme skew when the design is more complex than a 2X2 design (Sawilowsky, 1990). Sawilowsky states that, "The ANOVA is a flexible procedure that can be used in a variety of sophisticated designs, such as to test complex interaction effects. Historically, however, it has been maintained there are no satisfactory non-parametric tests for interaction" (p. 101). In today's research world where modeling complex interactions is a necessity, non-parametric techniques are lacking. This limits the complexity of the models that can be specified using non-parametric tests.

In the simple case, where testing for simple differences in variances is concerned, the use of the rank transformation appears to be acceptable. For the purpose of testing for equality of variances, using the rank transformation is appropriate because it retains information related to the variance of the sample. For example, imagine two groups with two subjects in each group. The two scores from the first group are 1 and 30, and the two scores from the second group are 10 and 11. The relative ranking of the first group are 1 and 4, and the relative ranking of the second group is 2 and 3. This demonstrates that the variances of the two groups remain unequal; however, as noted by Zimmerman (1996), not nearly as dissimilar as the original scores. Thus ranking the original scores and entering them into the analysis is an acceptable practice for testing simple hypotheses.

1.2.4 Simulation Methods in Research

Much of statistical theory is based upon analytic mathematics, which helps predict the performance of statistical tests under various conditions. However, much of analytic mathematics is dependent on certain conditions being satisfied. When they are not, the results from mathematics provide no information about the performance of a statistical

method (Mooney, 1997). Simulations are often used because they are persuasive for empirical researchers as a rhetorical device. As you will see in the first manuscript of this dissertation, simulation is used to inform about a mathematical result. Simulations are also of use when mathematic results are possible, yet the results are not disseminated into practice because of the inaccessibility of results to non-mathematically trained researchers. Simulation studies allow one to communicate the results to other audiences than mathematicians and statisticians. In this context, the conditions refer to the assumptions of the statistical test. If one or more of the assumptions are violated, then the conditions are sub-optimal for analytic mathematics. Thus, because this dissertation focuses on violating assumptions, a simulation study was selected as the method of choice. Also, asymptotic mathematical results are not usually readily useful to everyday researchers.

Simulation studies provide a method for testing the bias of a statistical test when one or several of its assumptions are violated. Central to the process of simulation is the sampling distribution, from which random numbers are generated in order to test the performance of the statistical test under investigation. The sampling distribution is essentially a possible range of numbers and the probabilities of those numbers occurring.

Knowledge of the probabilities associated with the occurrence of the numbers allows statistical researchers to empirically investigate a wide variety of hypothetical distributions under varying conditions. For example, a statistical researcher wanted to conduct a simulation study to investigate the effects of kurtosis on the Type I error rate of the Student's t-test. The researcher would repeatedly draw random samples of numbers from known populations of simulated data that have incrementally varying levels of

kurtosis, run the t-test on each of the simulated random samples, and count the frequency of Type I errors. Then the researcher would compare the number of Type I error rates across the varying levels of kurtosis. The researcher would then know the impact of kurtosis on the Type I error rate of Student's t-test under those conditions.

It should be noted that, even though simulation is a powerful tool for empirically investigating the bias of a statistical test, it does have limitations. Simulations are conducted on theoretical sampling distributions that do not necessarily occur in real data. When a sampling distribution is selected and some parameter, like kurtosis, is varied, the change in that distribution is predictable due to the mathematic theorems related to that hypothetical distribution. In the social, behavioral, and health sciences, the factors that may cause the distribution to become more kurtotic are many and not based on an algorithm allowing them to be accurately predicted. Often researchers are unaware of outside influences on their data, which is difficult to simulate. With that in mind, a keen researcher must be cautious when generalizing simulation results to real world problems, being aware that their theoretical distributions are only an approximation of real world conditions. However, if hypothetical distributions are near to or mirror some real world situation, then the results of simulations are very useful and appropriate. The key is for statistical researchers to build upon the knowledge that is gained through simulation work to improve everyday research practice and modify simulation studies based on information about the sampling distribution that can be gained through theory that is provided in the everyday research setting. Together, simulation and practice generated theory provide a framework for investigating the performance of statistical methods and theory building.

It should be noted that in this dissertation I will be referring to various versions of the Levene test. To clarify, I will be referring to the original Levene test as the mean version of the Levene test and the SPSS version of the Levene test. In addition, reference will be made to the median version of the Levene test which is sometimes referred to in the literature as the Brown/Forsythe test. Finally, in chapter 3, I will introduce a new non-parametric Levene test that will be referred to interchangeably as the non-parametric Levene or the rank based Levene test.

1.3 Dissertation Objectives

There are several objectives that will be met throughout the course of this dissertation. The first objective is to investigate the performance of the mean version of the Levene test across various conditions. The purpose of this objective is to demonstrate what is happening in current statistical practice because, as mentioned previously, the mean version of the Levene test (described in detail in Chapter 2) is the default test provided by SPSS for testing equality of variances when conducting a t-test using that software. With that in mind, the first paper in the dissertation will focus on the mean version of the Levene test and its statistical power and Type I error rates when data are sampled from population distributions that range from being normally distributed to distributions that are extremely skewed. This will demonstrate what can occur when the normality assumption for the mean version of the Levene test is violated to different degrees.

A second objective of this dissertation is to investigate the statistical properties of the median version of the Levene test (described in detail in Chapter 3). This test has become somewhat of a gold standard (Brown & Forsythe, 1974; Conover et al., 1984)

when testing for equal variances when population distributions are non-normal. However, to this point, there have only been a few studies that have investigated its statistical power and Type I error rates; therefore a wider spectrum of conditions will be simulated to increase understanding of the statistical properties of the median version of the Levene test.

A third objective is to introduce and investigate the statistical properties of a new rank based Levene test for equal variances (described in detail in Chapter 3) under distributional properties that range from normality to extreme skew. This test has the potential to be the most efficient statistical test (i.e., maintaining Type I error rates with sufficient power under a wide range of conditions) for equality of variances of all the tests investigated in this dissertation. In all, this dissertation aims to expose insufficiencies in current statistical practice, and then provide a practical solution for data analysts to employ in their daily research activities.

Figure 1.1 Demonstration of the utility of using point by point power values

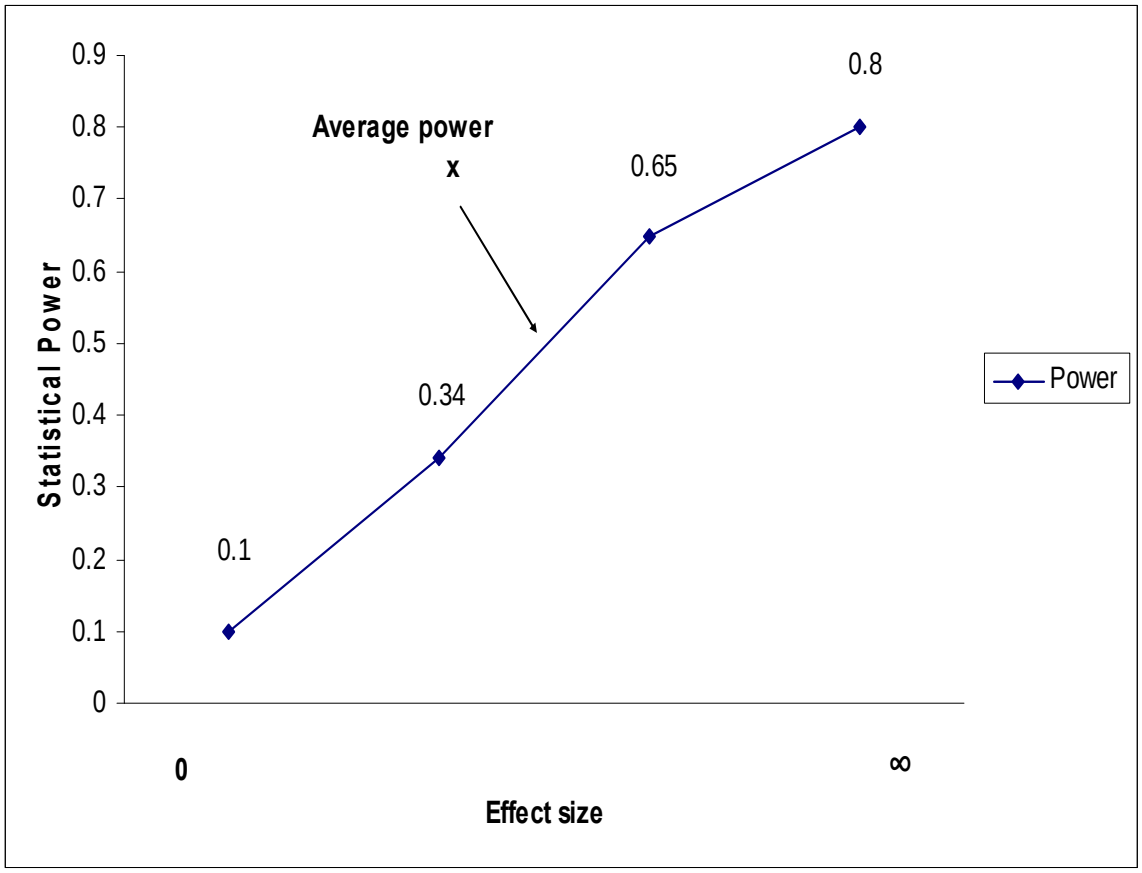


Table 1.1 Possible outcomes of a hypothesis test

		Actual Case	
		H0 is true	H0 is false
Researcher Decision	Retain H0	1-α	β (Type II error)
	Reject H0	α (Type I error)	1-β (power)

1.4 References

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. & Higgins, J.J. (1980). A comparison of the power of the Wilcoxon's rank-sum statistic to that of the student's t statistic under various non-normal distributions. *Journal of Educational Statistics*, *5*(*4*), 309-335.

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

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

Box, G. E. P. (1953). Non-normality and tests on variance. *Biometrika, 40*, 318–335.

Bradley, J.V. (1977). A common situation conductive to bizarre distribution shapes. *The American Statistician*, *31*(*4*), 147-150.

Brown, M.B. & Forsythe, A.B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*(2), 364-367.

Conover,W.J. (1999). *Practical Non-parametric Statistics: 3rd Edition*. Toronto: John Wiley & Sons, Inc.

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

Conover, W.J., Johnson, M.E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, *23*(*4*), 351- 361.

Fisher, R.A. (1973). *Statistical Methods and Scientific Inference 3rd Edition*. Toronto: Collier-Macmillan.

Gaito, J. (1959). Non-parametric methods in psychological research. *Psychological Reports*, *5*, 115-125.

Glass, G.V. (1966). Testing homogeneity of variances. *American Education Research Journal*, *3*(*3*), 187-190.

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

Gravetter, F.J. & Wallnau, L.B. (2005). *Essentials of Statistics for the Behavioral Sciences 5th Edition*. Toronto: Thomson/Wadsworth.

Hayes, W.L. (1988). *Statistics 4th Edition*. Toronto: Holt, Rinehart and Winston, Inc.

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

Mooney, C.Z. (1997). *Monte Carlo Simulation: Quantitative Applications in the Social Sciences*. Thousand Oaks, CA: Sage Publications Inc.

Potivn, C. & Roff, D.A. (1993). Distribution-free and robust statistical methods: Viable alternatives to parametric statistics?. *Ecology*, *74*(*6*), 1615-1628.

Reber, A. & Reber, E. (2001). *The Penguin Dictionary of Psychology 3rd Edition.* Toronto: Penguin Press.

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

Sawilowsky, S.S. & Blair, R.C. (1992). A more realistic look at the robustness of and

Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin*, *111*(*2*), 352-360.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, *103*, 849-856.

Tomarken, A.J. & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin, 99*(*1*), 90-99.

Upton, G. & Cook, I. (2004). *Oxford Dictionary of Statistics*. Toronto: Oxford University Press.

Zimmerman, D.W. (1996). A note on homogeneity of variance of scores and ranks. *Journal of Experimental Education*, *64*(*4*), 351-362.

Zimmerman, D. W., & Zumbo, B.D. (1993). Rank transformations and the power of the Student t-test and Welch's t-test for non-normal populations with unequal variances. Canadian Journal of Experimental Psychology, 47, 523-539.

Zumbo, B. D., & Coulombe, D. (1997). Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. Canadian Journal of Experimental Psychology, 51, 139-150.

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

Zumbo, B.D. & Zimmerman, D.W. (1993). Is the selection of statistical methods governed by level of measurement?. *Canadian Psychology*, *34*(*4*), 390-400.

2 FIRST MANUSCRIPT CHAPTER

2.1 A Cautionary Tale about Levene's Tests for Equal Variances[1]

2.1.1 Introduction

When comparing groups in educational, social, behavioral, and policy research, a common tacit, yet essential, statistical assumption is that the variances of the dependent variable for each group are equal. This assumption is referred to as 'homogeneity of variances' when using statistics like the t-test or analysis of variance to compare group means. For example, as is widely seen in educational and policy research, one may use the independent samples t-test to compare boys and girls in terms of their average mathematics achievement test scores, and hence one is, sometimes unknowingly, assuming that the boys and girls have equal mathematics score variances.

The matter of unrecognized, or ignored, statistical assumptions and their impact on research practice are exaggerated during what one of the founding editors of this journal, Professor Sean Mulvenon, aptly describes as our "Era of Point-and-Click Statistics" wherein easy to use statistical software often masks and hides the complex statistical assumptions and realities of day-to-day research practice in educational, social, behavioral, and policy studies. This matter of meeting complex statistical questions and procedures with deceptively easy to use statistical software, and in turn its impact on research practice, is a theme that runs throughout this paper.

Zumbo and Coulombe (1997, p. 147) remind us that there are, at least, two situations in which one cannot assume equality of variances: (a) when the groups of participants (i.e., subjects or experimental units) are formed by domain differences such

[1] A version of this chapter has been published. Nordstokke, D. W., & Zumbo B. D. (2007). A cautionary tale about Levene's tests for equal variances. *Journal of Educational Research and Policy Studies, 7*, 1-14.

as age groups, gender, or educational level, and/or, (b) when the participants (knowingly or unknowingly to the researcher) differ on some important, possibly unmeasured variable. In either situation, one cannot necessarily assume that the participants are homogeneous or exchangeable and so there is no basis to assume equality of variances when testing the null hypothesis of no difference between means or medians – nonparametric tests are also susceptible to issues of unequal variances when testing for equal medians (Harwell, Rubinstein, Hayes, & Olds, 1992; Zimmerman & Zumbo, 1993a; 1993b). It can be easily argued that either of these situations occurs commonly in educational, behavioral, social, and policy research. One then cannot assume equal variances and hence needs to regularly test for equality of variances before testing for equal means (or medians).

2.1.2 Review of statistical notation important for testing for equal variances

Common understanding, as documented in statistical and methodological research papers, textbooks, and codified in widely used statistical software, is that the F test for equality of variances is problematic in terms of its inflated Type I error rate with non-normal population data. As a reminder, the hypothesis for the F test of variances is

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 \\ H_1 &: \sigma_1^2 \neq \sigma_2^2 \end{aligned}. \qquad \text{(H1)}$$

The test statistic to test H_0 against H_1 is

$$F = \frac{s_1^2}{s_2^2}. \qquad \text{(T1)}$$

When the H_0 in (H1) is true, the sampling distribution of $F(\upsilon_1, \upsilon_2)$ from (T1) is the F-family of distributions with $\upsilon_1 = n_1 - 1$ and $\upsilon_2 = n_2 - 1$ degrees of freedom, and the

sample variances and sample sizes are s_1^2, s_2^2, n_1 and n_2, respectively. The reader should see Glass & Hopkins (1984, p. 263) for a detailed description. It has been known for over half a century, however, that the test of (H1) by (T1) is notoriously sensitive to, and largely invalidated by, non-normally distributed (population) dependent variable scores (Box, 1953).

Building on the work of Box, Scheffe, and others, Levene (1960) introduced a methodological approach that was meant to resolve Box's concern for the F-test being so sensitive to population non-normality when investigating equality of variance. In short, Levene's approach involves using the usual F-test for equality of means computed on what we will refer to as intermediary scores, which one defines as the absolute deviations of the data points from an estimate of the center of the group – i.e., a one-way ANOVA of the centered original data. Levene's original proposal was to compute these intermediary (centered) scores by centering at the sample mean. In short, the original Levene's test involves one conducting a one-way, j-group, ANOVA of the transformed original data, $\left|X_{ij} - \overline{X}_j\right|$, for each i individual in the j groups, where $\overline{X}_j$ denotes the mean of the j^{th} group; and, for our purposes, Levene's original test will be denoted as

$$\text{ANOVA}(\left|X_{ij} - \overline{X}_j\right|). \qquad \text{(T2)}$$

The original Levene's test, (T2), was initially found to be quite robust to departures from normality (Levene, 1960). It was this initial finding that drew attention to (T2) as a useful alternative to the F-test (T1). It has, however, been shown using computer simulation that violations of normality increases the Type I error rate of the Levene test (T2) (e.g., Shoemaker, 2003; Zimmerman, 2004). Carroll and Schneider

(1985) showed mathematically that Levene's test involving means, (T2), maintains its nominal Type I error rate only for symmetric distributions – distributions that are non-normal but yet still symmetric obviously fall within this category; for example, the uniform distribution. They also described a modified Levene's test (Brown & Forsythe, 1974a; 1974b) incorporating the sample median, rather than the mean,

$$\text{ANOVA}(\left|X_{ij} - Mdn_j\right|), \qquad \text{(T3)}$$

where Mdn_j denotes the sample median for the j^{th} group and the remaining notation is the same as above. They went on to show that (T3) maintains its Type I error rate for asymmetric distributions. That is, Carroll and Schneider show that, asymptotically, Levene's approach has the correct Type I error rate whenever the estimate of group 'center' is an estimate of group median, (T3). They went on to show that this explains why published Monte-Carlo studies have found that Levene's original proposal of centering at the sample mean, (T2), has the correct Type I error rate only for symmetric distributions, while centering at the sample median has correct Type I error rate both for symmetric and for asymmetric distributions (Brown and Forsythe, 1974). Interestingly, it was this median-based approach, (T3), and not the mean-based approach, (T2), that was found to be the most robust and useful of 56 possible tests for homogeneity in extensive simulations done by Conover, Johnson, and Johnson (1981).

What becomes evident from the simulations and mathematical work is that one needs to be precise about which Levene test is being used, (T2) or (T3). In fact, Levene introduced a strategy for data analysis, by centering and then applying the ANOVA, so there really is no <u>one</u> Levene test, per se, but instead an approach or strategy to the problem. Curiously, research papers and textbooks, as well as the codified methods in

widely used statistical software, such as SPSS, continue to use the original Levene's test, (T2), without even mentioning that alternatives have been developed, or warning the data analyst that (T2) may be problematic. In many textbooks and software documentation, it is stated that the (unspecified) "Levene test" is robust to non-normality and should be used instead of the notorious F-test from (T1).

2.1.3 Review of information found in textbooks

To take our discussion a step further, textbooks going back 20 years, including recently published introductory statistics and research methodology textbooks for the social and educational sciences, were consulted to obtain information regarding the assumption of equal variances, for two independent groups, and how to test that assumption for one's data. Nearly all of the textbooks recommended using what they refer to as Levene's test for equality of variances and most suggested the use of SPSS (e.g., Cohen & Lea, 2004; Cramer, 1996; Tabachnick & Fidell, 2007; Vaughan, 1998). What is even more troubling is that one widely used and influential textbook suggested that, if the sample sizes are equal then the assumption of equal variances can be disregarded (Hays, 1988), and yet another, Ferguson and Takane (1989), suggested to conduct the F test of (T1), without reference to the over half a century old finding by Box. In fact, as Keyes and Levy (1997) note, the Levene's test involving means, (T2), is available in many widely used statistical software packages such as BMDP, MINITAB, and SPSS and, in some cases (e.g., SPSS t-test), it is the only test made available to the software user.

To provide a concrete example of the analytic results noted above, we conducted a simulation study of the Type I error rate of the Levene test, (T2), provided by software

packages like SPSS. In addition, we also included the F-test, (T1), to show its comparative performance to (T2) – i.e., how does the Levene's test compare to using the notoriously bad F-test? This comparison of (T1) to (T2) is somewhat novel and really meant to be a pointed contrast of the much-advocated use of Levene's test, by which is typically meant (T2). Therefore, the purpose of the simulation is to document the Type I error rates (and, if appropriate, the statistical power) of Levene's test, (T2), with an eye toward cautioning researchers who implement tests of equal variances using Levene's test, by which is meant (T2), in their day-to-day research activities. In fact, much to our chagrin, in what Mulvenon describes as our era of point-and-click statistics, (T2) is embodied in day-to-day research activities by default in statistical software packages.

It should be noted that Carroll and Schneider's (1985) results make a simulation study, per se, unnecessary for the mathematically (and statistically) inclined who can decode those findings and incorporate them into their research practice. However, as we show above, given that 20 years after its publication Carroll and Schneider's results evidently have yet to enter the consciousness of textbook writers and statistical software designers in the social and behavioral sciences, perhaps a more persuasive argument is required. In what Mulvenon refers to as our era of point-and-click statistics (and, hence, in the hands of the so-called point-and-click educational and social science 'statistician'), this simulation study was, in its essence, intended to be a persuasive demonstration of why we should tend to the warnings in Carroll and Schneider (1985) and others in the statistical and methodological literature, and a reminder that when one hears reference to the "Levene test" one should naturally then ask: which one?

2.2 Methods

2.2.1 Data Generation

Given our study purposes, a computer simulation was performed using SPSS software. Throughout the remainder of this paper, we will use the term "SPSS Levene's test" as shorthand for the original Levene test in (T2). Following standard simulation methodology (e.g., Zimmerman, 1987; 2004), population distributions were generated using a pseudo random number sampling method to produce χ^2 distributions. The design of the simulation study was a 4 x 3 x 3 x 9 completely crossed design with: (a) four levels of skew of the population distribution, (b) three levels of sample size, (c) three levels of sample size ratio, n_1/n_2, and (d) nine levels of ratios of variances. The dependent variables in the simulation design are the Type I error rates (when the variances are equal), and power under the eight conditions of unequal variances. Of course, we will only investigate statistical power in those conditions wherein the nominal Type I error rate (in our study .05) is maintained.

2.2.2 Shape of the population distribution

I investigated four levels of skew: 0, 1, 2, and 3 and used the family of χ^2 distributions to simulate the population data. As is well known, as the degrees of freedom of a χ^2 distribution increases, it more closely approximates a normal distribution[2]. The skew of the distributions for both groups were always in the same direction in all replications and are shown in Figure 2.1 (reading from top left to bottom right) for skew values of 0, 1, 2, and 3 respectively.

[2] It should be noted that the population skew was determined empirically for large sample sizes of 100,000 simulees with 10000, 7.4, 2.2, and 0.83 degrees of freedom resulting in skew values of 0.03, 1.03, 1.92, and 3.06, respectively.

2.2.3 Sample Sizes

Three different sample sizes, $N = n_1 + n_2$, were investigated: 24, 48, and 96. Three levels of ratio of group sizes (n_1 / n_2 : 1/1, 2/1, and 3/1) were also investigated.

2.2.4 Population variance ratios

Nine levels of variance ratios were investigated (σ_1^2 / σ_2^2 : 5/1, 4/1, 3/1, 2/1, 1/1, 1/2, 1/3, 1/4, 1/5). The design was created so that there were direct pairing and inverse pairing in relation to unbalanced groups and direction of variance imbalance. Direct pairing occurs when the larger sample sizes are paired with the larger variance and inverse pairing occurs when the smaller sample size is paired with the larger variance (Tomarken & Serlin, 1986). This was done to investigate a more complete range of data possibilities. In addition, Keyes and Levy (1997) drew our attention to concern with unequal sample sizes, particularly in the case of factorial designs – see also O'Brien (1978, 1979) for discussion of Levene's test in additive models for variances. Findings suggest that the validity and efficiency of a statistical test is somewhat dependent on the direction of the pairing of sample sizes with the ratio of variance (Tomarken & Serlin, 1986).

As a whole, the complex multivariate variable space represented by our simulation design captures many of the possibilities found in day-to-day research practice.

2.2.5 Determining Type I Error Rates & Power

The frequency of Type I errors was tabulated for each cell in the design. In all there were 324 cells in the simulation design. As a description of our methodology, the following will describe the procedure for completing the steps for one cell in the design.

First, two similarly distributed populations are produced; for this example, it is two normally distributed populations that are sampled to create two groups. In this case, each group has 12 members, and the population variances of the two groups are equal. An independent samples t-test using SPSS is then performed on the two groups; a Levene's test for equality of variances, by which we mean (T2), is reported in this procedure as a default test to determine if the variances are statistically significantly different at the nominal alpha value of .05 (±.01). Again, note that we intend to mimic day-to-day research practice. This procedure was replicated 5000 times for each cell in the design.

In the cells that maintained their Type I error rates, statistical power is represented by the percentage of times that the Levene's test, (T2), correctly rejected the null hypothesis.

2.3 Results

Type I error rates for the Levene's mean test is presented in Table 2.1. Table 2.1 has four columns: (i) total sample size, N, (ii) ratio of sample sizes, n_1/n_2, (iii) Type I error rate of SPSS's Levene test, (T2), and (iv) the Type I error rate for the F-test, (T1). Within the table there are the four levels of skew of the population distribution. As an example, the Type I error rate of SPSS's Levene test for a skew of 0, and a total sample size of 24 (with 12 per group) is 6.0%.

For symmetric distributions (i.e., skew of zero), the Type I error rates, for both the SPSS Levene's test and the F-test, were near the nominal alpha level of .05. Furthermore, for these symmetric distributions, the SPSS Levene's test and the F-test were not influenced by either total sample sizes or unequal group sizes.

When the distribution had a skew of 1, 2, or 3, (i.e., the non-normal distributions), the Type I error rate of both the SPSS Levene's test and the F-test were inflated above the nominal level of .05. In fact, one finds that the skew and sample size inequalities together lead to even further Type I error rate inflation. Although both are quite inflated above their nominal Type I error rates, SPSS Levene's test appears to be less effected by unequal group sizes.

The statistical power results of the SPSS Levene's test and F-test under zero skew (symmetric distribution) conditions are presented in Table 2.2. Note that power was only reported for those cells in the simulation design for which the nominal Type I error rate was protected. Table 2.2 is structured so that the first column lists the two statistical tests, either SPSS Levene's test or the F-test. Furthermore, columns two and three list the total sizes and the ratio of sample sizes, respectively. The ratio of samples sizes, n_1 / n_2: 1/1, 2/1, and 3/1, are also paired with the ratio of population variances, σ_1^2 / σ_2^2, resulting in 1/2, 1/3, 1/4, 1/5 being inversely paired, and 5/1, 4/1, 3/1, 2/1 are directly paired. Therefore, as an example, in the case of a total sample size of 24, with 16 in group one and 8 in group two (i.e., a 2/1 sample size ratio), the statistical power of the SPSS Levene's test is 62.0% and the F-test 67.5% in the variance ratio of one to five (group one to group 2, hence an inverse pairing).

It is evident from Table 2.2, that when comparing the SPSS Levene's test to the corresponding F-test, in 66 of the possible 72 such comparisons in Table 2.2, the F-test is more powerful than the SPSS Levene's test. And, in fact, the F-test is more powerful than the corresponding SPSS Levene's test for all cases of direct pairings – i.e., when the

larger sample size comes from a population with the larger variance. The power superiority of the F-test for normal distributions is expected from mathematical statistics – i.e., the F-test is most powerful for the normal population distribution.

2.4 Discussion

Several points are important to take away from this study. When speaking of "Levene's test" it is important to be precise about which of the family of possible Levene tests one is referring. Furthermore, most elementary textbooks in the social and behavioral sciences, as well as some of the widely used statistical software packages, are referring to the original test proposed by Levene in 1960, (T2), based on means.

The widely discussed Levene's test, (T2), has been shown to be sensitive to non-normality of the population distribution (e.g., Carroll & Schneider, 1985). Not surprisingly our simulation study also showed the same inflation in Type I error rates as Carroll and Schneider discussed. However, it should be noted that the inflation is not minimal and depends, to some degree, on degree of skew and sample size ratio. In fact, what is interesting is that the Levene's test, (T2), actually performs on par, in terms of invalidity, with the notorious F-test of (T1), known widely since 1950 to be problematic. However, although both are quite inflated above their nominal Type I error rates, Levene's test, (T2), appears to be less affected by unequal group sizes – hardly a consolation when the Type I error rates tend to be between two to four times the nominal alpha!

In those situations wherein the Levene's test (and hence the F-test) maintain their nominal alpha, the statistical power findings show that, in most situations, the F-test is

more powerful than the Levene's test, (T2), as is expected from results in mathematical statistics.

Given the current state of knowledge, following Brown and Forsythe (1974) and Conover, Johnson, and Johnson (1981), we would recommend that day-to-day researchers use the median-based Levene's test, (T3); however, this median-based Levene test is not currently available in the widely used software packages such as SPSS. To provide an alternative, an easily implemented new statistical technique we have developed entitled the 'nonparametric Levene test' is, however, showing very promising results in terms of maintaining its nominal Type I error rate and having substantial statistical power in all the conditions studied in this current paper. This nonparametric Levene test uses Conover and Iman's (1981) notion of the rank transformation as a bridge between parametric and nonparametric statistics and simply involves (i) pooling the data and replacing the original scores by their ranks and then (ii) separating the data back into their groups, and (iii) applying the mean-based Levene test (T2) to the ranks. This can be easily accomplished using widely available software such as SPSS. A forthcoming paper will describe this test in detail and provide further information about its statistical performance.

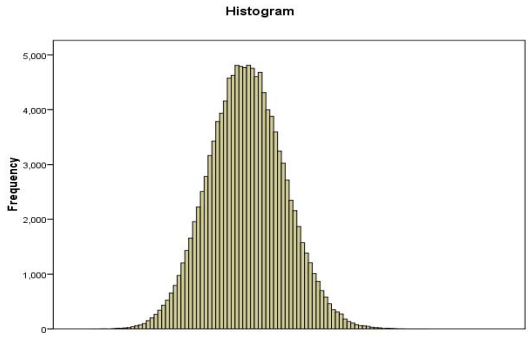
This paper, therefore, is a cautionary tale about Levene's tests for homogeneity of variances. If one is using the original variation of Levene's test, a mean-based test, (T2), such as that found in SPSS, one may be doing as poorly (or worse) than the notorious F test of equal variances. We hope that Carroll and Schneider's caution about the mean version of the Levene test will soon become as widely recognized, and adopted in

textbooks and statistical software, as was Box's tale in 1953 about the F-test, and that the median-based Levene's test will be more widely used.

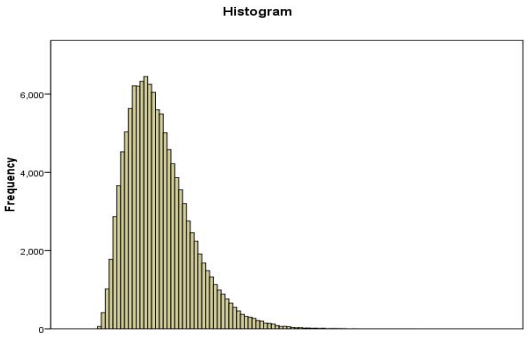
In closing, then, one needs to keep in mind Ted Micceri's (1989) observation that, in real data situations, the normal curve appears nearly as often as the mythical unicorn (Micceri, 1989). Therefore, to George Box's (1953) well-known quip in his influential paper on tests of variances that the preliminary test on variances is rather like putting to sea in a row boat to find out whether conditions are sufficiently calm for an ocean liner to leave port, we would add that using the mean-based Levene's, (T2), is akin to sending out a dinghy instead.

Figure 2.1 Shape of population distributions used in simulations

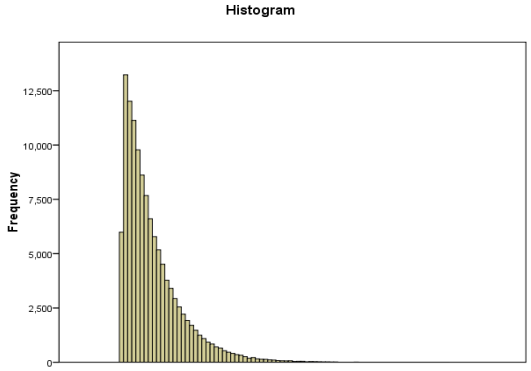
Skew = 0



Skew = 1



Skew = 2



Skew = 3

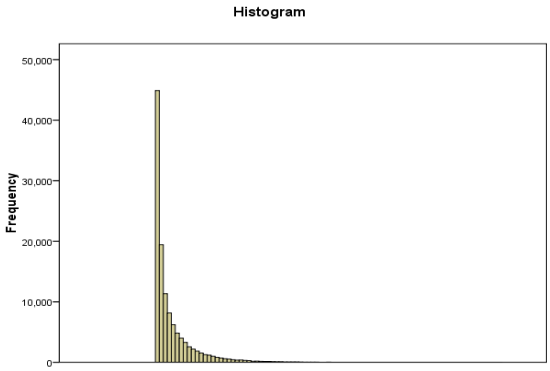


Table 2.1 Empirical Type I Error Rates for the SPSS Levene's and the F Tests

N	n1/n2	SPSS's Levene's Test	F test
Skew = 0			
24	1/1	6.0	5.1
24	2/1	5.9	5.6
24	3/1	5.8	5.2
48	1/1	5.3	5.4
48	2/1	5.6	4.9
48	3/1	5.5	4.9
96	1/1	4.8	4.6
96	2/1	5.1	4.8
96	3/1	4.7	4.9
Skew = 1			
24	1/1	8.1	8.5
24	2/1	8.0	8.1
24	3/1	8.3	8.5
48	1/1	8.0	8.7
48	2/1	7.7	9.6
48	3/1	8.5	9.1
96	1/1	8.3	10.8
96	2/1	8.2	10.2
96	3/1	7.1	10.0
Skew = 2			
24	1/1	14.4	16.4
24	2/1	13.5	16.2
24	3/1	13.7	15.3
48	1/1	14.6	17.5
48	2/1	13.0	17.3
48	3/1	13.0	18.9
96	1/1	12.8	18.8
96	2/1	13.3	18.4
96	3/1	12.8	20.1
Skew =3			
24	1/1	22.8	24.4
24	2/1	23.4	27.7
24	3/1	19.7	28.0
48	1/1	21.0	24.8
48	2/1	20.4	27.8
48	3/1	19.2	29.9
96	1/1	20.3	27.5
96	2/1	20.2	29.2
96	3/1	19.5	29.9

Table 2.2 Statistical Power for SPSS Levene's Test and the F-test

			Population Variance Ratio, σ_1^2 / σ_2^2							
Test	N	n1/n2	**1/5**	**1/4**	**1/3**	**1/2**	**2/1**	**3/1**	**4/1**	**5/1**
			Inverse Pairings				**Direct Pairings**			
Levene	24	1/1	60.6	50.7	35.7	18.2	18.2	35.7	50.7	60.6
F	24	1/1	81.0	73.1	54.4	29.3	29.3	54.4	73.1	81.0
Levene	24	2/1	62.0	50.8	36.1	17.7	13.8	26.0	40.1	49.7
F	24	2/1	67.5	54.6	38.5	17.3	35.5	58.2	75.6	84.0
Levene	24	3/1	57.0	45.7	33.3	16.4	11.7	19.5	28.4	36.6
F	24	3/1	49.1	37.0	23.0	9.0	37.9	57.5	73.4	83.7
Levene	48	1/1	92.4	84.3	64.9	31.3	31.3	64.9	84.3	92.4
F	48	1/1	98.6	95.0	81.2	48.5	48.5	81.2	95.0	98.6
Levene	48	2/1	90.3	80.3	63.1	31.9	27.1	55.3	76.4	87.7
F	48	2/1	95.1	88.2	72.2	37.7	51.7	82.8	94.7	98.3
Levene	48	3/1	85.2	75.0	57.3	29.4	21.4	45.7	66.2	78
F	48	3/1	88.0	78.8	60.2	27.8	51.0	80.2	93.2	97.6
Levene	96	1/1	99.9	98.9	92.4	58.6	58.6	92.4	98.9	99.9
F	96	1/1	100.0	99.8	97.8	76.4	76.4	97.8	99.8	100.0
Levene	96	2/1	99.5	98.0	89.4	56.9	51.0	89.0	98.4	99.8
F	96	2/1	99.8	99.4	94.8	68.3	74.6	97.8	99.9	100.0
Levene	96	3/1	98.7	96.1	85.9	48.4	41.0	81.9	96.1	99.1
F	96	3/1	99.4	98.2	90.7	56.2	70.6	97.0	99.8	100.0

2.5 References

Box, G. E. P. (1953). Non-normality and tests on variance. *Biometrika, 40*, 318–335.

Brown, M. B. & Forsythe, A. B. (1974a). The small sample behavior of some statistics which test for the equality of several means. *Technometrics*, *16*, 129-132.

Brown, M.B. & Forsythe, A.B. (1974b). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*(2), 364-367.

Carroll, R.J. & Schneider, H. (1985). A note on Levene's tests for equality of variances. *Statistics and Probability Letters*, *3*, 191-194.

Cohen, B.H. & Lea, R.B. (2004). *Essentials of Statistics for the Social and Behavioral Sciences*. Hoboken: John Wiley & Sons.

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

Conover, W.J., Johnson, M.E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, *23*(*4*), 351- 361.

Cramer, D. (1996). *Basic Statistics for Social Research*. New York: Routeledge.

Ferguson, G.A. & Takane, Y. (1989). *Statistical Analysis in Psychology and Education 6th Edition*. Toronto: McGraw-Hill Book Company.

Glass, G. V., and Hopkins, B.K. (1984). *Statistical Methods in Education and Psychology (2nd ed.)*, New York: Prentice-Hall.

Harwell, M.R., Rubinstein. E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics, 17,* 315-339.

Hayes, W.L. (1988). *Statistics 4th Edition*. Toronto: Holt, Rinehart and Winston, Inc.

Keyes, T. M., & Levy, M. S. (1997). Analysis of Levene's test under design imbalance. *Journal of Educational and Behavioral Statistics, 22,* 227-236.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin et al. (Eds.), *Contributions to probability and statistics: Essay in honor of Harold Hotelling* (pp. 278-292). Stanford, CA: Stanford University Press.

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

O'Brien, R. G. (1978). Robust Techniques for Testing Heterogeneity of Variance Effects in Factorial Designs. *Psychometrika, 43*, 327-344.

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.

Shoemaker, L. H. (2003). Fixing the *F* Test for Equal Variances. *American Statistician*, *57*(*2*), 105-114.

Tabachnick, B. G., & Fidell, L. S. (2007). *Experimental Designs Using ANOVA.* Belmont, CA: Thomson, Brooks-Cole.

Tomarken, A.J. & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin, 99*(*1*), 90-99.

Vaughan, E.D. (1998). *Statistics: Tools for Understanding Data in the Behavioral Sciences*. Upper Saddle River: Prentice-Hall.

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. (2004). A note on preliminary test of equality of variances. *British Journal of Mathematical and Statistical Society*, *57*, 173-181.

Zimmerman, D. W., & Zumbo, B.D. (1993a). Rank transformations and the power of the Student t-test and Welch's t-test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology, 47,* 523-539.

Zimmerman, D. W., & Zumbo, B.D. (1993b). The relative power of parametric and nonparametric statistical methods. In G. Keren and C. Lewis (Eds.) *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues* (pp. 481-517). Hillsdale, NJ: Erlbaum.

Zumbo, B. D., & Coulombe, D. (1997). Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. *Canadian Journal of Experimental Psychology, 51*, 139-150.

3 SECOND MANUSCRIPT CHAPTER

3.1 A New Non-parametric Levene Test for Equal Variances [3]

3.1.1 Introduction

In current statistical practice, there is no consensus of whether or not there is any one statistical test that has robustness of validity and statistical efficiency when data are sampled from skewed population distributions. In hypothesis testing, the test for equality of variances is often used as a preliminary test to assess the degree of similarity between two samples.

The widely used hypothesis for the test of equal variances, when there are two groups, is

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 \\ H_1 &: \sigma_1^2 \neq \sigma_2^2 \end{aligned}, \qquad \text{(H1)}$$

wherein, a two-tailed test of the null hypothesis (H_0) that the variances are equal is tested against the alternative hypothesis (H_1) that the variances are not equal. The matter of testing this hypothesis can be approached in numerous ways. The variance is a representation of the spread of the data; the purpose of this hypothesis test is to essentially make sure that each group has the equal amount of spread in their distribution of scores. Nordstokke and Zumbo (2007) investigated the parametric mean based Levene test for testing equal variances. As a reminder, the mean version of the Levene test is

$$\text{ANOVA}\left(\left|X_{ij} - \overline{X}_j\right|\right), \qquad \text{(T1)}$$

wherein, a one-way analysis of variance is conducted on the absolute deviation value, which is calculated by subtracting (jth) group mean from each individual's (i) score. As a

[3] A version of this chapter will be submitted for publication. Nordstokke, D. W., & Zumbo, B. D. A new non-parametric test for equal variances.

general rule (i) denotes the individual score and (j) denotes group membership, and will expressed as such for the remainder of this paper.

The primary goal of this study is to compare the Type I error rates and the statistical power of the median version of the Levene test and the nonparametric Levene test that was briefly introduced by Nordstokke and Zumbo (2007). Nordstokke and Zumbo remind readers that the mean version of the Levene test for equality of variances does not maintain its nominal Type I error rate when the underlying population distribution is skewed and, in so doing, introduced the nonparametric version of the Levene test that is intended to be more robust under the conditions where samples are collected from population distributions that are skewed.

This problem has been approached from many perspectives. As reviewed by Conover and his colleagues (1984), the top performing test for equality of variances was the median based Levene test. The median based version of the Levene test for equal variances is

$$\text{ANOVA}\left(\left|X_{ij} - Mdn_j\right|\right), \tag{T2}$$

wherein, the analysis of variance is conducted on the absolute deviations of each (i) score from the median (Mdn_j) for each of the (j) groups. This test was shown to perform well in situations wherein data were skewed by maintaining the nominal Type I error and power. Browne and Forsythe (1974), who are widely recognized as the developers of the median based version of the Levene test, also demonstrated that this test was suitable for use with skewed distributions under the distributions that were investigated. They investigated several distributional forms in their study: a Gaussian (normal) distribution, a Student's t distribution with 4 degrees of freedom (symmetric), and a Chi-square

distribution with 4 degrees of freedom (skew equals approximately 1.5). In the skewed distribution used in their study, the Type I error rates were maintained in all of the conditions and had power values above .80 when the effect size was large (4/1), the ratio of sample sizes was 1/1, and the sample size was 40. This result suggests that the median version of the test could potentially become the widely used standard if these results hold across a broader ranger of conditions.

Therefore, another purpose of this paper is to investigate the performance of the Levene median test for equal variances under a wider range of conditions than studied by Browne and Forsythe (1974). Conover, Johnson, and Johnson (1981) also showed that the Levene median test maintains its Type I error rates under an asymmetric double exponential distribution, but had average power values of .10. It is important that the Levene median test be studied further to assess its usefulness across varying research situations. For these reasons, the Levene median test for equal variances is used as a comparison for the proposed non-parametric Levene test.

As Nordstokke and Zumbo (2007) describe it, the non-parametric Levene test involves pooling the data from both groups, in the two group situation, ranking the scores, placing the rank values back into their original groups, and running the Levene test on the ranks.

$$\text{ANOVA}\left(\left|R_{ij} - \overline{X}_{j}\right|\right), \tag{T3}$$

wherein R_{ij} is calculated by pooling the values from each of the (j) groups and ranking the scores. An analysis of variance is conducted on the absolute value of the mean of the ranks for each group ($\overline{X}_{j}$) subtracted from each individual's rank (R_{ij}). This is based on the principle of the rank transformation. When the data are extremely non-normal,

perhaps caused by several outliers or some other intervening variables, the transformation changes the distribution and makes it uniform. Conover and Iman (1981) suggested conducting parametric analyses, for example, the analysis of variance, on rank transformed data. Thus the non-parametric Levene test is a parametric analysis of variance conducted on rank transformed data.

It should be noted that the null hypothesis for both the median and non-parametric Levene test is not the same as for the mean version of the Levene test. The null hypothesis of these two tests is that the population distributions are identically distributed, and the alternate hypothesis is that they are not identically distributed. If two or more distributions are identically distributed, then it is implied that the variances are equal. Thus the overlap between the hypotheses of parametric and non-parametric tests allows for interchangeability between them when testing for equal variances because implicit in the assumption of equal variances is identical distributions. This overlap allows one to test for equal variances using the non-parametric hypothesis of identical distributions.

Rank transformations are appropriate for testing for equal variances because, if the rankings between the two groups are widely disparate, it will be reflected by a significant result. For example, if the ranks of one of the groups tend to have values whose ranks are clustered near the top and bottom of the distribution and the other group has values whose ranks cluster near the middle of the distribution, the result of the non-parametric Levene test would lead one to conclude that the variances are not homogeneous.

3.1.2 Steps for calculating the Levene Median and the non-parametric Levene

The comparison between the median Levene (T2) and nonparametric Levene (T3) tests will be based on the Type I error rates and power of both tests under a wide range of sampling conditions, ranging from nearly normally distributions to skewed distributions. It is important for the test of equal variances to maintain its Type I error rates while maintaining adequate power to detect a difference. To provide a concrete hypothetical example, consider 24 scaled math scores between two groups with the sample size balanced equally between the groups (12/12). Table 3.1 illustrates the steps of the Levene median test (T2).

In this example, there is a 5/1 ratio of variances between the groups, with data that has a skew value of 3.0. (See the next section on how data were generated) These data were simulated, so the effect size is known to be large and H0 should be rejected. The first column in Table 3.1 is the grouping variable for the analysis, the number in the second column is the scaled math score, the third column lists the median values for each group, and the fourth column is the absolute value of each score subtracted from the median. There is clearly a difference between variances; however, the median version of the Levene test failed to reject the null hypothesis, $F(1, 22) = 1.57$, $p = .23$.

The non-parametric Levene (T3) test was conducted on the same data, as illustrated in Table 3.2. The first column in Table 3.2 is the grouping variable for the analysis, the number in the second column is the scaled math score, the third column is the rank of the pooled scores, the fourth column is the mean rank for each group, and the fifth column is the absolute value of the individual rank minus the mean rank. Unlike the

median version, the non-parametric version of the Levene test rejected the null hypothesis, $F(1, 22) = 11.01$, $p = .003$.

This example clearly illustrates that there is much potential for the non-parametric Levene test to be used in daily research practice, and that more conditions should be investigated to determine the usefulness of both the median and non-parametric versions of the Levene test for equal variances. The remainder of this paper will describe the methods for the simulations and the results comparing the Type I error and power of both the non-parametric and the median versions of the Levene test for equal variances.

3.2 Methods

3.2.1 Data Generation

A computer simulation was performed using SPSS software. Following standard simulation methodology (e.g., Nordstokke & Zumbo, 2007; Zimmerman, 1987; 2004), population distributions were generated using a pseudo random number sampling method to produce χ^2 distributions. Building from Nordstokke and Zumbo (2007), the design of the simulation study was a 4 x 3 x 3 x 9 completely crossed design with: (a) four levels of skew of the population distribution, (b) three levels of sample size, (c) three levels of sample size ratio, n_1/n_2, and (d) nine levels of ratios of variances. The dependent variables in the simulation design are the Type I error rates (when the variances are equal), and power under the eight conditions of unequal variances. Staying consistent with Nordstokke and Zumbo (2007), we will only investigate statistical power in those conditions wherein the nominal Type I error rate (in our study .05) is maintained.

3.2.2 Shape of the population distributions[4]

I investigated four levels of skew: 0, 1, 2, and 3 and used the family of χ^2 distributions to simulate the population data. As is well known, as the degrees of freedom of a χ^2 distribution increases it more closely approximates a normal distribution[4]. The skew of the distributions for both groups were always in the same direction in all replications and are shown in Figure 3.1 (reading from top left to bottom right) for skew values of 0, 1, 2, and 3 respectively.

3.2.3 Sample Sizes

Three different sample sizes, $N = n_1 + n_2$, were investigated: 24, 48, and 96. Three levels of ratio of group sizes (n_1 / n_2 : 1/1, 2/1, and 3/1) were also investigated.

3.2.4 Population variance ratios

Nine levels of variance ratios were investigated (σ_1^2 / σ_2^2 : 5/1, 4/1, 3/1, 2/1, 1/1, 1/2, 1/3, 1/4, 1/5). The design was created so that there were direct pairing and inverse pairing in relation to unbalanced groups and direction of variance imbalance. Direct pairing occurs when the larger sample sizes are paired with the larger variance and inverse pairing occurs when the smaller sample size is paired with the larger variance (Tomarken & Serlin, 1986). This was done to investigate a more complete range of data possibilities. In addition, Keyes and Levy (1997) drew our attention to concern with unequal sample sizes, particularly in the case of factorial designs – see also O'Brien

[4] It should be noted that the population skew was determined empirically for large sample sizes of 120,000 simulees with 10000, 7.4, 2.2, and .83 degrees of freedom resulting in skew values of 0.03, 1.03, 1.92, and 3.06, respectively; because the degrees of freedom are not whole numbers, the distributions are approximations.

(1978, 1979) for discussion of Levene's test in additive models for variances. Findings suggest that the validity and efficiency of a statistical test is somewhat dependent on the direction of the pairing of sample sizes with the ratio of variance.

As a whole, the complex multivariate variable space represented by our simulation design captures many of the possibilities found in day-to-day research practice.

3.2.5 Determining Type I Error Rates & Power

The frequency of Type I errors was tabulated for each cell in the design. In all, there were 324 cells in the simulation design. As a description of our methodology, the following will describe the procedure for (T2) and (T3) for completing the steps for one cell in the design. First, for both tests, two similarly distributed populations were produced; for this example, it was two normally distributed populations that were sampled to create two groups. In this case each group had 12 members, and the population variances of the two groups are equal. So this example tests the Type I errors for the two tests under the current conditions. For (T2), the absolute deviation from the median is calculated for each value in the sampled distribution and an ANOVA is performed on these values to test if the variances are significantly different at the nominal alpha value of .05 (±.01). For (T3), values are pooled and ranked, then partitioned back into their respective groups. An independent samples t-test using SPSS is then performed on the ranked data of the two groups. A Levene's test for equality of variances, by which we mean (T3), is reported in this procedure as a default test to determine if the variances are statistically significantly different at the nominal alpha value of .05. Again, note that

we intend to mimic day-to-day research practice, hence the number of cells under varying conditions. This procedure was replicated 5000 times for each cell in the design.

In the cells where the ratio of variances was not equal and that maintained their Type I error rates, statistical power is represented by the percentage of times that the Levene's median test, (T2) and the non-parametric Levene's test (T3), correctly rejected the null hypothesis.

3.3 Results

The Type I error rates for the Levene median test (T2) and the non-parametric Levene test (T3) for all of the conditions in the study are illustrated in Table 3.3. For example, the first row in Table 3.3 (reading across the row left to right), for a skew of 0, and a sample size of 24 with equal group sizes each containing 12 per group, the Type I error rate for the non-parametric Levene test is 4.9% and the Type I error rate for the Levene median test is 3.8%. In all of the conditions of the simulation, both tests maintain their Type I error rate, with the Levene median test (T2) being somewhat conservative in some of the conditions.

As mentioned previously, the power values of the Levene median test (T2) and the non-parametric Levene (T3) will only be investigated if the nominal Type I error rate was maintained. It was the case that the Type I error rates of both tests was maintained in all of the conditions of the present study. Table 3.4 reports the power values of the Levene median test (T2) and the non-parametric Levene tests when the population skew is equal to 0. In nearly all of the cells of the Table 3.4 the Levene median test (T2) has slightly higher power values. For example, in the first row of the table are the results for the non-parametric Levene test (T3), which, for a sample size of 24 with equal groups

and a ratio of variances of 5/1, the power is.42 (or 42% of the null hypotheses were correctly rejected). In comparison, the power of the Levene median (T2) test (the next row in the table) under the same conditions was.50 (or 50% of the null hypotheses were correctly rejected). In 61 of the 72 cells in Table 3.4, the median test had higher power than the non-parametric test.

The values for the power of the non-parametric Levene test (T3) and the Levene median test (T2) when the population distributions have a skew equal to 1 are illustrated in Table 3.5. Again, in most of the cases, the Levene median test (T2) had slightly higher power values than the non-parametric Levene test (T3); however the discrepancy between the scores is reduced. The power values are much closer than when the population skew was equal to 0. For example, in the first row of Table 3.5 are the power values for the non-parametric Levene (T3). For a sample size of 24 with equal groups and a ratio of variances 5/1, the power value is .474 (or 47.4% of the null hypotheses were correctly rejected. In comparison, the Levene median test (T2) under identical conditions has a power value of .434 (or 43.4% of the null hypotheses were correctly rejected). The median test was more powerful than the non-parametric test in 25 of the 72 cells in Table 3.5.

The power of the two tests when population skew is equal to 2 is listed in Table 3.6. In a great number of the cells of the table, the non-parametric Levene (T3) has higher power values than the Levene median test (T2). For example, the power for the non-parametric Levene test (T3) is present in the first row of Table 3.6. For a sample size of 24 with equal group sizes and a ratio of variances of 5/1, the power value is .572 (or 57.2% of the null hypotheses were correctly rejected. In comparison, the power of the

Levene median test under the same conditions is .296 (or 29.6% of the null hypotheses were correctly rejected. The non-parametric test was more powerful than the median test in every cell of Table 3.6.

When population skew was equal to 3, the greatest differences between the power values of the two tests were present and are illustrated in Table 3.7. The non-parametric Levene test (T3) has notably higher power values than the Levene median test (T2). For example, the first row of Table 3.7 lists the power values for the non-parametric Levene test (T3). For a sample size of 24 with equal group sizes and a ratio of variances of 5/1, the power value is .667 (or 66.7% of the null hypotheses were correctly rejected). In comparison, the power of the Levene median test (T2) is .155 (or 15.5% of the null hypotheses were correctly rejected. The non-parametric test was more powerful than the median in every cell of Table 3.7.

At this point, the results of some of the variables that were manipulated in this study will be investigated in a more in depth manner. In particular, the result of the main effects (overall sample size (N), ratio of sample sizes (n_1/n_2), inverse vs. direct pairings, and skew) will be illustrated graphically. As well, several of the interactions of these factors will be shown. For comparative purposes, especially when summarizing data across a number of conditions, the power difference between the two tests will be used. A power difference is simply the difference in statistical power between the two tests. The Levene median test will be used as a base because it is generally accepted as the gold standard. Essentially, if the power difference is negative, the Levene median test is performing better in terms of statistical power; if the power difference is positive, the non-parametric Levene test is performing better in terms of statistical power, and if the

difference score is zero, the two tests are equal as far as statistical power. This section of the results is intended to be a more direct comparison between the two tests.

3.3.1 2/1 variance ratio

Figure 3.2 illustrates the difference in power between the two versions of the test for equal variances across the 3 sample sizes (24, 48, and 96) for each of the four levels of skew (0, 1, 2, and 3) when the ratio of sample sizes are equal ($n_1/n_2 = 1/1$). It should be noted here that the ratio of variances (effect size) that is being described represents the smallest ratio of variances that was tested (2/1) because, in practice, it is the smaller effect sizes that usually are of interest to researchers and they also usually present more of a challenge when trying to detect differences than larger effect sizes.

It can be seen in Figure 3.2 that, when sample sizes are equal and small (24) and the distribution has a skew of 0 the median test has a slight power advantage over the non-parametric test and as sample size increases the non-parametric test enjoys a slight power advantage over the non-parametric test. However, as the skew of the population distribution increases, the non-parametric test has more power than the median version and this power advantage increases as the skew of the population distribution increases. For example, in the condition where skew=3, the power difference favors the non-parametric version of the test with a power difference of nearly .40 when the sample size is 24 and increasing to a power advantage of nearly .80 when the sample size is 96.

Figure 3.3 illustrates the power differences between the two tests across the three levels of sample sizes and four levels of skew when the ratio of variances is 2/1. As sample size increases, generally there is an increase in the power difference between the two tests that favors the non-parametric version of the test. Notice however, that when

the skew of the population distribution equals 0 and the pairing of sample size to variance ratio is inverse (i.e., the larger sample size with the smaller variance), the median version of the test has a slight power advantage over the non-parametric test that increases slightly as the sample size increases. Interestingly, under the same conditions except with a direct ratio of sample sizes and variance ratio (i.e., the larger sample size is associated with the larger variance) the power advantage of the median test over the non-parametric test is reduced. That is to say, when the larger sample size is associated with the smaller variance, the median test performs better compared to itself, under the same, condition than when the pairing is direct. Also, when the larger sample size is associated with the larger variance, the non-parametric test performs better compared to itself, under the same condition, when the sample size and variance are inversely paired. This is consistent across all of the levels of skew in Figure 3.2. For example, when skew=3 the non-parametric test clearly has more power than the median test, but when the ratio of sample size/variance ratio pairing are inverse the power differences are slightly less than when the pairing is direct demonstrating that the median test performs better in terms of power when there are inverse pairings. For example, see Table 3.7 for the condition where the skew=3, the sample size of 48, with a ratio of sample sizes 2 to 1 (16/8), and a 2 to 1 ratio of variances, the non-parametric Levene test's power is .677 (67.7%) when the sample size and variance are inversely related. When they are directly related, the power is .829 (82.9%). In contrast, the median version of the Levene test, under the same conditions, has a power of .123 (12.3%) when sample size and variance are inversely related. When sample size and variance are directed related the power of the median test is .060 (6%).

The power differences between the two tests across the three sample sizes and four levels of skew when the ratio of variances is 3 to 1 are shown in Figure 3.4. When the skew is equal to 0 and the ratio of sample sizes and variances ratios are inversely related, as the sample size increases, the median test gains a slight power advantage. This relation ship is reversed as the level of skew increases. Much like the conditions when there is a sample size ratio of 2 to 1, as the sample size and the skew increase, the power difference between the median test and the non-parametric test increases in favor of the non-parametric test. Much like in the result of the 2 to 1 sample size ratio, the power advantage was more pronounced when sample size ratios and variance ratios were directly paired and less so when sample sizes and variance ratios were inversely paired. This was due to the fact the median test performed generally better in terms of power when sample sizes and variance ratios were inversely paired than when sample sizes and variance ratios were directly paired.

3.3.2 4/1 variance ratio

This section of results focuses on the condition where the ratio of variances is 4 to 1. These results are report the same set of conditions as in the preceding section, but with a larger ratio of variances to add continuity and completeness to the results. Figure 3.5 illustrates the power difference between the two tests when sample sizes are equal across the three sample sizes. When the skew=0 the median test has a slight power advantage over the non-parametric test. This is reasonably stable across the three levels of sample size. As the skew increases, the power advantage begins to favor the non-parametric test and when the skew reaches 3, the power advantage of the non-parametric test is quite pronounced across all the sample sizes.

The reduction in the power difference between sample size of 48 and 96 when the skew is equal to 2 and 3 is explained by an increase in the power of the median test when the sample size becomes larger (i.e., 96). This is not due to a loss of power by the non-parametric test. It can be confirmed by looking at Tables 3.4 and 3.5 that the non-parametric test is correctly rejecting the null hypothesis nearly 100 percent of the time when the sample size is 96.

Figure 3.6 illustrates the power difference between the two tests when the ratio of variances is 4 to 1 across the three different sample sizes categorized by inverse versus direct pairing at each level of skew. The median test once again shows a power advantage when distributions are not skewed across all sample sizes in the simulation. As shown in previous results of this dissertation, as the skew of the population distributions increases the power advantage of the non-parametric test becomes more pronounced. As well, the power difference is less pronounced when the pairing of sample size ratios and variance ratios are inverse instead of direct.

Figure 3.7 shows the power difference for a 4 to 1 variance ratio across each of the sample size conditions when the sample size ratio is 3 to 1. Again the median test shows a power advantage when distributions are not skewed across all sample sizes in the simulation. As reported in previous conditions described in the results, as the skew of the population distributions increases, the power advantage of the non-parametric test becomes more pronounced. The power difference is more pronounced between the two tests when the pairing of sample size ratios and variance ratios is direct instead of inverse.

Figure 3.8 highlights how the ratio of sample sizes has an effect on the power difference between the non-parametric and median tests across the four different levels of

distributional skew when the ratio of variances is 2 to 1 and the sample size is 24. When the sample sizes are equal and the degree of skew in the population distribution is 0, the power difference between the two tests favors the median test slightly and as the skew increases, the power difference shifts in the favor of the non-parametric test. Using the case where the sample sizes are equal as a comparison, in both cases where the sample size ratios (2/1 and 3 /1) and variance ratios are inversely related, the power difference lines are below the comparison line. In the cases where the sample size ratios (2/1 and 3/1) and variance ratios are directly related, the power difference lines are above the comparison lines.

Figure 3.9 shows how the ratio of sample sizes has an effect on the power difference between the non-parametric and median tests across the four different levels of distributional skew when the ratio of variances is 4 to 1 and the sample size is 24. When the ratio of sample sizes is equal and the skew is 0, the power difference between the two tests is in the direction of the median test. Once again, as the skew of the population distribution increased, the power difference shifted in favor of the non-parametric test. Again using the condition where the sample sizes are equal as a comparison, when sample sizes (2/1 and 3/1) and the ratio of variances are inversely paired, the lines are below the comparison line, and when the ratio of sample sizes (2/1 and 3/1) and the ratio of variances are directly paired, the plotted lines are above the comparison line.

To investigate the relationship between the ratio of sample sizes and statistical power further, the power values of each test were plotted against each other. Figure 3.10 illustrates the results of inverse pairing between the sample size and the variance from the cells of the design where (N = 24, skew = 3, and the ratio of variances is 1/5. The

power of the non-parametric test was influenced by the ratio of sample sizes when the sample size and the variance were inversely paired. As the numbers in each of the groups became more unbalanced the power of the non-parametric test was reduced. The median test was more robust against unbalanced numbers in each of the groups when sample sizes and variances are inversely paired. As the sample size ratio becomes larger, the power of the non-parametric test is reduced by .213 (21.3%), whereas the power of the median test experiences a slight increase in power of .066 (6.60%). The non-parametric test was more powerful across all of the levels in Figure 3.10.

Figure 3.11 illustrates the results of direct pairing between the sample size and the variance from the cells of the design where (N = 24, skew = 3, and the ratio of variances is 1/5. As the sample size ratio increases from 1/1 to 3/1, the non-parametric test maintains its power with minor fluctuations in power values. For the median test, as the sample size ratio increases from 1/1 to 3/1, the power values decrease, with a power loss of .112 (11.2%) when the sample size ratio is 3/1. Again, the non-parametric test was more powerful across all of the levels in Figure 3.11.

Figure 3.12 illustrates the power comparison between the two tests across four variance ratios (2/1, 3/1, 4/1, 5/1 from left to right) when sample sizes are small (24) and equal (12 per group) and the population distributions are heavily skewed (3). When data are heavily skewed (3), the non-parametric based test is consistently more powerful than the median test, and becomes more powerful as the variance ratio increases. It is evident from the results that as the population distribution becomes more skewed, the non-parametric Levene becomes more powerful and the Levene median test becomes less powerful across all the levels of variance ratios.

3.4 Discussion

When data come from heavily skewed population distributions, the non-parametric version of the Levene test performs quite well in terms of maintaining its Type I error rate and statistical power. The median version of the test consistently showed a power advantage over the non-parametric test when population distributions had skew=0. This is interesting because, when data come from skewed population distributions, the median test lacks substantive power compared to the non-parametric test. This leaves the situations when the population distributions are normal where the median test is more powerful than the non-parametric test. Recalling the results of Nordstokke and Zumbo (2007), when data are sampled from normal or symmetrically distributed populations, the Levene mean test has suitable statistical power, leaving the gold standard of tests for equal variances "out in the cold" so to speak. If one was to use Levene mean test when distributions are considered normally distributed and the non-parametric version of the Levene test when distributions are considered to be skewed, then this leaves very few options to use the median version of the Levene test. This is only generalizable to the limits of the conditions investigated in the simulations.

An interesting finding is that, when sample sizes were inversely paired with the ratio of variances (i.e., large sample size paired with the smaller variance), the median test has an increase in power, compared to itself when the pairing is direct, holding all other conditions constant. When the ratio of sample sizes were directly paired with the ratio of variances (i.e., large sample size paired with the larger variance), the non-parametric test experiences an increase in power, compared to itself when the pairing is inverse, holding all other conditions used in the simulation constant. This may be

attributed to the fact that, when sample sizes are directly paired with the ratio of variances, the sums of squares between (SS_B) becomes distorted. In the case of direct pairing, it becomes attenuated, thus leading to a smaller value for the SS_B, hence leading to relatively larger sums of squares within (SS_W). The relationship can be expressed as ($SS_W = SS_T - SS_B$), where SS_T is the total sums of squares in the model. Direct pairing will affect the values of the mean squares within (MS_W) and between (MS_B), resulting in a reduction in the MS_B and an inflation of the MS_W, resulting in a reduction of power because a reduction in the MS_B and an inflation of the MS_W will lead to fewer rejections of the null hypothesis even when true differences are present (Type II errors).

Interestingly, as the skew of the population distribution increased, the median version of the test for equal variances was affected more by the direct pairing of the sample sizes and ratio of variances and became less powerful. The opposite occurred for the non-parametric test; it was less affected by the direct pairing as distributions became more skewed. This could be related to the nature of the rank transformation that may moderate the effect of design imbalance when calculating the mean of the ranks and the SS_B (i.e., controlling for the attenuation of the SS_B). This suggests that, even when designs are unbalanced and population distributions are heavily skewed, the non-parametric test possesses good statistical properties and should be implemented by researchers.

As pointed out by Bridge and Sawilowsky (1999), it often may be the case that the applied researcher does not know the shape of the population distribution that they are sampling from and thus should more often choose a non-parametric version of tests to maintain efficiency by increasing the odds that the test selected has sufficient power to

correctly reject the null hypothesis. As noted by Kruskal and Wallis (1952), the advantage of ranks is that only very general assumptions are made about the kind of distributions from where the observations come, which is that the distributions have the same form. This provides researchers and statisticians a great deal of flexibility with their analyses. Put in the context of a test for equal variances, if applied researchers are unsure of the shape of the population distribution, they should employ the non-parametric Levene test for equality of variances.

Even though, in many cases, the Levene median test has higher power under non-skewed distributions, both tests have quite low power values under these conditions suggesting that neither of these tests should be used when there is evidence to suggest a normally distributed dependent variable. Selection of another test such as the mean version of the Levene test is recommended when the normality assumption is tenable. It is imperative that data analysts and researchers use such test selection strategies when analyzing data because it reduces the chance that incorrect decisions are made based on incorrect results from a statistical test. Investigating empirical distributions provides some evidence about the nature of the population distribution. This, plus prior knowledge (previous empirical work) of the dependent variable, should help guide statistical practices and allow an approximate estimation of the shape of the population distribution. Based on this information, the most appropriate version of the test for equality of variances may be selected.

One limitation of this study is that the two tests were compared on only one distributional form. All of the distributions used in the simulations were based on χ^2, thus limiting the range of distributional properties investigated. This is not really a limitation,

per se, because it does not disqualify the results of the study, but implies that future work should focus on other distributions, for example, multimodal distributions.

To summarize, this paper investigated the Type I error rates and statistical power of the median version of the Levene test and the new non-parametric version of the Levene test for equal variances. In cases where samples were generated from population distributions with increasing skew, the non-parametric version of the Levene test was superior in statistical power to the median version of the Levene test. It is recommended that data analysts and researchers use the non-parametric Levene test when there is evidence that data come from populations with skewed distributions.

Table 3.1 Steps in calculating the Levene median test

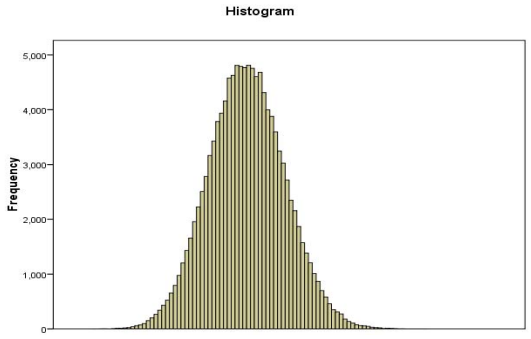
Levene median test			
Group	Score (Xij)	Median (Mdnj)	|(Xij) - (Mdnj)|
1	9.91	9.76	.15
1	9.27	9.76	.49
1	11.91	9.76	2.15
1	9.93	9.76	.17
1	9.21	9.76	.55
1	9.17	9.76	.59
1	9.62	9.76	.14
1	10.33	9.76	.57
1	9.25	9.76	.51
1	9.93	9.76	.17
1	9.90	9.76	.14
1	9.32	9.76	.44
2	8.16	8.64	.48
2	8.76	8.64	.12
2	8.27	8.64	.37
2	9.68	8.64	1.04
2	11.22	8.64	2.58
2	10.03	8.64	1.39
2	10.07	8.64	1.43
2	8.18	8.64	.46
2	8.53	8.64	0.11
2	8.29	8.64	0.35
2	8.17	8.64	0.47
2	21.40	8.64	12.76

Table 3.2 Steps in calculating the non-parametric Levene test

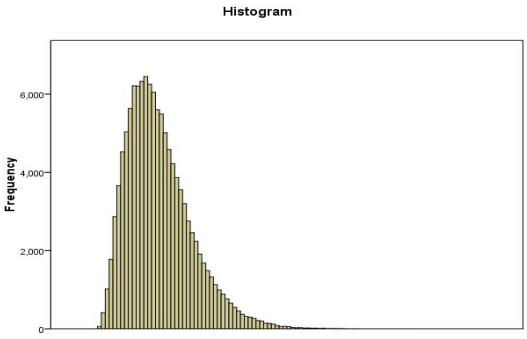
Non-parametric Levene test				
Group	Score (Xij)	Rank (Rij)	Mean rank	|Rij - mean rank|
1	9.91	16.00	14.42	1.58
1	9.27	11.00	14.42	3.42
1	11.91	23.00	14.42	8.58
1	9.93	18.00	14.42	3.58
1	9.21	9.00	14.42	5.42
1	9.17	8.00	14.42	6.42
1	9.62	13.00	14.42	1.42
1	10.33	21.00	14.42	6.58
1	9.25	10.00	14.42	4.42
1	9.93	17.00	14.42	2.58
1	9.90	15.00	14.42	0.58
1	9.32	12.00	14.42	2.42
2	8.16	1.00	10.58	9.58
2	8.76	7.00	10.58	3.58
2	8.27	4.00	10.58	6.58
2	9.68	14.00	10.58	3.42
2	11.22	22.00	10.58	11.42
2	10.03	19.00	10.58	8.42
2	10.07	20.00	10.58	9.42
2	8.18	3.00	10.58	7.58
2	8.53	6.00	10.58	4.58
2	8.29	5.00	10.58	5.58
2	8.17	2.00	10.58	8.58
2	21.40	24.00	10.58	13.42

Figure 3.1 Shape of population distributions used in simulations

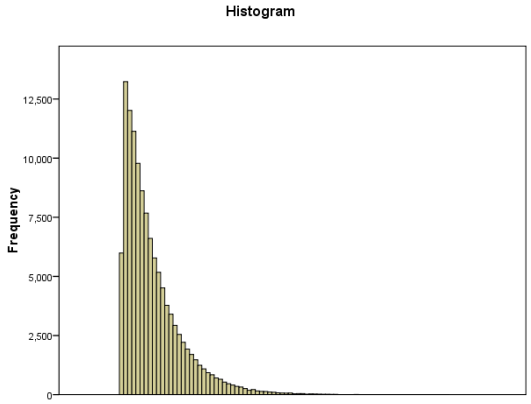
Skew = 0



Skew = 1



Skew = 2



Skew = 3

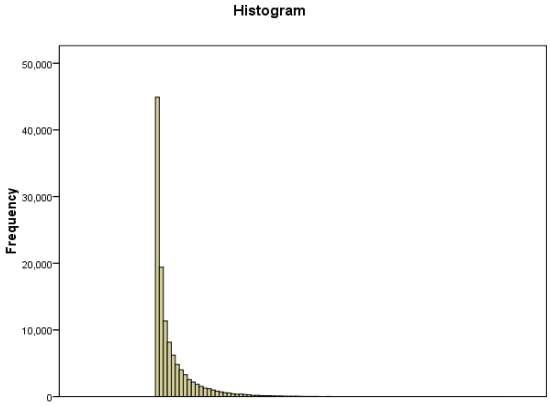


Table 3.3 Type I error rates of the Non-parametric and Median versions of the Levene tests

N	n1/n2	Non-parametric Levene	Levene Median
Skew = 0			
24	**1/1**	4.9	3.8
24	**2/1**	5.0	3.7
24	**3/1**	4.7	3.9
48	**1/1**	4.4	3.9
48	**2/1**	5.3	4.3
48	**3/1**	5.4	4.6
96	**1/1**	4.7	4.0
96	**2/1**	5.1	4.3
96	**3/1**	5.1	4.3
Skew = 1			
24	**1/1**	4.3	4.0
24	**2/1**	4.8	4.1
24	**3/1**	4.9	3.9
48	**1/1**	4.6	4.1
48	**2/1**	4.8	4.0
48	**3/1**	5.8	4.4
96	**1/1**	5.0	5.2
96	**2/1**	4.8	4.4
96	**3/1**	4.7	4.2
Skew = 2			
24	**1/1**	5.1	5.0
24	**2/1**	4.9	4.7
24	**3/1**	5.4	5.0
48	**1/1**	5.3	5.5
48	**2/1**	5.2	4.7
48	**3/1**	5.3	4.5
96	**1/1**	5.1	5.1
96	**2/1**	4.9	5.0
96	**3/1**	5.4	4.8
Skew =3			
24	**1/1**	4.9	5.3
24	**2/1**	5.4	5.0
24	**3/1**	4.6	4.9
48	**1/1**	4.9	4.5
48	**2/1**	4.6	4.3
48	**3/1**	5.0	4.4
96	**1/1**	4.5	4.4
96	**2/1**	5.4	4.8
96	**3/1**	5.0	4.3

Table 3.4 Power values of the non-parametric and median versions of the Levene test for equality of variances for skew of zero

			Population Variance Ratio, σ_1^2 / σ_2^2							
Test	N	n1/n2	**1/5**	**1/4**	**1/3**	**1/2**	**2/1**	**3/1**	**4/1**	**5/1**
Skew = 0			**Inverse Pairings**				**Direct Pairings**			
Non-parametric Levene	24	1/1	42.0	35.4	23.9	12.4	12.4	23.9	35.4	42.0
Levene Median	24	1/1	50.0	40.7	27.2	13.2	13.2	27.2	40.7	50.0
Non-parametric Levene	24	2/1	32.6	26.0	18.3	9.5	14.0	24.6	37.0	45.9
Levene Median	24	2/1	50.9	40.1	26.7	12.2	9.8	19.7	32.0	39.3
Non-parametric Levene	24	3/1	25.6	20.6	14.5	7.3	13.0	22.2	31.4	40.2
Levene Median	24	3/1	46.6	35.1	24.4	11.0	8.8	13.9	21.4	28.5
Non-parametric Levene	48	1/1	78.3	67.3	48.4	22.2	22.2	48.4	67.3	78.3
Levene Median	48	1/1	89.9	80.5	59.5	27.2	27.2	59.5	80.5	89.9
Non-parametric Levene	48	2/1	66.1	54.6	39.1	18.8	23.1	48.0	68.1	79.6
Levene Median	48	2/1	87.4	75.5	56.6	26.3	24.2	51.5	72.8	85.1
Non-parametric Levene	48	3/1	53.5	44.8	31.5	15.6	21.1	44.3	63.1	74.6
Levene Median	48	3/1	80.5	69.8	50.8	23.7	19.3	42.5	62.5	75.5
Non-parametric Levene	96	1/1	98.0	94.3	78.7	43.8	43.8	78.7	94.3	98.0
Levene Median	96	1/1	99.8	98.8	91.3	56.0	56.0	91.3	98.8	99.8
Non-parametric Levene	96	2/1	95.0	87.6	69.9	37.8	42.6	79.8	94.1	98.0
Levene Median	96	2/1	99.4	97.6	87.9	53.2	49.3	88.4	98.1	99.7
Non-parametric Levene	96	3/1	87.5	77.9	60.7	29.3	37.4	73.9	91.1	97.0
Levene Median	96	3/1	98.4	95.1	83.3	43.9	40.1	81.5	95.7	98.9

Table 3.5 Power values of the non-parametric and median versions of the Levene test for equality of variances for skew of one

			Population Variance Ratio, σ_1^2 / σ_2^2							
Test	N	n1/n2	**1/5**	**1/4**	**1/3**	**1/2**	**2/1**	**3/1**	**4/1**	**5/1**
Skew = 1			**Inverse Pairings**				**Direct Pairings**			
Non-parametric Levene	24	1/1	47.4	38.5	27.8	14.2	14.2	27.8	38.5	47.4
Levene Median	24	1/1	43.4	34.0	22.9	12.0	12.0	22.9	34.0	43.4
Non-parametric Levene	24	2/1	35.7	29.3	20.2	10.5	15.4	29.6	41.6	50.5
Levene Median	24	2/1	43.6	34.8	23.4	11.7	9.0	16.7	25.3	32.1
Non-parametric Levene	24	3/1	28.5	23.0	16.0	8.7	14.3	24.4	35.6	43.9
Levene Median	24	3/1	42.4	31.2	21.4	11.3	6.9	11.2	16.8	22.1
Non-parametric Levene	48	1/1	83.6	73.0	56.6	27.6	27.6	56.6	73.0	83.6
Levene Median	48	1/1	82.0	70.5	51.5	24.1	24.1	42.4	70.5	82.0
Non-parametric Levene	48	2/1	71.5	60.9	43.9	21.5	26.1	55.8	73.2	85.2
Levene Median	48	2/1	80.4	68.1	48.5	22.8	18.2	41.8	58.7	74.0
Non-parametric Levene	48	3/1	58.6	48.5	35.0	17.8	25.3	52.4	70.1	81.5
Levene Median	48	3/1	73.5	61.3	42.8	20.6	15.6	33.0	50.0	63.4
Non-parametric Levene	96	1/1	99.1	96.6	87.8	52.2	52.2	87.8	96.6	99.1
Levene Median	96	1/1	99.1	96.5	85.2	46.6	46.6	85.2	96.5	99.1
Non-parametric Levene	96	2/1	96.6	91.3	77.4	42.3	49.2	86.0	97.0	99.3
Levene Median	96	2/1	98.7	94.1	80.7	43.0	38.9	77.4	93.7	98.0
Non-parametric Levene	96	3/1	90.5	83.7	67.3	35.9	46.3	82.2	94.7	98.6
Levene Median	96	3/1	96.4	90.8	75.2	39.0	33.5	68.9	87.3	95.9

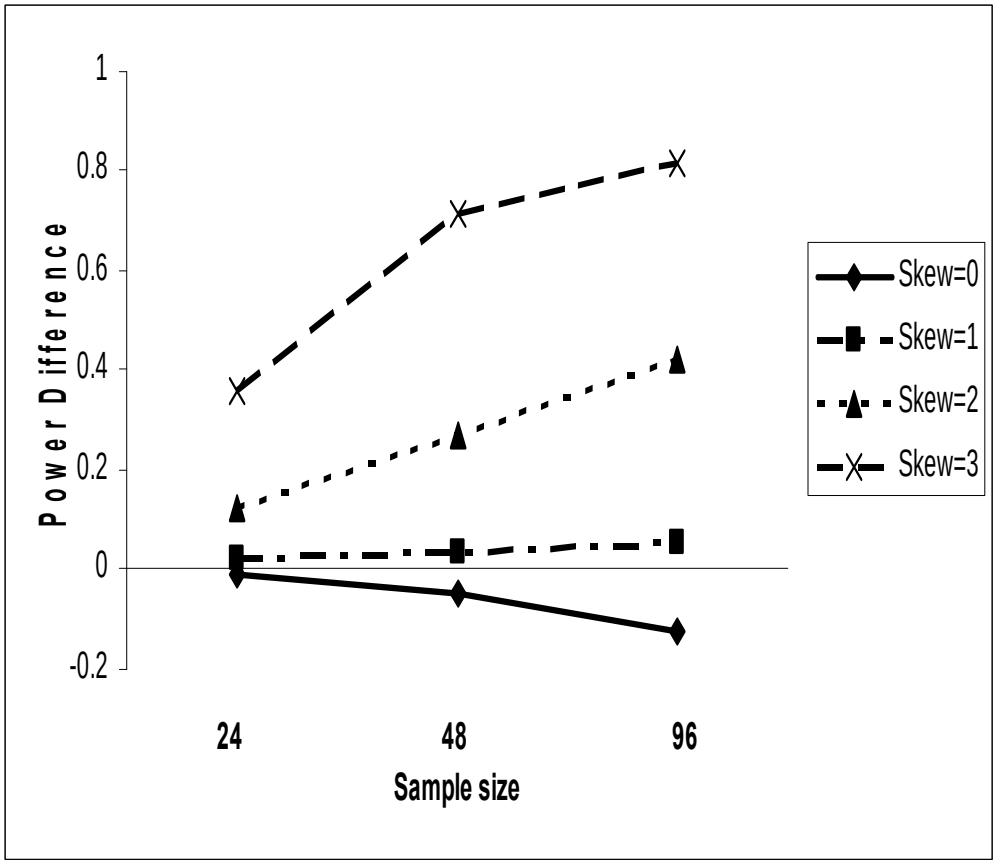
Table 3.6 Power values of the non-parametric and median versions of the Levene test for equality of variances for skew of two

			Population Variance Ratio, σ_1^2 / σ_2^2							
Test	N	n1/n2	**1/5**	**1/4**	**1/3**	**1/2**	**2/1**	**3/1**	**4/1**	**5/1**
Skew = 2			**Inverse Pairings**				**Direct Pairings**			
Non-parametric Levene	24	1/1	57.2	49.9	37.6	21.4	21.4	37.6	49.9	57.2
Levene Median	24	1/1	29.6	23.8	16.6	9.1	9.1	16.6	23.8	29.6
Non-parametric Levene	24	2/1	43.1	36.4	27.5	15.1	22.2	40.5	51.7	61.2
Levene Median	24	2/1	34.2	26.6	19.5	10.1	6.7	11.8	14.3	19.3
Non-parametric Levene	24	3/1	33.3	29.8	22.9	12.8	20.3	35.8	46.6	55.1
Levene Median	24	3/1	32.6	26.7	18.9	11.2	5.6	7.4	10.3	12.7
Non-parametric Levene	48	1/1	92.0	86.4	72.3	42.8	42.8	72.3	86.4	92.0
Levene Median	48	1/1	62.9	52.2	34.6	15.5	15.5	34.6	52.2	62.9
Non-parametric Levene	48	2/1	79.5	72.3	60.2	35.1	43.6	74.6	88.4	94.0
Levene Median	48	2/1	62.2	51.4	35.3	18.1	12.2	26.8	39.0	49.5
Non-parametric Levene	48	3/1	66.3	59.6	47.2	25.3	39.5	70.4	84.6	91.7
Levene Median	48	3/1	58.3	47.2	32.6	15.6	9.3	18.5	27.1	34.2
Non-parametric Levene	96	1/1	99.9	99.5	96.4	74.0	74.0	96.4	99.5	99.9
Levene Median	96	1/1	92.6	83.4	66.2	31.5	31.5	66.2	83.4	92.6
Non-parametric Levene	96	2/1	98.8	97.2	90.7	64.8	74.1	96.5	99.6	99.9
Levene Median	96	2/1	90.9	81.3	62.4	30.8	24.8	53.8	74.9	86.4
Non-parametric Levene	96	3/1	94.6	91.4	82.0	54.7	68.8	95.0	99.2	99.9
Levene Median	96	3/1	87.1	76.8	57.4	27.4	19.6	44.7	64.6	77.9

Table 3.7 Power values of the non-parametric and median versions of the Levene test for equality of variances for skew of three

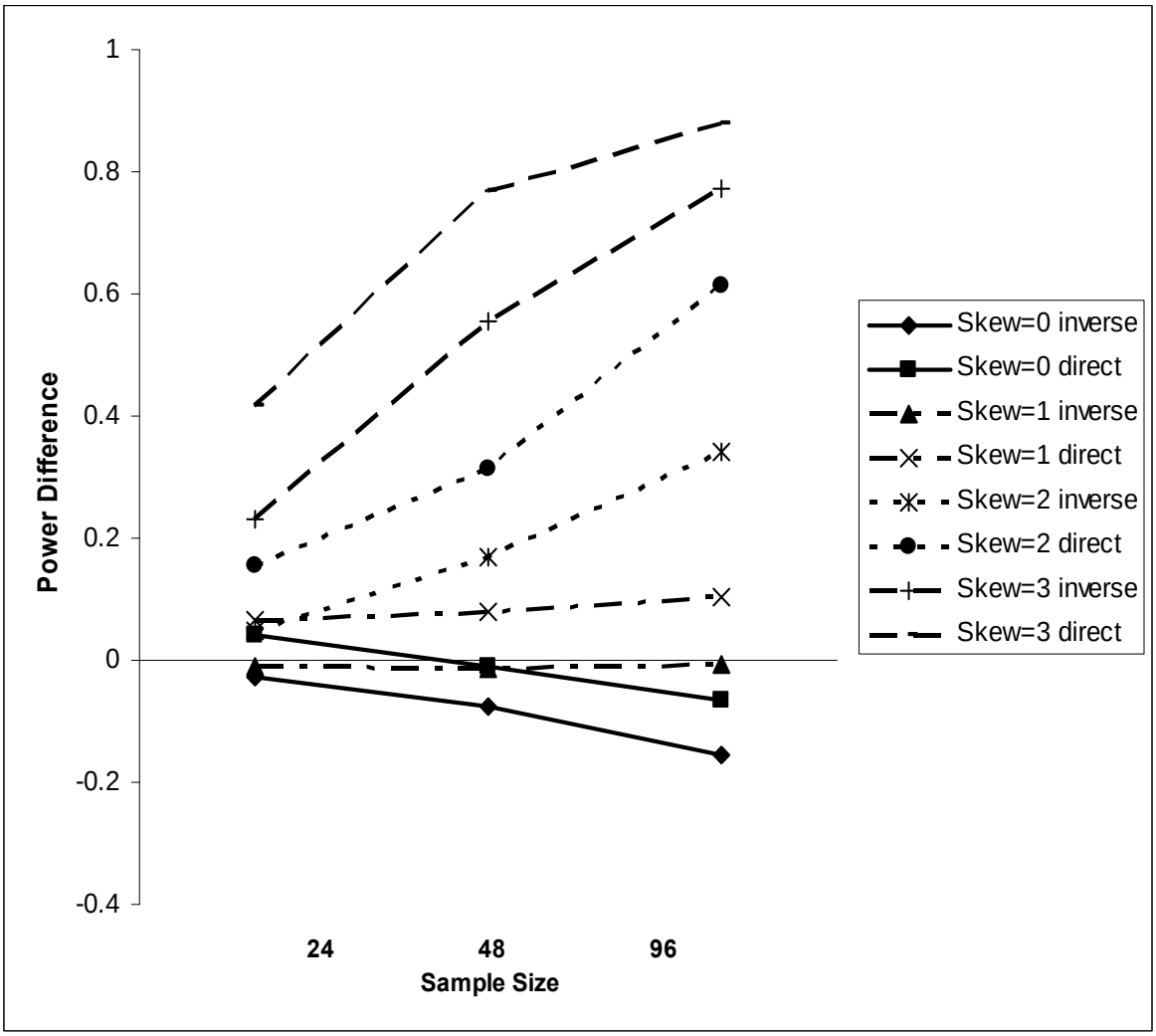
			Population Variance Ratio, σ_1^2 / σ_2^2							
Test	N	n1/n2	**1/5**	**1/4**	**1/3**	**1/2**	**2/1**	**3/1**	**4/1**	**5/1**
Skew = 3			**Inverse Pairings**				**Direct Pairings**			
Non-parametric Levene	24	1/1	66.7	62.2	56.5	44.3	44.3	56.5	62.2	66.7
Levene Median	24	1/1	15.5	12.4	9.4	7.3	7.3	9.4	12.4	15.5
Non-parametric Levene	24	2/1	50.4	48.5	41.9	32.0	46.1	60.3	67.6	71.2
Levene Median	24	2/1	21.5	18.3	12.7	8.8	4.4	5.3	7.1	7.6
Non-parametric Levene	24	3/1	45.4	40.5	33.3	25.5	41.1	56.8	62.7	69.5
Levene Median	24	3/1	22.1	18.7	14.4	9.6	3.1	3.2	3.4	4.3
Non-parametric Levene	48	1/1	95.4	94.4	91.9	81.0	81.0	91.9	94.4	95.4
Levene Median	48	1/1	31.9	25.8	17.8	9.3	9.3	17.8	25.8	31.9
Non-parametric Levene	48	2/1	84.3	82.9	78.9	67.7	82.9	95.6	98.3	99.3
Levene Median	48	2/1	37.3	29.9	23.0	12.3	6.0	10.4	14.6	18.5
Non-parametric Levene	48	3/1	73.5	69.7	67.1	52.5	79.2	93.7	97.7	99.0
Levene Median	48	3/1	37.6	29.0	21.0	11.8	4.0	5.5	8.2	10.5
Non-parametric Levene	96	1/1	99.9	99.9	99.9	98.7	98.7	99.9	99.9	99.9
Levene Median	96	1/1	64.8	49.9	35.1	16.8	16.8	35.1	49.9	64.8
Non-parametric Levene	96	2/1	98.4	98.6	98.4	95.0	98.9	99.9	99.9	99.9
Levene Median	96	2/1	65.7	54.0	38.2	17.8	10.9	24.1	37.7	49.4
Non-parametric Levene	96	3/1	94.6	94.7	93.8	86.7	98.4	99.9	99.9	99.9
Levene Median	96	3/1	61.5	51.1	35.2	18.6	8.1	17.4	26.2	33.9

Figure 3.2 Power difference (2 to 1 variance ratio) for equal sample size ratios



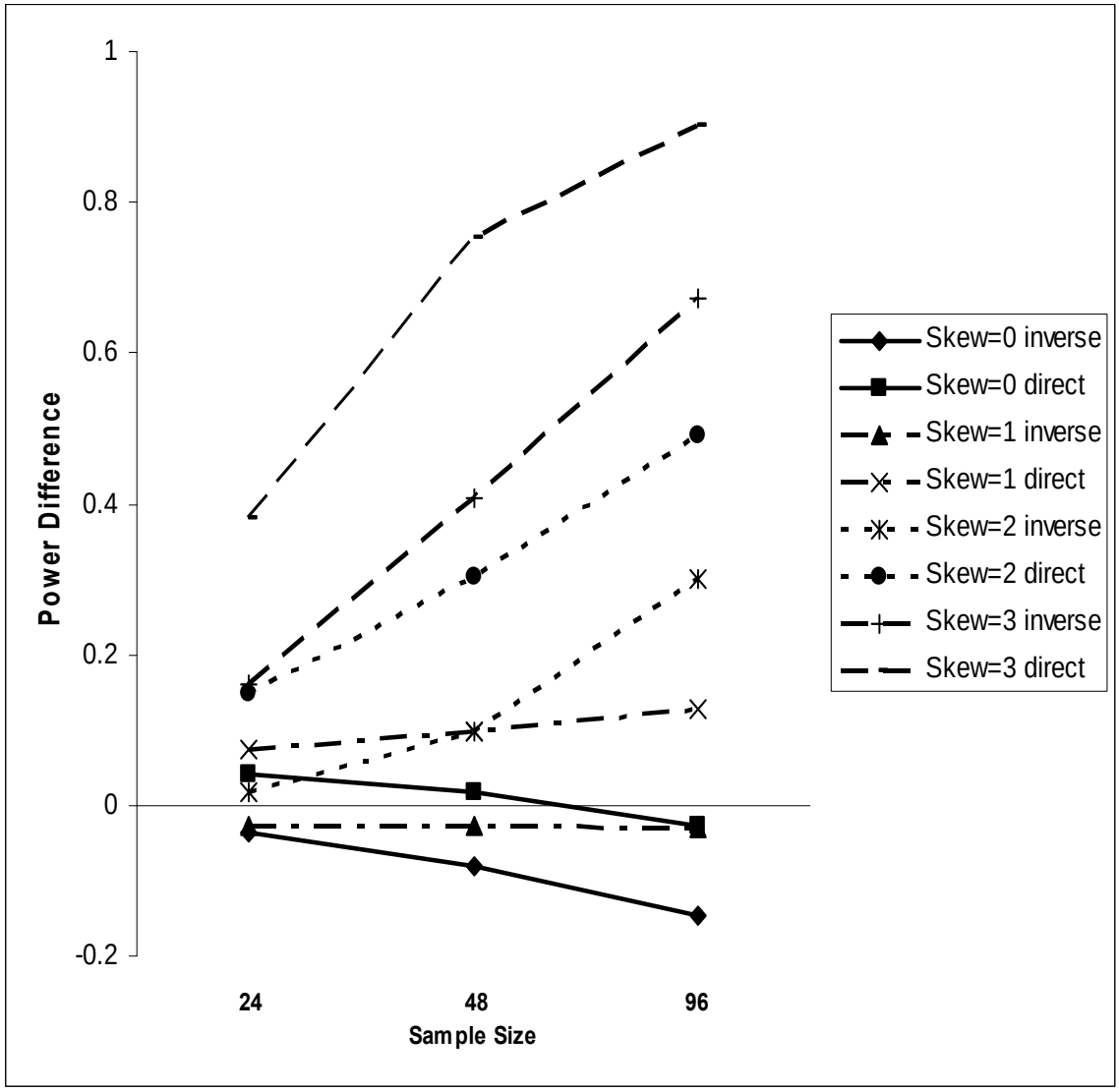
Note: Power differences are based on the non-parametric test minus the median test with a negative value representing superior power for the median test and a positive value representing superior power for the non-parametric test.

Figure 3.3 Power difference (2 to 1 variance ratio) values for sample size ratio of 2 to 1



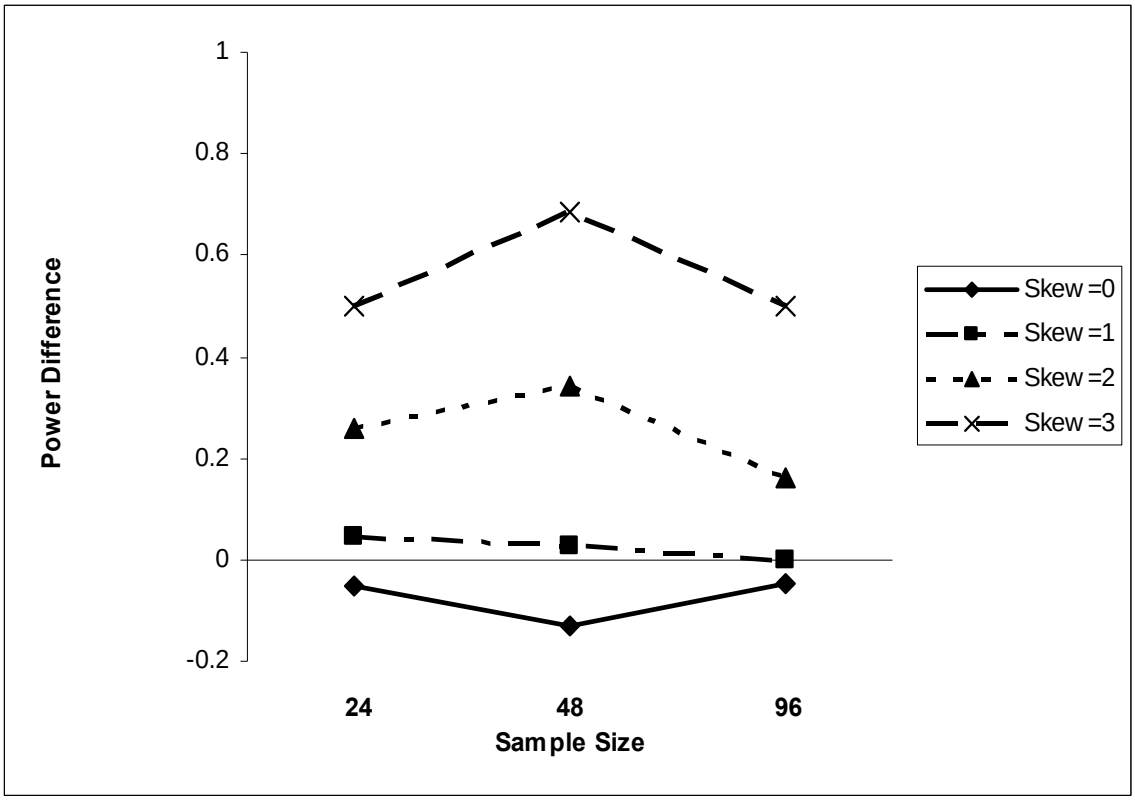
Note: Power differences are based on the non-parametric test minus the median test with a negative value representing superior power for the median test and a positive value representing superior power for the non-parametric test.

Figure 3.4 Power difference (2 to 1 variance ratio) values for sample size ratio of 3 to 1



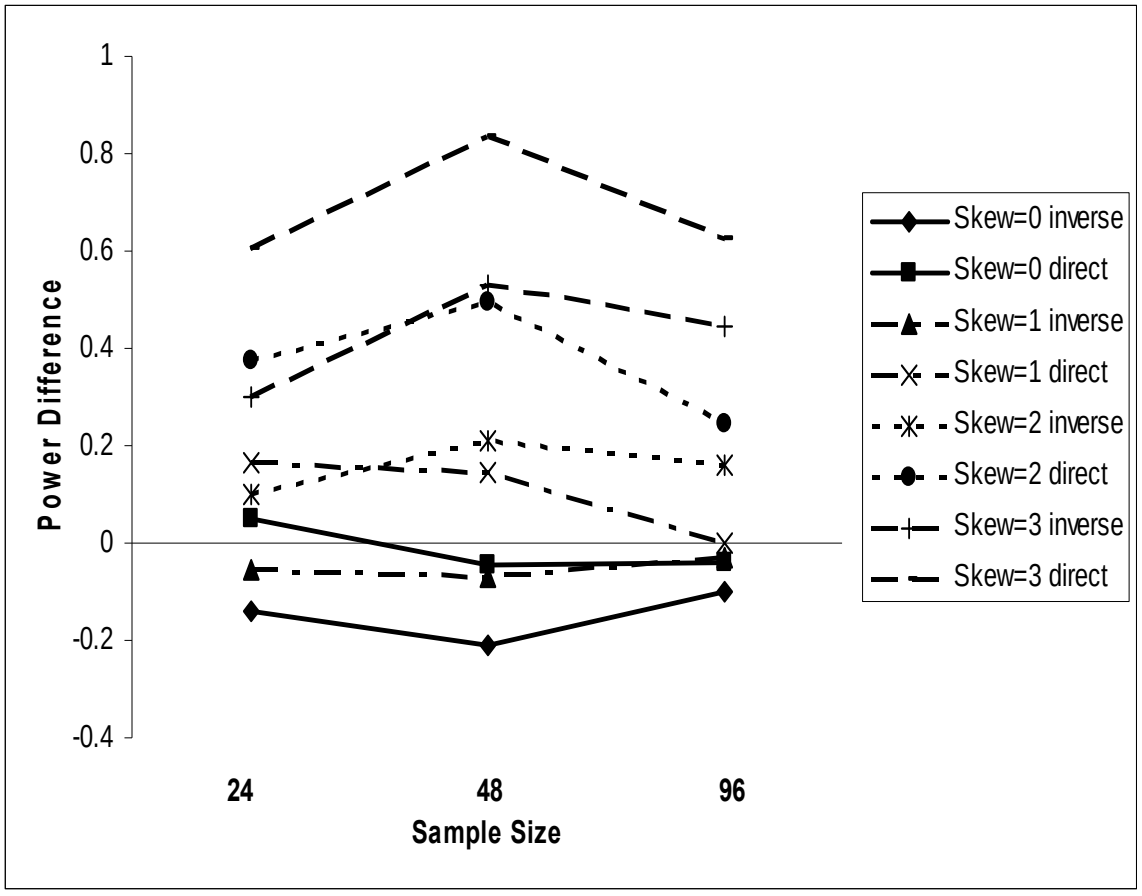
Note: Power differences are based on the non-parametric test minus the median test with a negative value representing superior power for the median test and a positive value representing superior power for the non-parametric test.

Figure 3.5 Power difference (4 to 1 variance ratio) for equal sample size ratios



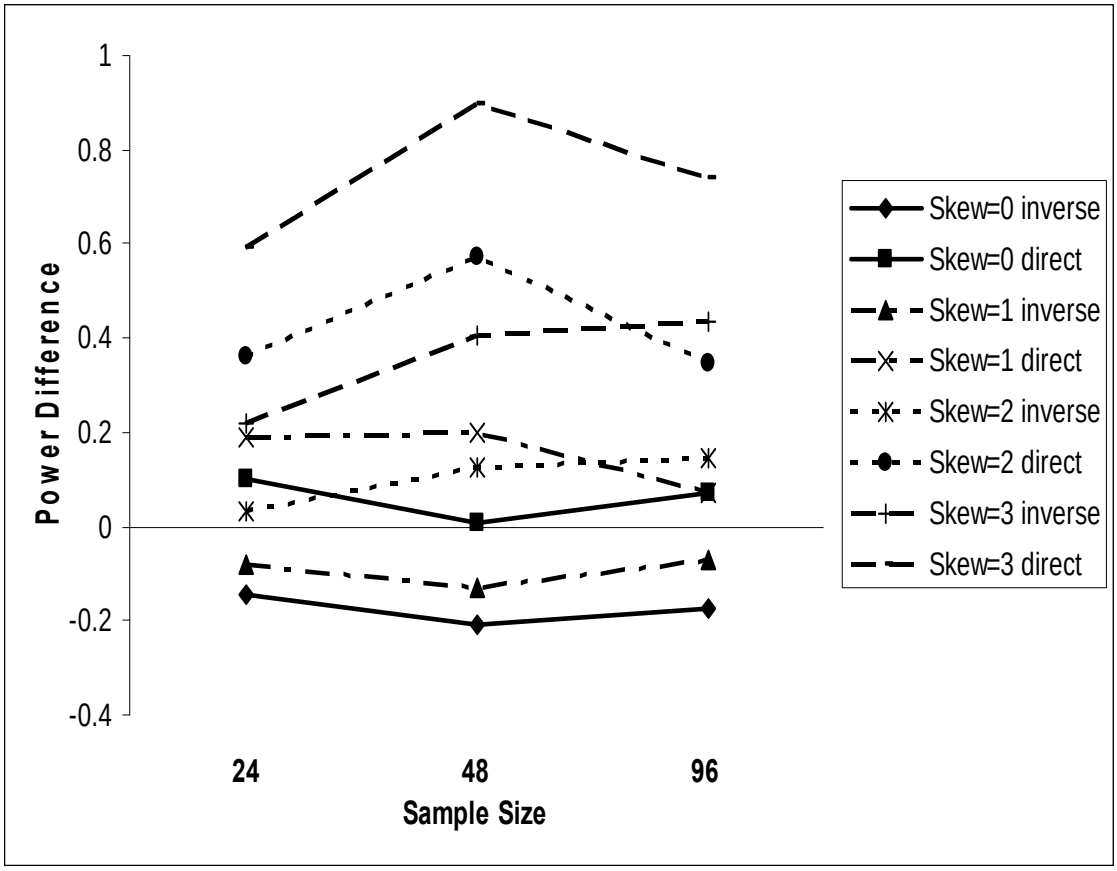
Note: Power differences are based on the non-parametric test minus the median test with a negative value representing superior power for the median test and a positive value representing superior power for the non-parametric test.

Figure 3.6 Power difference (4 to 1 variance ratio) values for sample size ratio of 2 to 1



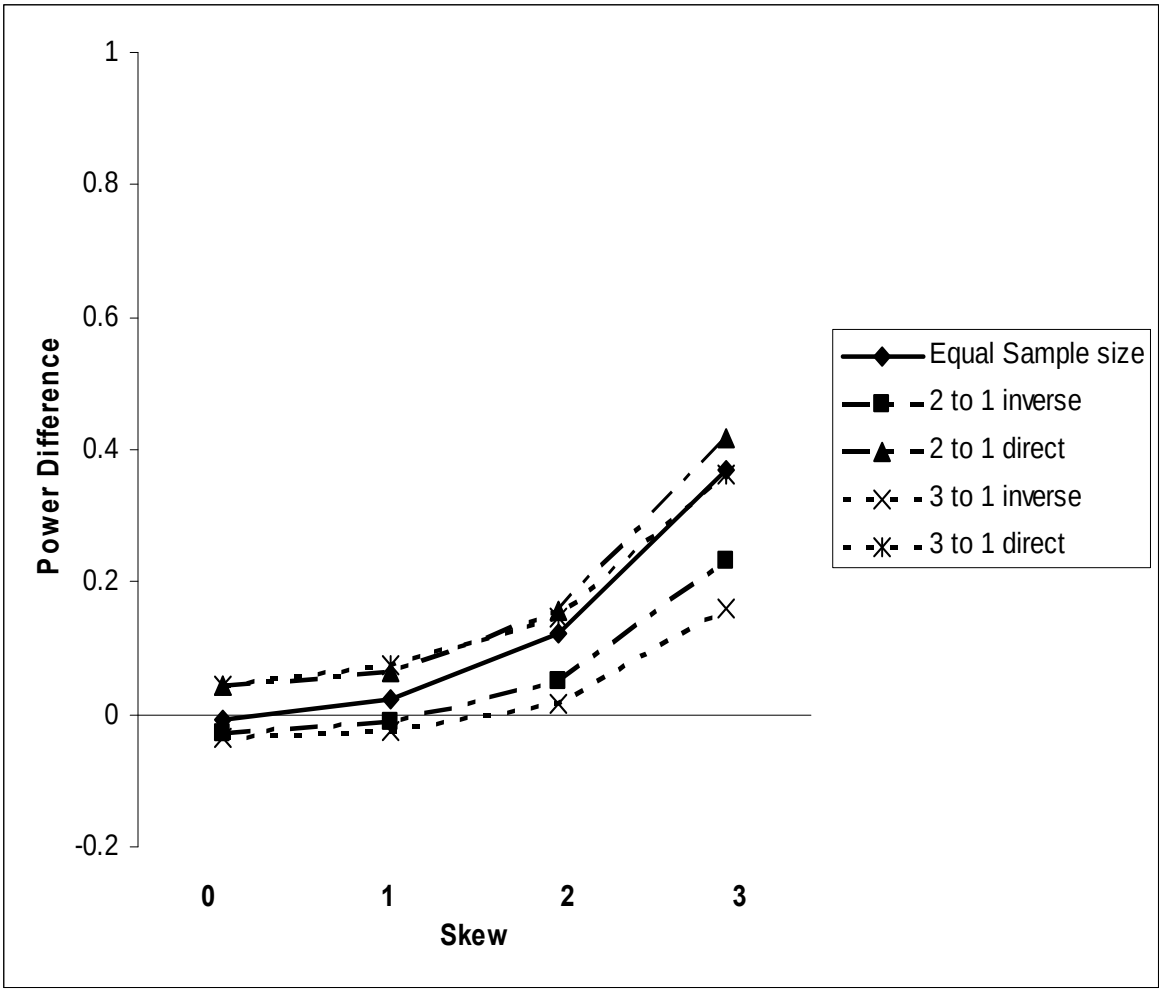
Note: Power differences are based on the non-parametric test minus the median test with a negative value representing superior power for the median test and a positive value representing superior power for the non-parametric test.

Figure 3.7 Power difference (4 to 1 variance ratio) values for sample size ratio of 3 to 1



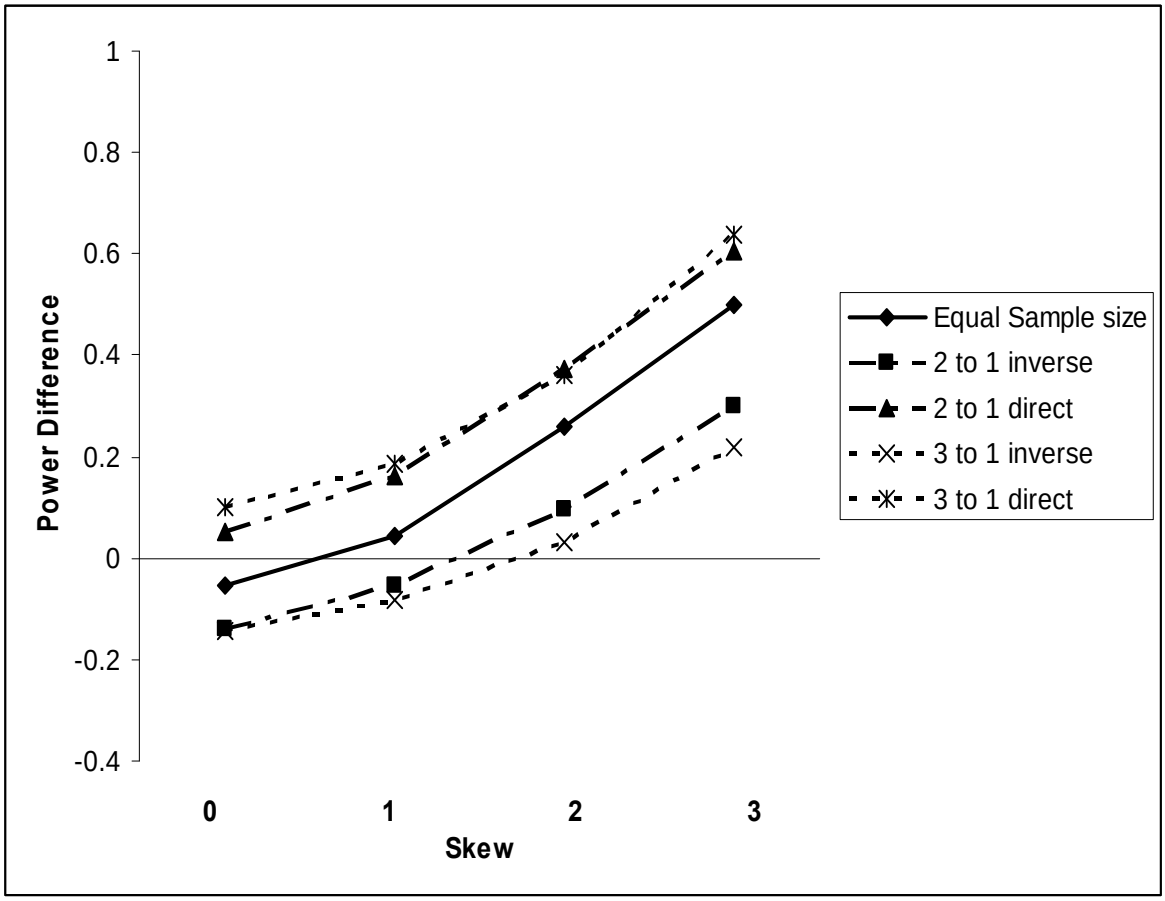
Note: Power differences are based on the non-parametric test minus the median test with a negative value representing superior power for the median test and a positive value representing superior power for the non-parametric test.

Figure 3.8 Power Difference (2 to 1 variance ratio) across levels of skew



Note: Power differences are based on the non-parametric test minus the median test with a negative value representing superior power for the median test and a positive value representing superior power for the non-parametric test.

Figure 3.9 Power Difference (4 to 1 variance ratio) across levels of skew



Note: Power differences are based on the non-parametric test minus the median test with a negative value representing superior power for the median test and a positive value representing superior power for the non-parametric test.

Figure 3.10 Power comparisons between the Levene median and the non-parametric Levene tests across simulated sample size ratios for inverse pairings

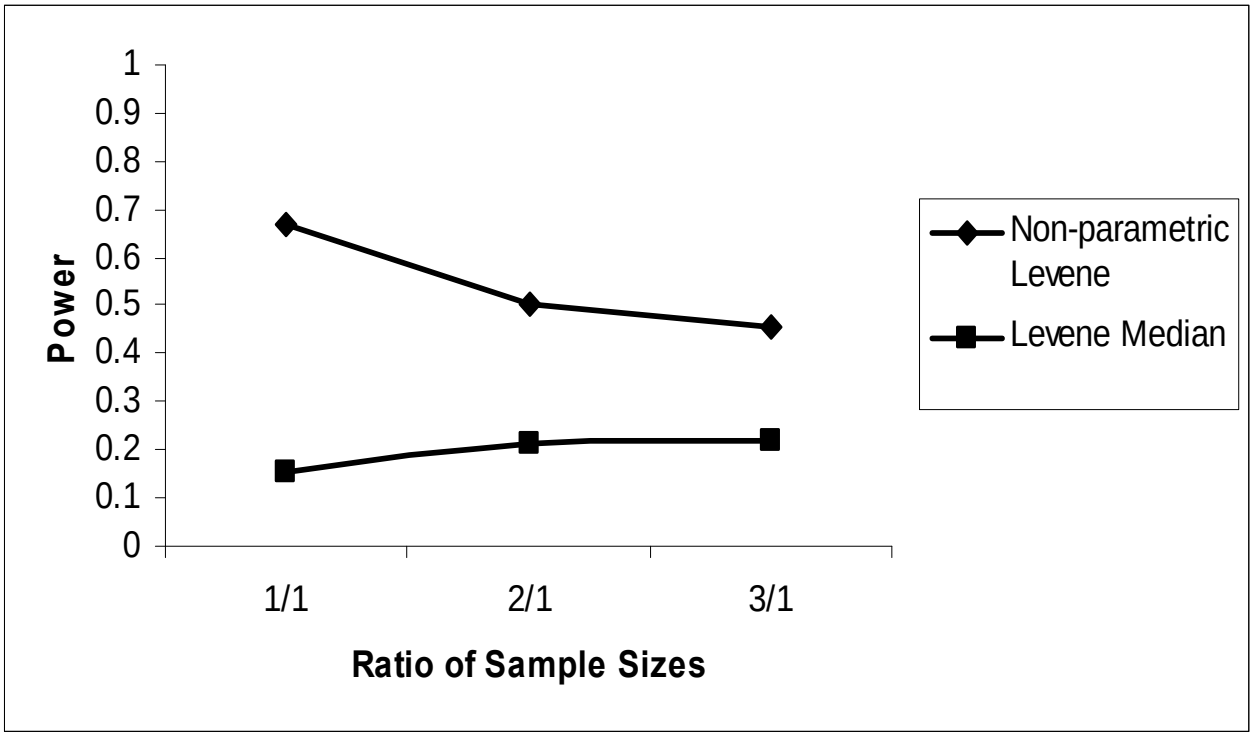


Figure 3.11 Power comparisons between the Levene median and the non-parametric Levene tests across simulated sample size ratios for direct pairings.

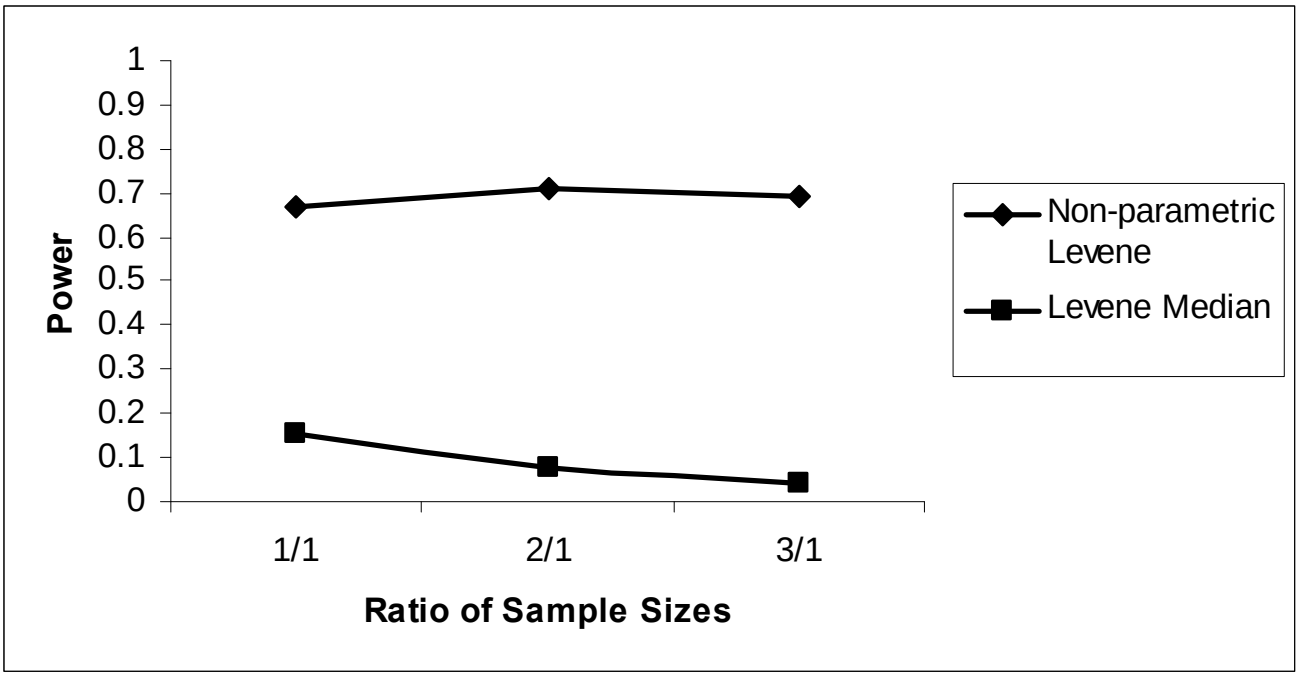
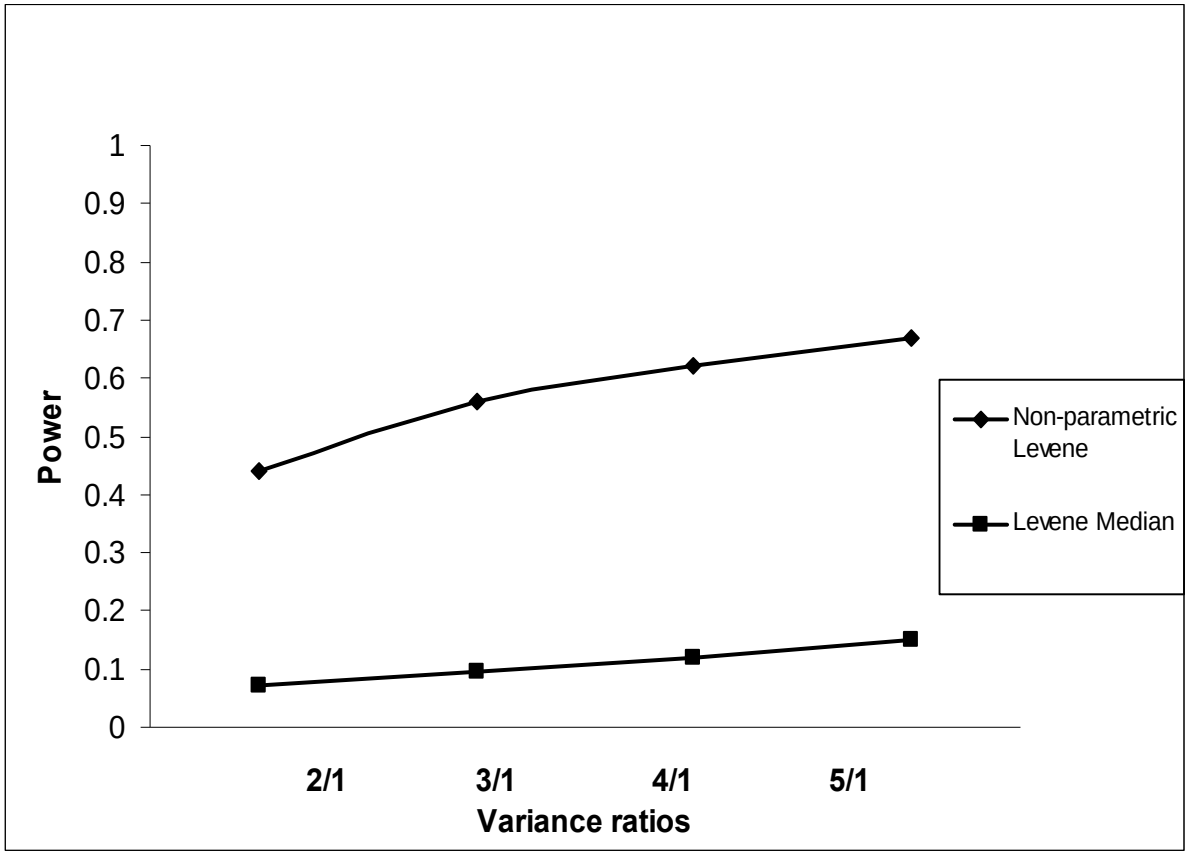


Figure 3.12 Power comparison between the Levene median and the non-parametric Levene test across simulated variance ratios when population distributions have skew=3.



3.5 References

Bridge, P.D. & 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 Epidemiology*, *52*(*3*), 229-235.

Brown, M.B. & Forsythe, A.B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*(2), 364-367.

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

Conover, W.J., Johnson, M.E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, *23*(*4*), 351- 361.

Keyes, T. M., & Levy, M. S. (1997). Analysis of Levene's test under design imbalance. *Journal of Educational and Behavioral Statistics, 22,* 227-236.

Kruskal, W.H. & Wallis, W.A. (1952). Use of rands in one-criterion variance analysis. *Journal of the American Statistical Association*, *260*(*47*), 583-621.

Nordstokke, D.W. & Zumbo, B.D. (2007). A cautionary tale about Levene's tests for equality of variances. *Journal of Educational Research and Policy Studies*, *7*(*1*), 1-14.

O'Brien, R. G. (1978). Robust Techniques for Testing Heterogeneity of Variance Effects in Factorial Designs. *Psychometrika, 43*, 327-344.

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.

Tomarken, A.J. & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin, 99*(*1*), 90-99.

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. (2004). A note on preliminary test of equality of variances. *British Journal of Mathematical and Statistical Society*, *57*, 173-181.

4 CONCLUDING CHAPTER

4.1 Review of the purpose of the dissertation

The purpose of this dissertation was to investigate the Type I error rates and statistical power of four tests for equality of variances between two groups. This was centered around the idea that, in practice, many researchers and data analysts use a conditional test of equal variances to inform them of whether or not to use a pooled variances (Student's) t-test or a separate variances (Welch's) t-test to test for differences in means between two groups. It is important to understand how statistical tests perform under a variety of situations because it helps to inform daily practice.

The central message of the first manuscript (Chapter 2) is that when considering tests of equal variances one needs to be cautious about what is being referred to as "Levene's test" by textbooks writers and statistical programmers, and that it refers to a statistical approach and represents a family of techniques. Depending on which Levene test that is being used, one may be using a statistical test that is as bad, if not worse, that the F-test, which it was intended to replace. In this manuscript (Chapter 2), the performance of the commonly used mean version of the Levene test was investigated under a variety of conditions to test its performance when the assumption of normality has been violated. The performance of the F-test for equal variances was included as a persuasive example of the poor performance of the mean version of the Levene test.

The purpose of this manuscript (Chapter 2) was to not only inform data analysts and researchers about the performance of the mean version of the Levene test, but also to make them aware of the fact that the tests provided by statistical software packages as default tests may not be the best version of the test that is available. It was also

illustrated how statistical software packages manipulate statistical practice and, in combination with their advocation by statistical textbook writers, are essentially shaping the nature of scientific inquiry.

In the second manuscript (Chapter 3) of this dissertation, the performance of a new rank based non-parametric test for equal variances was investigated and compared it to the median version of the Levene test, which is considered the gold standard by many statisticians. The primary goals of this manuscript (Chapter 3) were to (a) introduce a new rank based non-parametric test for homogeneity of variances, and (b) investigate the Type I error rate and power of the new rank based test and compare it to that of the median version of the Levene test. To date, there has been a limited number of simulation studies that studied the performance of the median version of the Levene test, which was another motivation for Paper two (Chapter 3).

The main contribution of this work is that the results found in the first paper (Chapter 2) informs data analysts and researchers about a problem that is occurring in current everyday statistical practice and cautions them to beware of thoughtlessly applying statistical tests. The second paper (Chapter 3) builds on the information that was gained in the first paper (Chapter 2) and provides a possible solution to the problem of using the mean version of the Levene test by default. Together, these two papers highlight a problem occurring in practice and make an attempt to rectify the problem through a new approach to testing for equal variance. It is obvious that the solution will not result in an instant change in statistical practices but, when awareness of the results of this dissertation become known to statistical software programmers, textbook writers, and data analysts, there is great potential for widespread change.

4.2 Review of dissertation results

Manuscript one (Chapter 2) demonstrated that the mean version of the Levene test for equal variances performed poorly, in terms of Type I errors, when the assumption of normality is violated, in particular, when population distributions are skewed. As population distributions became increasingly skewed, the Type I error rates increased sharply to the point where about 25% of the results were statistically significant, thus rejecting the null hypothesis of no significant differences when there were no differences present. When the normality assumption was not violated, the F-test outperformed the mean version of the Levene test for equal variances; this is because under the normal distribution the F-test is the uniformly most powerful test. The problem with the F-test is that it is the most powerful when the conditions are optimal (i.e., normal distribution, equal group sizes), so when one or several of these assumptions are violated, it performs poorly.

It was demonstrated that, when sample sizes were small (N = 24), the mean version of the Levene test was more robust, in terms of Type I errors, to unequal sample sizes. It should be noted that when the mean version of the Levene had lower Type I error rates comparatively to the F test, it generally occurred during times of great inflation of the nominal alpha. Nonetheless, in general when sample sizes were unequal the Levene mean test outperformed the F-test.

Manuscript one concludes with several recommendations, one being that researchers and data analysts avoid using the F test or the mean version of the Levene test for equal variances when there is evidence that the population distribution is not normal.

The second paper (Chapter 3) investigated the Type I error rates and statistical power of the median version of the Levene test and a new rank based test for equal variances. Both tests maintained their nominal Type I error rates across every condition simulated, which then allowed for power comparisons across every condition simulated. Results also revealed that the median version of the Levene test has higher power than the non-parametric test under normality; however, as the population distributions become more skewed, the non-parametric test had higher power.

In all, the first manuscript revealed that currently in textbooks and in common statistical software packages, the mean version of the Levene test is being used in an inappropriate manner. Including it as a default test, as in SPSS, when an independent samples t-test is carried out is misguided because, as demonstrated in the first manuscript, when population distributions are skewed, the Type I error rates increase to unacceptable levels.

The second manuscript demonstrates that the direction of the pairing between the sample size ratio and the ratio of variances has an effect on the Type I error rates and statistical power of tests for equal variances. When the pairing between the ratio of variances and the ratio of sample sizes was inverse (i.e., the larger sample size with the smaller variance and the smaller sample size with the larger variance) the median version of the Levene test performed more efficiently than when the pairing was direct (i.e., the larger sample size with larger variance and smaller sample size with smaller variance). The non-parametric test performed better when the pairing was direct, but was not as affected as the median test by the direction of the pairings.

Manuscript two (Chapter 3) investigates a possible approach for dealing with skewed data in an easily implemented manner. The rank transformation used in conjunction with the Levene approach to testing for equal variances provides an opportunity for data analysts and researchers to properly analyze a wider variety of population distributions than just the Gaussian distribution. This is illustrated in the examples provided in the second manuscript (Chapter 3) that demonstrates to the reader that, when data are severely skewed, the rank transformation is the preferred method for testing for equal variances.

One interesting result found in paper two (Chapter 3) was that, even though both the median version of the Levene test and the rank based Levene test maintained their Type I error rates across all of the conditions, the power of the median test decreased as the distributions became more skewed whereas the power of the rank based test increased as distributions became more skewed. The reason for this seems uncertain. Perhaps it was due to the sampling methods used to generate the data or perhaps it is a naturally occurring property of the rank transformation. Nonetheless, in the distributions investigated in this dissertation, the rank transformation becomes more powerful as data becomes more skewed.

4.3 Implications for statistical practices

The findings of this dissertation have the capacity to have a substantial effect on day-to-day statistical practices. As mentioned in the introduction, when researchers test for equality of variances, it is usually as a preliminary test to aid researchers in the decision of whether to use the pooled variances or the separate variances version of the t-test. In an ideal situation, any statistical test that maintains its nominal Type I error rate with high

statistical power is sought. Anyone using a statistical test should strive to maintain their nominal Type I error rate because, when Type I error rates are inflated, the information that is carried in the result is false and misleading.

Building on the previous section, currently in research practice it is standard practice to use the mean version of the Levene test to determine whether variances are equal. As mentioned earlier in this dissertation, the mean version of the Levene test was initially implemented to replace the F-test for equal variances when data come from population distributions that depart from normality and its use is widely supported in textbooks. The first paper of this dissertation illustrates that, when population distributions are skewed, the mean version of the Levene test performs as bad as or worse than the F-test for equal variances. It is obvious that continuing down this path of using inappropriate statistical tests will lead to many mis-represented results, thus making the process of analyzing group differences very difficult. This puts the process of analyzing group differences into an awkward position under current statistical methods because it is impossible to determine whether differences are true or statistical artifacts representing inflation of Type I error rates. A more pragmatic approach must be utilized to avoid this continuing in day-to-day research practice.

Based on the results of the second paper of this dissertation, it is evident that the non-parametric Levene test is appropriate for use when samples come from skewed population distributions. However, this is not to suggest that the non-parametric test is the uniformly most powerful under all population distribution. Rather, it is to say that there is a need for a test of equal variances when population distributions are skewed, which is the non-parametric Levene test. In practice, researchers testing for group

differences where random assignment to groups is not possible (e.g., gender) must gather as much information about the nature of the population distribution of the dependent variable prior to testing for equality of variances. This can be achieved through prior knowledge (theoretical information) of the dependent variable. With this information, the proper choice of a test for equal variances can be selected based on knowledge of the population distribution. For example, take a very difficult section of an achievement test, that is known by the test makers and invigilators that most of the test takers get only a few of the items correct and occasionally a respondent will get all of the items correct. This results in a skewed population distribution. If the researcher knows the general shape of the population distribution, they can choose the appropriate test for equality of variances; in this case, it would be the non-parametric version of the Levene test.

Quite often researchers do not know the true nature of the population distributions, especially when there is little work done in the area. Thus, they must rely on other methods for determining whether or not the variances are equal between the groups. Empirical frequency distributions can be used to attain a general sense of the shape of the population distribution. An important assumption related to the use of empirical frequency distributions is that they are measured with some sense of accuracy. If there is a great deal of measurement error affecting the shape of the empirical distribution, then the frequency plots will not be useful for estimating the shape of the parent distribution.

By using all of the information at hand, a researcher can select the most appropriate test. For example, if a researcher has evidence that two comparison groups are normally distributed (i.e., via normally distributed empirical frequency plots), then the best test to determine whether variances are equal is the F-test because it is uniformly most powerful

across the normal distribution. If there is evidence that the two groups are sampled from heavily skewed population distributions (i.e., via heavily skewed empirical frequency plots), then the best test to use is the non-parametric version of the Levene test. It is important here to strongly express the point that it is up to the researcher to gather evidence about what they are studying and not expect there to be an "all purpose" statistical test that is robust across all possible conditions. This is where many of the statistical software packages fall short. In many cases (especially if you are using SPSS), a specific tests for equal variances is included as a default test, and there is only one option, that is selected by the computer program, that automatically provides a p-value to users whenever a t-test is conducted. The limitations of this approach are obvious based on the results of this dissertation. As a general rule, it can be stated that there is not any uniformly most powerful test for equal variances across all distributional forms. Therefore, a default test where it is the only test in statistical software packages is inappropriate for the purpose of testing for equal variances.

The simulation method used in this dissertation reflected that of an experimental design where factors (e.g., skew) are manipulated across conditions, as opposed to simulations that investigate very specific research parameters that are found in practice. This enables a descriptive approach for interpreting results. The strength of the experimental design used in this dissertation is that the results can be extended to other conditions. For example, in the parameters used in the simulations of this dissertation, a skew of 3 was the maximum amount that was investigated; however, based on the results, it can be assumed that skew increased, the directionality of the result would continue. The mean version of the Levene test had increased Type I error rates as the skew

increased and this trend is likely continued as skew increases beyond 3. In addition, the experimental nature of the simulations allow for prediction of what Type I error rate will be between the different values investigated. For example, one could approximate the Type I error rate of a test when the skew is somewhere between 2 and 3.

The nature and breadth of this experimental design strengthens the internal validity of this study. By manipulating factors such a skew across several conditions, it can be demonstrated that increased skew causes the Type I error rates of the mean version of Levene's test to increase, whereas the non-parametric version of the Levene test is robust to increasing skew. This allows for the conclusion to be made that, in some tests (i.e., the mean version of the Levene test), increased skew causes increased Type I error rates. The external validity of this study may be not as strong because of the experimental design where a broad range of conditions are investigated without a specific research context in mind thus allowing generalizability across a wide array of conditions, but without any specific research context in mind. This is not a considerable limitation, but instead speaks to the broadness of the design and gives a general sense of the operating characteristics of the tests that were investigated. Further studies could simulate more specific research contexts for the utility of these tests for equal variance.

4.4 Future directions in this program of research.

The next section of this dissertation will discuss some possible limitations that are present in this dissertation and use them to stage some directions for future work.

4.4.1 Discussion of the distributions that were used in simulations

One limitation of the current dissertation is the limited number of distributions that were investigated. The distributions simulated in papers one and two (Chapters 2 & 3)

were based on a Chi-squares distribution with the appropriate degrees of freedom to create the required distributional form. This technique was used to create continuous distributions upon which the statistical tests were performed. The results of this simulation are not generalizable beyond the distributions that were tested; however, the results do provide evidence that the non-parametric test for equal variances does seem to perform well in non-normal data situations as well as guidance for further studies.

Future work will investigate a wider variety of distributions. Micceri (1989) investigated the distributional characteristics of 440 real-world large-sample achievement and psychometric measures. Within these samples, he found several classes of contamination that can occur, including tail weights from the uniform to the double exponential, exponential-level asymmetry, severe digit preferences, multimodalities, and modes external to the mean/median interval. This demonstrates that, in real-world research, there are multiplicities of possible distributions with characteristics that deviate from normality in several ways. For example, the tails on the ends of a distribution can stray from normality by either being very thick at each end of the distribution representing a large number of values lie near the tails, or very thin, representing very few scores in the tails. Micceri's work can be thought of as a platform that can be used to launch a number of simulation studies to investigate the performance of the rank transformation for testing equality of variances under a wider variety of distributional forms.

4.4.2 Methodological considerations

A second potential limitation of this dissertation is that the only data that were investigated were generated through Monte-Carlo simulation methods. This is not

necessarily a limitation, but distributions that are generated by using Monte-Carlo simulation methods do not necessarily encompass the multi-influential nature of real-world data. To unpack this a bit further, real-world data is rich with variability that is influenced by many different and usually unknown reasons. When data are simulated, all of the parameters are specified by the researcher, thus limiting some of the generalizability of the results of simulation studies. However, this type of thinking is somewhat short-sighted because simulation results provide evidence related to the performance of a test under ideal (simulated) conditions where the researcher knows apriori, the properties of the distribution. From simulation results, tests on real-world data with can commence in an informed manner. With this in mind, further research will investigate the performance of the rank transformation under real-world conditions.

4.4.3 Number of groups investigated in simulations and more complex designs

A third limiting factor perhaps hampering generalizability is that all simulations in this study were conducted using only two groups for comparison. Future work will extend this to more than two groups and in situations where there are more complex factorial designs. As mentioned in the introductory chapter, Sawilowsky (1990) demonstrated that the rank transformation which, is the basis of the non-parametric Levene test, tends to break down when factorial designs become more complex than a 2X2 design. This does not mean that more complex designs should not be investigated but, instead, future work will use the 2X2 factorial design as a base model for simulation and attempts will be made via simulation to extend the utility of the rank transformation for use with more complex models. This work will not only provide information to data analysts and researchers about the usefulness of the rank transformation, but it will also

add to the current knowledge of the limits of the rank transformation in more complex designs with more than two groups.

In addition, the finding that, the power of the non-parametric Levene test increases when data are sampled from skewed population distributions requires further investigation to extend the current findings. Future work will generate data using multiple methods to ensure that the increase in power that occurs as skew increases is not merely a statistical artifact. Also, as mentioned earlier in the discussion, future work will simulate a larger variety of population distributions using multiple methods to provide stronger evidence that the results from the simulation studies are valid for the conditions being studied.

4.4.4 Ranking with ties

One problem not addressed in this dissertation is the issue of how ties are dealt with when ranking a set of values. When there are equivalent scores in a set of values, the method of dealing with these values may have an effect on the outcome of the analysis. Kendall (1945) suggests that allotting ties when ranking the mean rank. For example, if the third and the fourth members each is given a score of 3.5. By doing this, the average rank is assigned to the two scores to correct for the tie. This approach may be limited by the way the ties are distributed within and between the groups.

Future work in this area will investigate the nature of how the ties occur and their impact on the results of statistical tests. For example, what if all the ties are in one group? How does this have an effect on the results when using the rank transformation? In this vein, ties could be considered a source of contamination and the number of tied scores in the total set of values could be considered the 'amount' of contamination by

ties. This would enable statistical researchers to investigate a gradient of contamination starting from no ties up to the point where every score was tied if they wish to do so. By doing this, the effects of ties can be determined and researchers can proceed, using data with tied scores in an informed manner.

Another limitation related to the rank transformation, but not necessarily a limitation of this dissertation, is that the nature of the true metric (population values) of the dependent variable can potentially have a large impact on the results of statistical tests. For example, imagine a five point Likert response scale on a suicide checklist asking respondents about the nature of their feelings. The typical response for most respondents is the first choice in the response scale; however, the rest of the choices in the response scale must be made available to respondents because they are valid possibilities. The resulting population distribution will be highly kurtotic, which will lead to most of the scores being the same when rank transformed. Data that have more of a continuous nature will tend to have fewer ties than Likert type data, even in highly kurtotic distributions simply due to the fact that there are more overall possible values that data can be represented as. Future work will simulate a wide variety of population metrics to investigate the effects of true metric on the rank transformation.

4.4.5 Conditional tests

An overarching issue that encompasses this entire dissertation is the use of conditional test of equality of variances prior to selecting the pooled or the un-pooled sample t-test. Hayes and Cai (2007) note that, the conditional decision rule is quite often advocated by textbooks and statistical software packages, and is usually carried out using a single, usually arbitrarily selected, variance equality test. They point out that, in the

majority of the conditions investigated in their simulation study the unconditional tests (i.e., unpooled samples t-test) outperforms the conditional test and, should be the preferred method instead of using a conditional test of equal variances. In a simulation study, Zimmerman (2004) showed that using a conditional test of variances (i.e., mean version of the Levene test) to compare groups that have unequal group sizes and variances prior to choosing whether to use the pooled or un-pooled variances t-test, usually does not provide and protection from elevated Type I error rates and usually makes matters worse.

This is not to suggest that we abandon the use of conditional tests to select whether to use the pooled or un-pooled variances t-test, but instead it is to state that, current practice has perhaps come to the point where the use of simple statistical tests occurs without thought. One possible solution, is to inform researchers by carrying out a series of simulation studies that investigate the utility of using conditional tests of equal variances, including the new rank based test for equal variances that was introduced in Manuscript one (Chapter 2) and investigated in Manuscript two (Chapter 3). Future work stemming from this dissertation will investigate how conditional tests perform under various combinations of violations of its assumptions. Initial work will involve investigating conditional tests that are currently provided in widely used statistical software (i.e., SPSS). Subsequent studies will extend the findings of this dissertation into initial inquiries on conditional tests to investigate the usefulness of other tests for equal variances as preliminary tests for selecting the appropriate t-test. This work will help guide data analysts and researchers in their daily practice by providing information about the most appropriate test given the nature of the data (e.g., distributional form).

4.5 Summary

The purpose of this dissertation was to investigate the current statistical practice of testing for equal variances prior to conducting a t-test through conducting a series of monte carlo simulations (Manuscript 1), and to provide data analysts and researchers with an easily implemented method for testing for equality of variances (Manuscript 2) that they could employ in their daily practice. This dissertation highlights an important issue that can occur in our research practices that, when we are conducting statistical tests, we take the underlying algorithms provided by statistical software packages for granted and, utilize them without much thought of the underlying assumptions. This dissertation also highlights the persuasive nature of simulation work and, its potential usefulness as a teaching tool. The remainder of this dissertation will discuss the potential impact of findings of Manuscripts one and two on current research practice as well as some more global topics related to this dissertation.

4.5.1 Potential Impact of Dissertation Results on Current Statistical Practices

The two papers in this dissertation have great potential to have a large impact on daily research practice. The first Manuscript uses simulation as a persuasive tool to inform data analyst and researchers of the pitfalls that are occurring as they conduct their daily research, pertaining to the utilization of statistical software. The deceptively easy to use interface of widely used statistical software packages (i.e., SPSS) provides solutions to users without the mention of the assumptions. Manuscript one (Chapter 2) highlights the fact that, statistical software producers may not have statistical correctness in mind while they are writing the algorithms for statistical applications and, that data analysts and researchers must be responsible for knowing the algorithms that are being

implemented by the statistical software that they are using. Manuscript one also provides empirical evidence of previous mathematical results (i.e., Carroll & Schneider, 1984) to those data analysts and researchers who are not mathematically inclined.

Manuscript two (Chapter 3) uses simulation to investigate the performance (i.e., Type I errors and statistical power) of the median version of the Levene test and, a new rank based test to test for equality of variances. This paper has the potential to become an influential paper in the statistical literature, as well as have an impact on daily statistical practices. To date, there have been a limited number of studies that have investigated the performance of the median version of the Levene test (Brown & Forsythe, 1974; Conover, Johnson & Johnson, 1984). Thus, paper two expanded upon the current knowledge base of the performance this version of the Levene test. More importantly, paper two investigated the performance of a new rank based test for equality of variances and, demonstrated that this new test outperforms the median based Levene test in many of the conditions simulated, especially when the distributions are sampled from skewed populations. The impact of the results of paper two have the potential to be influential in the statistical literature. It is obvious that, more work must be conducted to extend the current findings of the performance of the new non-parametric test under a variety of conditions. Nonetheless, the results of paper two are quite convincing that this test may provide an alternative for testing for equal variances when there is an abundance of skew present in population distributions.

4.5.2 Simulation in Teaching

The empirical nature of simulation makes it an obvious choice as a teaching tool. Many students of statistics comment on the difficulty that they are having with the

abstract nature of many statistical concepts covered in textbooks. What follows, is an example of a simulation teaching module that could be delivered to students in an introductory statistics course when teaching them about the t-test. The central concept that will be discussed is that, Type I error rates are related to the violation of the assumptions of the test. It will be demonstrated, through using simulation, that, as the form of the population distribution shifts from being normally distributed to a more skewed population distribution the Type I error rate of the Student's t-test becomes inflated.

The teaching module includes (1) the data generation phase, (2) the visualization of the distribution, and (3) the analytic procedure. In the data generation phase students are introduced to the concept of frequency distributions and how different types of data will result in a variety of distributional forms. They will be introduced to the concept of data generation and provided with instruction on how to generate data. Once students are comfortable with the concepts and procedures, they can begin to investigate different distributional forms. It will be made clear to students that, the normal distribution is one of many distributional types that they may encounter in their daily research activities.

The second phase of the teaching module builds on the first phase, and is integrally related to it, the visualization of the distributions. Students will be taught how to illustrate empirical frequency distributions by using histograms and box-plots. These two graphical techniques are particularly useful for identifying the form of the frequency distribution and values that may increase skew. Through visualizing distributions, students can begin to appreciate the impact of the shape of the distribution on the range of scores that are the result of sampling from any given population distribution.

By this point in the module students should be comfortable with generating and visualizing various distributional forms and, will be poised to test the performance of statistical tests. The third phase of the teaching module takes the various data sets that were simulated by students and has them perform the Student's t-test. Data will have been generated with equal means (i.e., set to zero) so that a significant result represents a Type I error. The first step of phase three will be to generate data for two groups that are sampled from normally distributed population distribution, and then run the t-test on the data for several replications. The student will check to see how many Type I errors there were and whether it is close to the nominal alpha level of .05. Next, students will generate a data set that departs from normality in some way (e.g., increased skew) and then, run the t-test again and see whether the Type I error rate has increased, decreased, or remained the same. The student can continue to investigate how various distributional forms has an effect on the Type I error rates of the Student's t-test in this case, as well as other statistical tests. Simulation provides a powerful method for teaching students the impact of various distributional forms on the performance of statistical tests.

4.5.3 Accessibility to Researchers

Central to the dissemination of the findings of this dissertation to current research practice is the issue of accessibility. Data analysts and researchers must be made aware of the findings in Manuscript one (Chapter 2) that, the widely recommended Levene mean test is sensitive to violations of distributional assumptions and should be interpreted with caution. In addition, the findings from Manuscript two (Chapter 3) that, the new non-parametric test for equal variances performs well when data come from skewed population distributions must be made available to data analysts and researchers. This

could possibly happen through a web based application that data analysts and researchers can access and enter their data into for testing. Through a free easily accessible web resource data analysts and researchers could test for equal variances on their data.

4.5.4 Default Tests in Statistical Software

In current research practice, the advent of easy to use software that carries out all of our statistical analyses is both a blessing and a bane to scientific inquiry. Many statistical techniques require laborious calculations (e.g., maximum likelihood estimation) and, statistical software packages provide an opportunity for data analysts and researchers to use these techniques. Also, statistical software offers a wide variety of statistical tests to choose from; however, this flexibility does not come without a price. Some tests that are implemented in statistical software packages are provided in a default manner and, are often the only test made available for use. This is generally without any statement to users of the software about the assumptions of the test. Most statistical software packages can perform a simple t-test in the blink of an eye through a drop down window interface, allowing data analysts and researchers the opportunity to run a t-test without any thought of the consequences of violating the basic assumptions of the test.

Perhaps, the future of statistical software packages is that they become more flexible offer fewer default tests and, with more choices for data analysts and researchers to choose from. As well, provide a warning when the assumptions of tests are violated, or at least provide an easily accessible list of the assumption for users of the software. This may reduce the number of statistically illiterate software users permeating scientific knowledge with false reports; however, it could make it worse because there would just be more boxes for users to arbitrarily click, leading to further complications. Obviously

there is no easy answer but, in current statistical practice the idea of limiting the users' choices of statistical tests by providing default tests is problematic. This is because it narrows the breadth of possible analyses and should be addressed by software developers.

It is difficult, in many cases, to deliver the results of statistically based research to daily users of software because of the barriers in knowledge between the disciplines that conduct 'real-world' research and, those who work in the areas of mathematics and statistics. This is where writers of statistical textbooks and software packages are supposed to come in and provide assistance in disseminating the finding from analytic mathematics and statistical work to data analysts and researchers. The problems is, as it was demonstrated in Manuscript one, that, many of the textbooks consulted suggested using the mean version of the Levene test. This test has been known to be problematic for decades. Also, the developers of statistical software packages seem to be oblivious to the decades old result that demonstrated the susceptibility of the mean version of the Levene test to extremely exaggerated Type I error rates when the assumption of normality has been violated. Thus, this dissertation has revealed a very serious problem that is occurring in the research world, and those who should be responsible for making the findings of mathematical and statistical work accessible to researchers, are not. This leaves one with the feeling that, in the day-to-day world of research, findings that are based on statistical tests are questionable. The solution is relatively simple, but convincing textbook writer and software developer of this is quite another issue.

4.6 Concluding Remarks

The purpose of this dissertation was to investigate and demonstrate the statistical performance, based on Type I error rates and statistical power, of several tests for

equality of variances when data are sampled from population distributions that range from being normally distributed to being severely skewed. The usefulness of the non-parametric Levene in situations where samples come from moderate to highly skewed population distributions was demonstrated. Embedded within this dissertation were several other messages related to the usefulness of simulation for research and teaching, and the dissemination of analytical mathematics into practical applications through textbooks and statistical software packages.

4.7 References

Brown, M.B. & Forsythe, A.B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*(2), 364-367.

Carroll, R.J. & Schneider, H. (1985). A note on Levene's tests for equality of variances. *Statistics and Probability Letters*, *3*, 191-194.

Conover, W.J., Johnson, M.E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, *23*(*4*), 351- 361.

Hayes, A.F. & Cai, L. (2007). Further evaluating the conditional decision rule for comparing two independent means. *British Journal of Mathematical and Statistical Psychology*, *60*, 217-244.

Kendall, M.G. (1945). The treatment of ties in ranking problems. *Biometrika*, *33*(*3*), 239-251.

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

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

Zimmerman, D.W. & Zumbo, B.D. (1992). Parametric alternatives to the student t test under violations of normality and homogeneity of variance. *Perceptual and Motor Skills*, *74*, 835-844.