Psicológica (2010), 31, 401-430.

A New Nonparametric Levene Test for Equal Variances

David W. Nordstokke[*] (1) & Bruno D. Zumbo (2)

(1) University of Calgary, Canada
(2) University of British Columbia, Canada

Tests of the equality of variances are sometimes used on their own to compare variability across groups of experimental or non-experimental conditions but they are most often used alongside other methods to support assumptions made about variances. A new nonparametric test of equality of variances is described and compared to current 'gold standard' method, the median-based Levene test, in a computer simulation study. The simulation results show that when sampling from either symmetric or skewed population distributions both the median based and nonparametric Levene tests maintain their nominal Type I error rate; however, when one is sampling from skewed population distributions the nonparametric test has more statistical power.

Most studies in education, psychology, and the psycho-social and health sciences more broadly, use statistical hypothesis tests, such as the independent samples t-test or analysis of variance, to test the equality of two or more means, or other measures of location. In addition, some, but far fewer, studies compare variability across groups or experimental or non-experimental conditions. Tests of the equality of variances can therefore be used on their own for this purpose but they are most often used alongside other methods to support assumptions made about variances. This is often done so that variances can be pooled across groups to yield an estimate of variance that is used in the standard error of the statistic in question. In current statistical practice, there is no consensus about which statistical test of variances maintains its nominal Type I error rate and maximizes the statistical power when data are sampled from skewed population distributions.

[*] Address Correspondence to: David W. Nordstokke, Ph.D. Division of Applied Psychology. University of Calgary. 2500 University Drive NW. Education Tower 302. Calgary, Alberta, Canada T2N 1N4. Email: dnordsto@ucalgary.ca

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

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

wherein, a two-tailed test of the null hypothesis (H_0) that the variances are equal against the alternative hypothesis (H_1) that the variances are not equal is performed. Nordstokke and Zumbo (2007) recently investigated the widely recommended parametric mean based Levene test for testing equal variances and, like several papers before them (e.g., Carroll and Schneider, 1985; Shoemaker, 2003; Zimmerman, 2004), they highlighted that the Levene test is a family of techniques, and that the original mean version is not robust to skewness of the population distribution of scores. As a way of highlighting this latter point, Nordstokke and Zumbo showed that if one is using the original variation of Levene's test, a mean-based test, such as that found in widely used statistical software packages like SPSS and widely recommended in textbooks, one may be doing as poorly (or worse) than the notorious F test of equal variances, which the original Levene test (1960) was intended to replace.

As a reminder, the mean version of the Levene test (1960) is

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

wherein, equation (T1) shows that this test is a one-way analysis of variance conducted on the absolute deviation value, which is calculated by subtracting from each individual's score, denoted X_{ij}, from their group mean value, denoted $\overline{X}_j$, for each individual i in group j. The family of Levene tests can be applied to more than two independent groups but, without loss of generality, the current study focuses on the two-group case.

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 a new nonparametric Levene test (described in detail below) 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.

As Conover and his colleagues (1981) showed, 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

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

wherein, building on our notation above, the analysis of variance is conducted on the absolute deviations of individual's score, denoted X_{ij}, from their group median value, denoted Mdn_j, for each individual i in group j. The median based version has been shown to perform well in situations wherein data were skewed. This test is available in widely used statistical software programs such as SAS (using, for example, PROC GLM and MEANS sample / HOVTEST=BF) and R using the packages "lawstat" (Hui, Gel, & Gastwirth, 2008) or "car" (Fox, 2009). 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. In terms of skewed distributions, Browne and Forsythe investigated the Chi-square distribution with 4 degrees of freedom (skew equals approximately 1.5) and showed that 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 range 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 newly developed nonparametric Levene test.

As Nordstokke and Zumbo (2007) describe it, the nonparametric Levene test involves pooling the data from all groups, ranking the scores allowing, if necessary, for ties, placing the rank values back into their original groups, and running the Levene test on the ranks. Of course, if one were using SPSS (or some other program wherein the means version of the Levene test is computed) then one would merely have to apply the rank transformation and then submit the resulting ranks to the means Levene test

and one would have the nonparametric test, as described in Nordstokke and Zumbo (2007, pp. 11-12). Using our earlier notation, the nonparametric Levene test can be written as

$$\text{ANOVA}\left(\left|R_{ij} - \overline{X}_j^R\right|\right), \qquad \text{(T3)}$$

wherein a one-way analysis of variance is conducted on the absolute value of the mean of the ranks for each group, denoted $\bar{X}_j^R$, subtracted from each individual's rank R_{ij}, for individual i in group j. This nonparametric Levene test is based on the principle of the rank transformation (Conover & Iman, 1981). When the data are extremely non-normal, perhaps caused by several outliers, or the variable is genuinely non-normal (e.g., salary), or some other intervening variables, the transformation changes the distribution and makes it uniform. Conover and Iman suggested conducting parametric analyses, for example, the analysis of variance, on rank transformed data. The use of rank transformed data, although popularized by Conover and Iman, is an idea that has had currency in the field of statistics for many years as a way to avoid the assumption of normality in analysis of variance (see, for example, Friedman, 1937; 1940). Thus the nonparametric Levene test is a parametric Levene test on the rank transformed data.

It should be noted that the null hypothesis for both the median and nonparametric 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 populations are identically distributed in shape (not necessarily in location), and the alternate hypothesis is that they are not identically distributed in shape. If two or more distributions are identically distributed in shape, then it is implied that the variances are equal. That is, if the researcher can assume identical distributions, then they can assume homogeneity of variances. Thus the overlap between the hypotheses of parametric and nonparametric 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 nonparametric 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 nonparametric

Levene test would lead one to conclude that the variances are not homogeneous.

METHOD

Data Generation

A computer simulation was performed following standard simulation methodology (e.g., Nordstokke & Zumbo, 2007; Zimmerman, 1987; 2004). Population distributions were generated and the statistical tests were performed using the statistical software package for the social sciences, SPSS. A pseudo random number sampling method with the initial seed selected randomly was used to produce χ^2 distributions. An example of the syntax used to create the population distribution of one group belonging to a normal distribution is included in Appendix 1. 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 proportion of rejections of the null hypothesis in each cell of the design and, more specifically, 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 only investigated statistical power in those conditions wherein the nominal Type I error rate, in our study .05, is maintained.

Shape of the population distributions[1]

Four levels of skew 0, 1, 2, and 3 were investigated. As is well known, as the degrees of freedom of a χ^2 distribution increase it more closely approximates a normal distribution. The skew of the distributions for both groups were always the same in all replications and are shown in

[1] It should be noted that the population skew was determined empirically for large sample sizes of 120,000 values with 1000, 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. The well known mathematical relation is $\gamma_1 = \sqrt{8/df}$.

Figure 1 (reading from top left to bottom right) for skew values of 0, 1, 2, and 3 respectively.

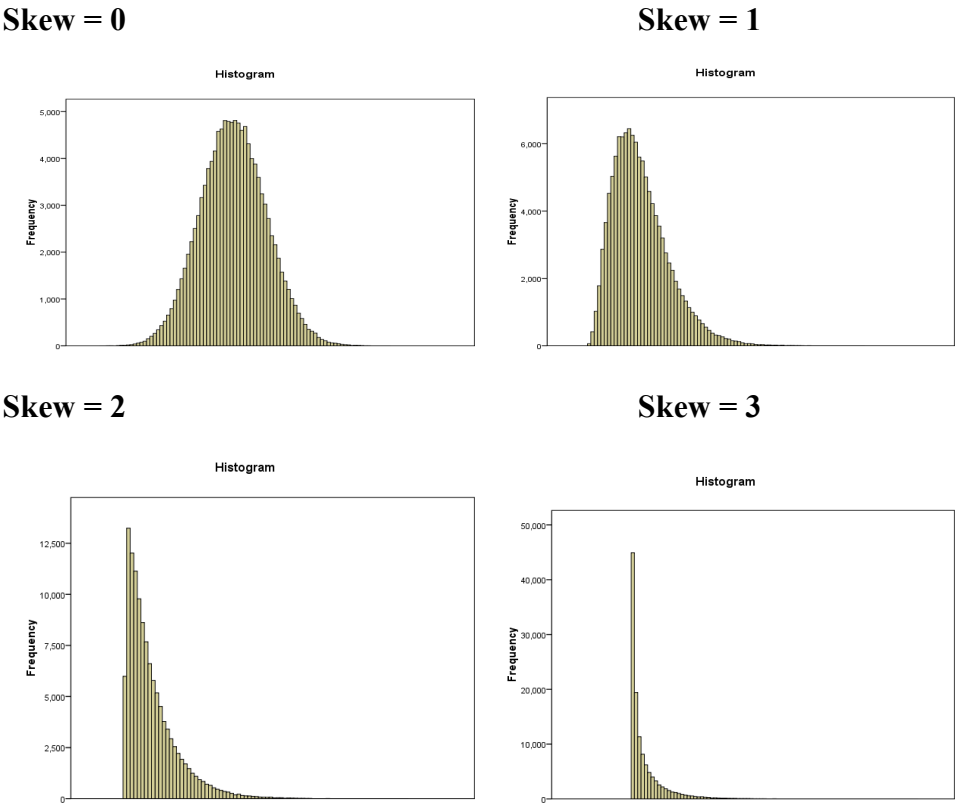

Figure 1. Shape of population distributions used in simulations

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.

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). Variance ratios were manipulated by multiplying the population of one of the groups in the design by a constant 2 (2/1, 1/2 ratios), 3 (3/1, 1/3 ratios), 4 (4/1, 1/4 ratios, and 5 (5/1, 1/5 ratios). The value of the constant was dependent on the amount of variance imbalance that was required for the cell of the design. For example, for a variance ratio of 2/1, the scores would be multiplied by 2. 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.

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.

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 generated and sampled from; 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 on the same set of data. 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

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 (±.01). The value of ±.01 represents moderate robustness and comes from Bradley (1978). The choice of Bradley's criterion is somewhat arbitrary, although it is a middle ground between his alternatives, and some of our conclusions may change with the other criteria. It should be noted that when Type I error rates are less than .05, the validity of the test is not jeopardized to the same extent as they are when they are inflated. This makes a test invalid if Type I errors are inflated, but when they decrease, the test becomes more conservative, reducing power. Reducing power does not invalidate the results of a test, so tests will be considered to be invalid only if the Type I error rate is inflated. 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 proportion of times that the Levene's median test, (T2) and the nonparametric Levene's test (T3), correctly rejected the null hypothesis.

RESULTS

The Type I error rates for the Levene median test (T2) and the nonparametric Levene test (T3) for all of the conditions in the study are illustrated in Table 1. For example, the first row in Table 1 (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 nonparametric Levene test is .049 and the Type I error rate for the Levene median test is .038. 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 nonparametric 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 2 reports the power values of the Levene median test (T2) and the nonparametric Levene tests when the population skew is equal to 0. In nearly all of the cells of the Table 2 the Levene median test (T2) has slightly higher power values. For example, in the first row of the table

are the results for the nonparametric 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; that is, 42 percent 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. In 61 of the 72 cells in Table 2, the median test had higher power than the nonparametric test.

The values for the power of the nonparametric Levene test (T3) and the Levene median test (T2) when the population distributions have a skew equal to 1 are illustrated in Table 3. Again, in most of the cases, the Levene median test (T2) had slightly higher power values than the nonparametric 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 are the power values for the nonparametric Levene (T3). For a sample size of 24 with equal groups and a ratio of variances 5/1, the power value is .474. In comparison, the Levene median test (T2) under identical conditions has a power value of .434. The median test was more powerful than the nonparametric 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 4. In a great number of the cells of the table, the nonparametric Levene (T3) has higher power values than the Levene median test (T2). For example, the power for the nonparametric Levene test (T3) is present in the first row of Table 4. For a sample size of 24 with equal group sizes and a ratio of variances of 5/1, the power value is .572. In comparison, the power of the Levene median test under the same conditions is .296. The nonparametric test was more powerful than the median test in every cell of Table 4.

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 5. The nonparametric Levene test (T3) has notably higher power values than the Levene median test (T2). For example, the first row of Table 5 lists the power values for the nonparametric 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. In comparison, the power of the Levene median test (T2) is .155. The nonparametric test was more powerful than the median in every cell of Table 5.

Table 1. Type I error rates of the Nonparametric and Median versions of the Levene tests.

N	n1/n2	Nonparametric Levene	Levene Median
Skew = 0			
24	**1/1**	.049	.038
24	**2/1**	.050	.037
24	**3/1**	.047	.039
48	**1/1**	.044	.039
48	**2/1**	.053	.043
48	**3/1**	.054	.046
96	**1/1**	.047	.040
96	**2/1**	.051	.043
96	**3/1**	.051	.043
Skew = 1			
24	**1/1**	.043	.040
24	**2/1**	.048	.041
24	**3/1**	.049	.039
48	**1/1**	.046	.041
48	**2/1**	.048	.040
48	**3/1**	.058	.044
96	**1/1**	.050	.052
96	**2/1**	.048	.044
96	**3/1**	.047	.042
Skew = 2			
24	**1/1**	.051	.050
24	**2/1**	.049	.047
24	**3/1**	.054	.050
48	**1/1**	.053	.055
48	**2/1**	.052	.047
48	**3/1**	.053	.045
96	**1/1**	.051	.051
96	**2/1**	.049	.050
96	**3/1**	.054	.048
Skew =3			
24	**1/1**	.049	.053
24	**2/1**	.054	.050
24	**3/1**	.046	.049
48	**1/1**	.049	.045
48	**2/1**	.046	.043
48	**3/1**	.050	.044
96	**1/1**	.045	.044
96	**2/1**	.054	.048
96	**3/1**	.050	.043

Table 2. Power values of the nonparametric 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**			
Nonparametric Levene	24	1/1	.420	.354	.239	.124	.124	.239	.354	.420
Levene Median	24	1/1	.500	.407	.272	.132	.132	.272	.407	.500
Nonparametric Levene	24	2/1	.326	.260	.183	.095	.14	.246	.37	.459
Levene Median	24	2/1	.509	.401	.267	.122	.098	.197	.32	.393
Nonparametric Levene	24	3/1	.256	.206	.145	.073	.130	.222	.314	.402
Levene Median	24	3/1	.466	.351	.244	.110	.088	.139	.214	.285
Nonparametric Levene	48	1/1	.783	.673	.484	.222	.222	.484	.673	.783
Levene Median	48	1/1	.899	.805	.595	.272	.272	.595	.805	.899
Nonparametric Levene	48	2/1	.661	.546	.391	.188	.231	.480	.681	.796
Levene Median	48	2/1	.874	.755	.566	.263	.242	.515	.728	.851
Nonparametric Levene	48	3/1	.535	.448	.315	.156	.211	.443	.631	.746
Levene Median	48	3/1	.805	.698	.508	.237	.193	.425	.625	.755
Nonparametric Levene	96	1/1	.98	.943	.787	.438	.438	.787	.943	.980
Levene Median	96	1/1	.998	.988	.913	.56	.56	.913	.988	.998
Nonparametric Levene	96	2/1	.95	.876	.699	.378	.426	.798	.941	.980
Levene Median	96	2/1	.994	.976	.879	.532	.493	.884	.981	.997
Nonparametric Levene	96	3/1	.875	.779	.607	.293	.374	.739	.911	.970
Levene Median	96	3/1	.984	.951	.833	.439	.401	.815	.957	.989

Table 3. Power values of the nonparametric 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**			
Nonparametric Levene	24	1/1	.474	.385	.278	.142	.142	.278	.385	.474
Levene Median	24	1/1	.434	.340	.229	.120	.120	.229	.340	.434
Nonparametric Levene	24	2/1	.357	.293	.202	.105	.154	.296	.416	.505
Levene Median	24	2/1	.436	.348	.234	.117	.090	.167	.253	.321
Nonparametric Levene	24	3/1	.285	.230	.160	.087	.143	.244	.356	.439
Levene Median	24	3/1	.424	.312	.214	.113	.069	.112	.168	.221
Nonparametric Levene	48	1/1	.836	.730	.566	.276	.276	.566	.730	.836
Levene Median	48	1/1	.820	.705	.515	.241	.241	.424	.705	.820
Nonparametric Levene	48	2/1	.715	.609	.439	.215	.261	.558	.732	.852
Levene Median	48	2/1	.804	.681	.485	.228	.182	.418	.587	.740
Nonparametric Levene	48	3/1	.586	.485	.350	.178	.253	.524	.701	.815
Levene Median	48	3/1	.735	.613	.428	.206	.156	.330	.500	.634
Nonparametric Levene	96	1/1	.991	.966	.878	.522	.522	.878	.966	.991
Levene Median	96	1/1	.991	.965	.852	.466	.466	.852	.965	.991
Nonparametric Levene	96	2/1	.966	.913	.774	.423	.492	.860	.970	.993
Levene Median	96	2/1	.987	.941	.807	.430	.389	.774	.937	.980
Nonparametric Levene	96	3/1	.905	.837	.673	.359	.463	.822	.947	.986
Levene Median	96	3/1	.964	.908	.752	.390	.335	.689	.873	.959

Table 4. Power values of the nonparametric 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**			
Nonparametric Levene	24	1/1	.572	.499	.376	.214	.214	.376	.499	.572
Levene Median	24	1/1	.296	.238	.166	.091	.091	.166	.238	.296
Nonparametric Levene	24	2/1	.431	.364	.275	.151	.222	.405	.517	.612
Levene Median	24	2/1	.342	.266	.195	.101	.067	.118	.143	.193
Nonparametric Levene	24	3/1	.333	.298	.229	.128	.203	.358	.466	.551
Levene Median	24	3/1	.326	.267	.189	.112	.056	.074	.103	.127
Nonparametric Levene	48	1/1	.920	.864	.723	.428	.428	.723	.864	.920
Levene Median	48	1/1	.629	.522	.346	.155	.155	.346	.522	.629
Nonparametric Levene	48	2/1	.795	.723	.602	.351	.436	.746	.884	.940
Levene Median	48	2/1	.622	.514	.353	.181	.122	.268	.390	.495
Nonparametric Levene	48	3/1	.663	.596	.472	.253	.395	.704	.846	.917
Levene Median	48	3/1	.583	.472	.326	.156	.093	.185	.271	.342
Nonparametric Levene	96	1/1	.999	.995	.964	.740	.740	.964	.995	.999
Levene Median	96	1/1	.926	.834	.662	.315	.315	.662	.834	.926
Nonparametric Levene	96	2/1	.988	.972	.907	.648	.741	.965	.996	.999
Levene Median	96	2/1	.909	.813	.624	.308	.248	.538	.749	.864
Nonparametric Levene	96	3/1	.946	.914	.820	.547	.688	.950	.992	.999
Levene Median	96	3/1	.871	.768	.574	.274	.196	.447	.646	.779

Table 5. Power values of the nonparametric 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**			
Nonparametric Levene	24	1/1	.667	.622	.565	.443	.443	.565	.622	.667
Levene Median	24	1/1	.155	.124	.094	.073	.073	.094	.124	.155
Nonparametric Levene	24	2/1	.504	.485	.419	.320	.461	.603	.676	.712
Levene Median	24	2/1	.215	.183	.127	.088	.044	.053	.071	.076
Nonparametric Levene	24	3/1	.454	.405	.333	.255	.411	.568	.627	.695
Levene Median	24	3/1	.221	.187	.144	.096	.031	.032	.034	.043
Nonparametric Levene	48	1/1	.954	.944	.919	.810	.810	.919	.944	.954
Levene Median	48	1/1	.319	.258	.178	.093	.093	.178	.258	.319
Nonparametric Levene	48	2/1	.843	.829	.789	.677	.829	.956	.983	.993
Levene Median	48	2/1	.373	.299	.230	.123	.060	.104	.146	.185
Nonparametric Levene	48	3/1	.735	.697	.671	.525	.792	.937	.977	.990
Levene Median	48	3/1	.376	.290	.210	.118	.040	.055	.082	.105
Nonparametric Levene	96	1/1	.999	.999	.999	.987	.987	.999	.999	.999
Levene Median	96	1/1	.648	.499	.351	.168	.168	.351	.499	.648
Nonparametric Levene	96	2/1	.984	.986	.984	.950	.989	.999	.999	.999
Levene Median	96	2/1	.657	.540	.382	.178	.109	.241	.377	.494
Nonparametric Levene	96	3/1	.946	.947	.938	.867	.984	.999	.999	.999
Levene Median	96	3/1	.615	.511	.352	.186	.081	.174	.262	.339

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 nonparametric 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.

2/1 variance ratio

Figure 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 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 nonparametric test and this is maintained as sample size increases. However, as the skew of the population distribution increases, the nonparametric 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 nonparametric 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 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 nonparametric 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 nonparametric 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 nonparametric 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 nonparametric 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. For example, when skew=3 the nonparametric 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 5 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 nonparametric Levene test's power is .677 when the sample size and variance are inversely related. When they are directly related, the power is .829. In contrast, the median version of the Levene test, under the same conditions, has a power of .123 when sample size and variance are inversely related. When sample size and variance are directed related the power of the median test is .060.

The power differences between the two tests across the three sample sizes and four levels of skew when the ratio of variances is 2 to 1 are shown in Figure 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 relationship is reversed as the level of skew increases. As the sample size and the skew increase, the power difference between the median test and the nonparametric test increases in favor of the nonparametric test. 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.

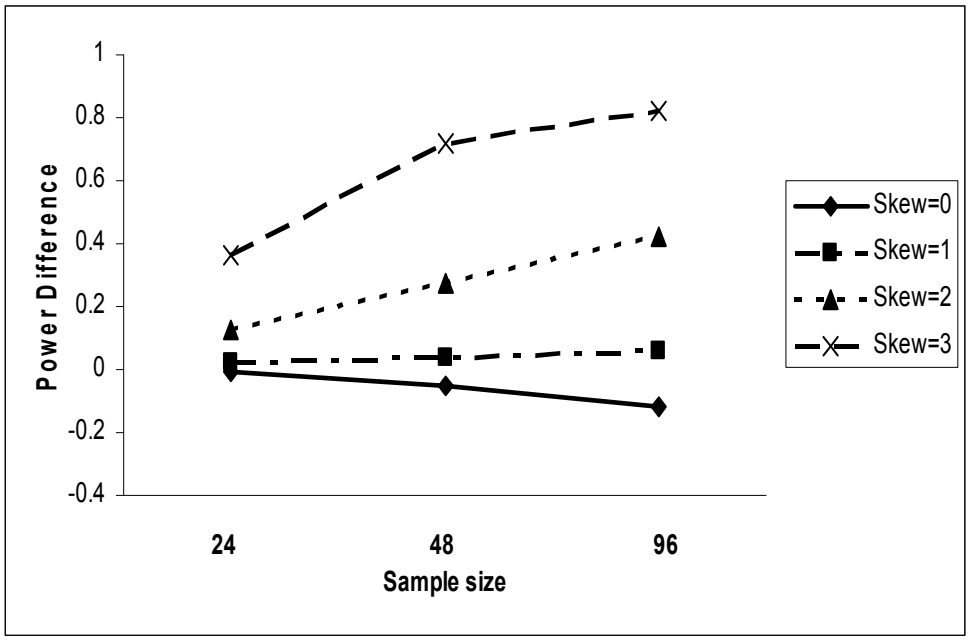

Figure 2. Power difference (2 to 1 variance ratio) for equal sample size ratios. Note: Power differences are based on the nonparametric test minus the median test with a negative value representing superior power for the median test and a positive value representing superior power for the nonparametric test.

4/1 variance ratio

This section of results focuses on the condition where the ratio of variances is 4 to 1. These results 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 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 nonparametric test. This is reasonably stable across the three levels of sample size. As the skew increases, the power advantage begins to favor the nonparametric test and when the skew reaches 3, the power advantage of the nonparametric 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 nonparametric test. It can be

confirmed by looking at Tables 2 and 3 that the nonparametric test is correctly rejecting the null hypothesis nearly 100 percent of the time when the sample size is 96.

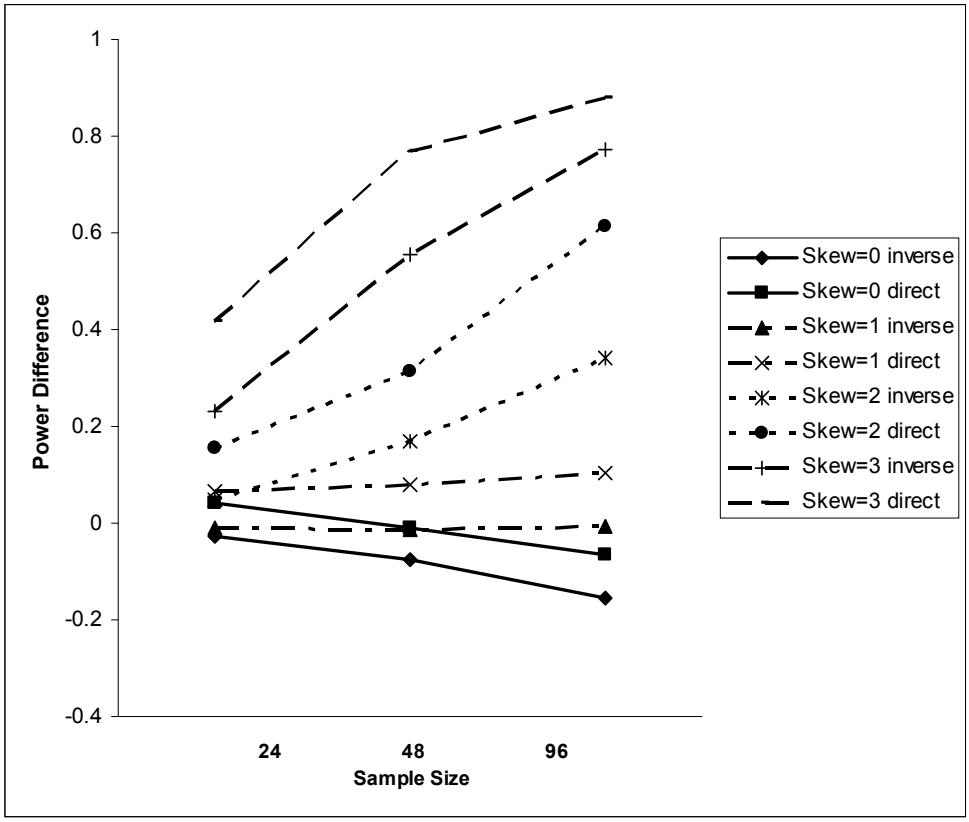

Figure 3. Power difference (2 to 1 variance ratio) values for sample size ratio of 2 to 1.

Figure 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 manuscript, as the skew of the population distributions increases the power advantage of the nonparametric 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.

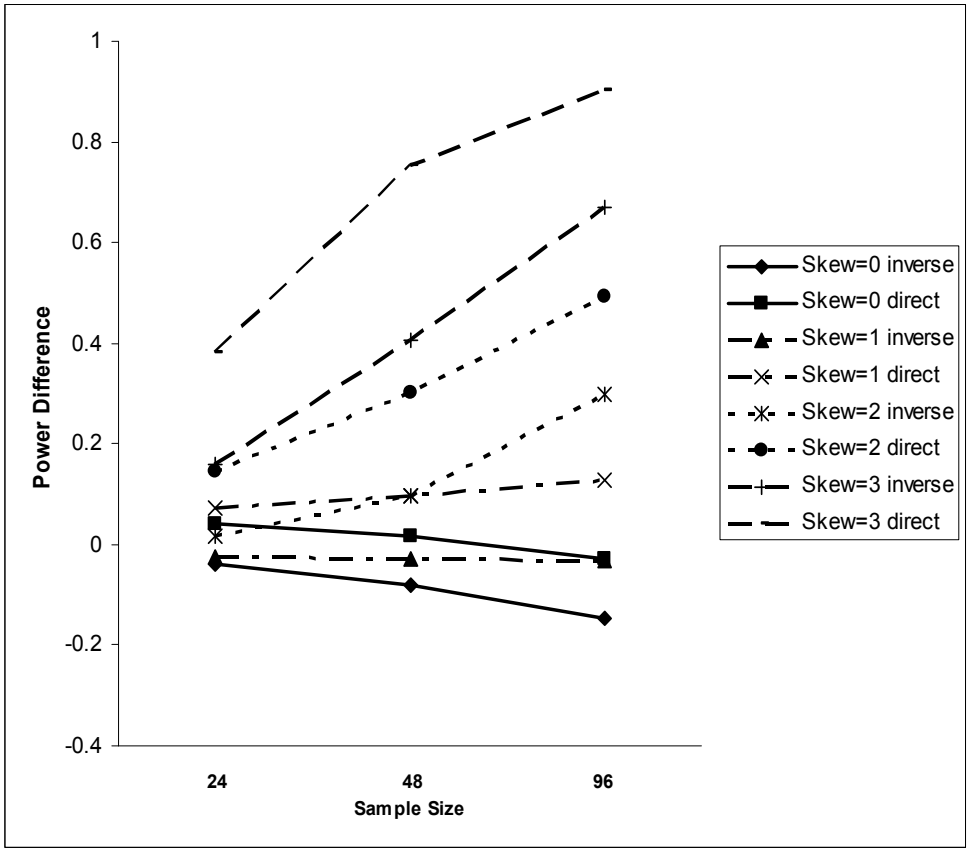

Figure 4. Power difference (2 to 1 variance ratio) values for sample size ratio of 3 to 1.

Figure 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 nonparametric 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.

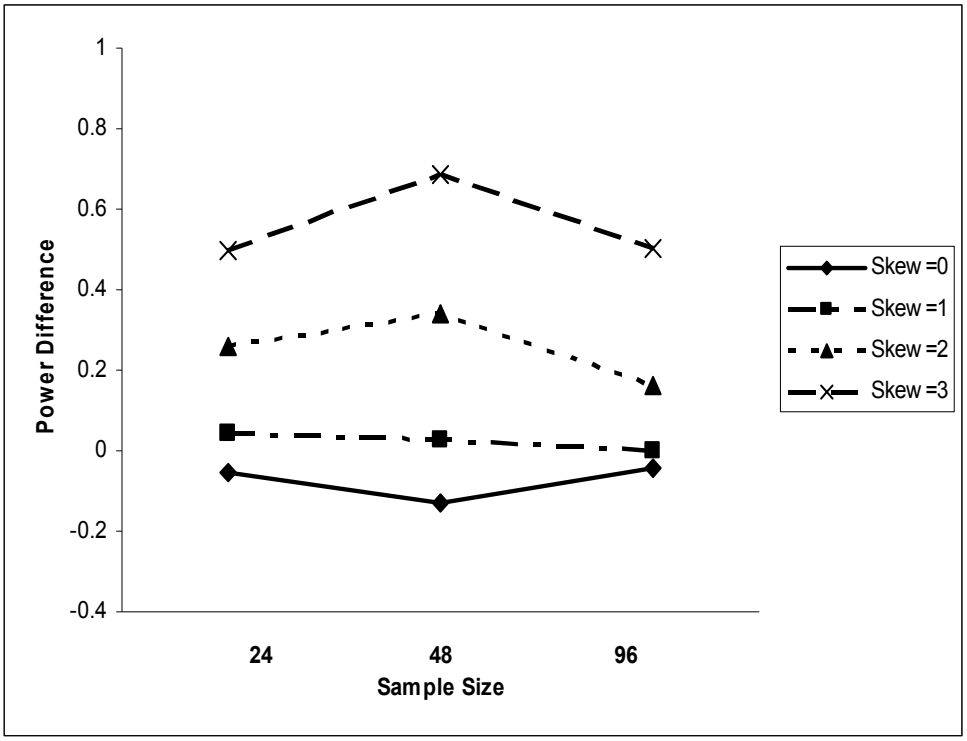

Figure 5. Power difference (4 to 1 variance ratio) for equal sample size ratios.

Figure 8 highlights how the ratio of sample sizes has an effect on the power difference between the nonparametric 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 nonparametric 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.

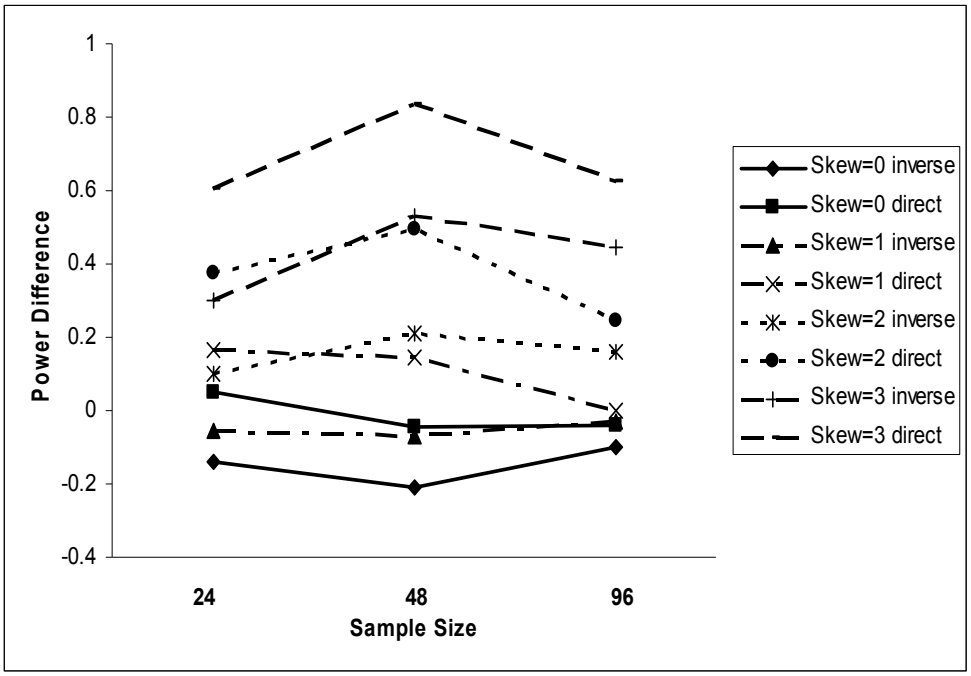

Figure 6. Power difference (4 to 1 variance ratio) values for sample size ratio of 2 to 1.

Figure 9 shows how the ratio of sample sizes has an effect on the power difference between the nonparametric 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 nonparametric 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.

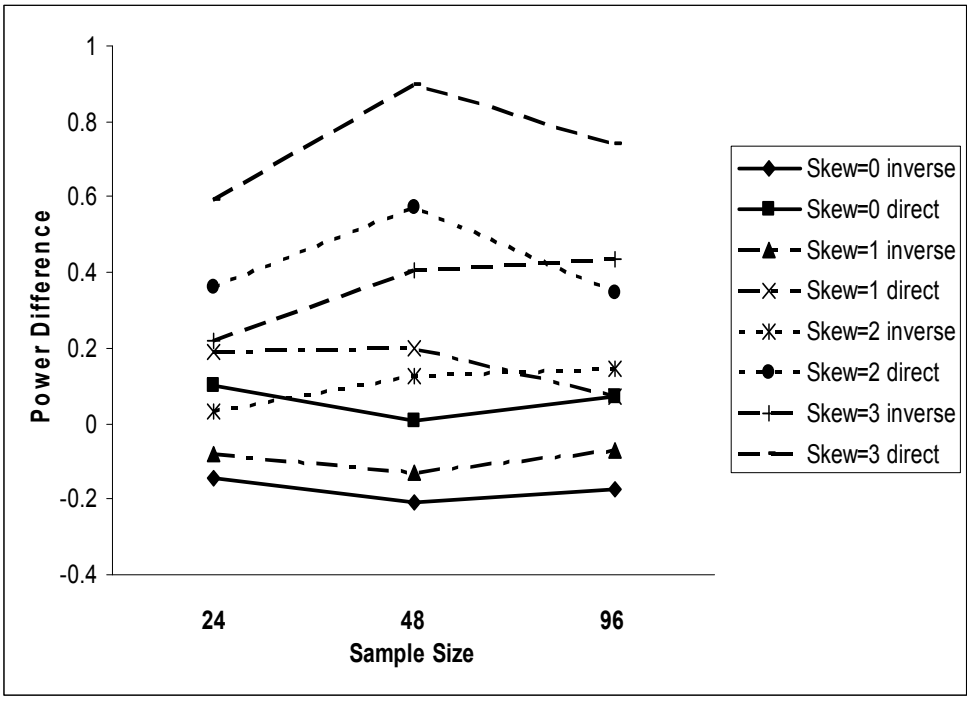

Figure 7. Power difference (4 to 1 variance ratio) values for sample size ratio of 3 to 1.

To further investigate the relationship between the ratio of sample sizes and statistical power, the power values of each test were plotted against each other. Figure 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 nonparametric 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 nonparametric 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 nonparametric test is reduced by .213, whereas the power of the median test experiences a slight increase in power of .066. The nonparametric test was more powerful across all of the levels in Figure 10.

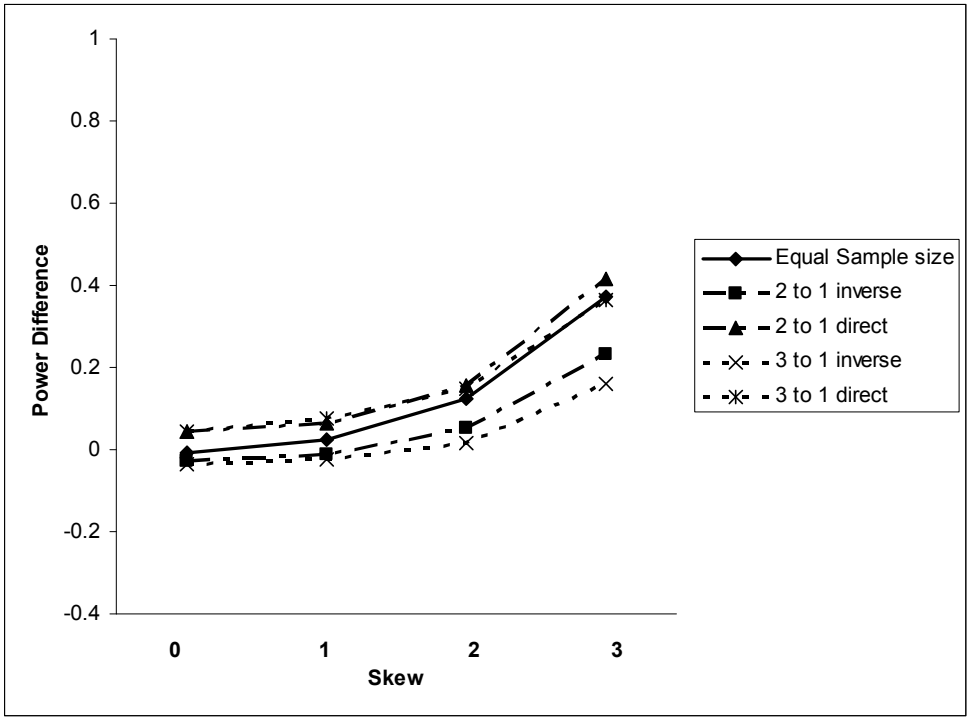

Figure 8. Power Difference (2 to 1 variance ratio) across levels of skew.

Figure 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 nonparametric 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 when the sample size ratio is 3/1. Again, the nonparametric test was more powerful across all of the levels in Figure 11.

Figure 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 nonparametric 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 nonparametric Levene becomes more powerful and the Levene median test becomes less powerful across all the levels of variance ratios.

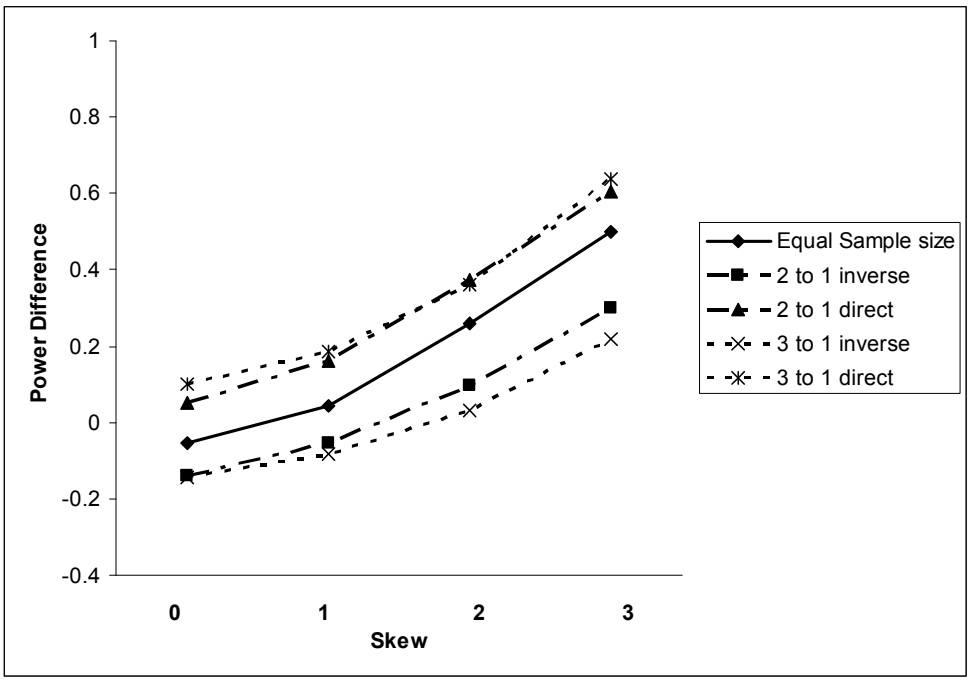

Figure 9. Power Difference (4 to 1 variance ratio) across levels of skew.

DISCUSSION

When data come from heavily skewed population distributions, the nonparametric 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 nonparametric 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 nonparametric test. This leaves the situations when the population distributions are normal where the median test is more powerful than the nonparametric 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 nonparametric 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.

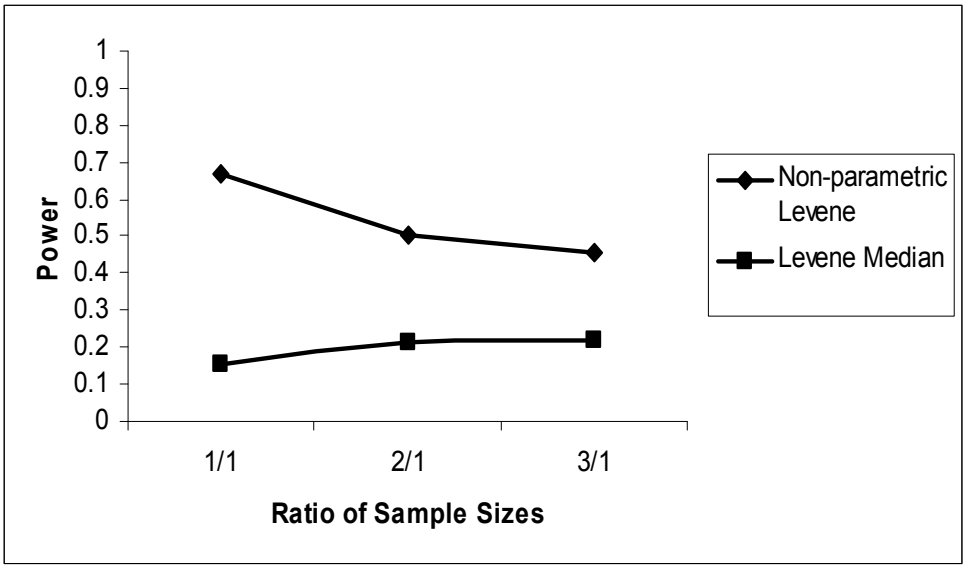

Figure 10. Power comparisons between the Levene median and the nonparametric Levene tests across simulated sample size ratios for inverse pairings.

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 it 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 nonparametric 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).

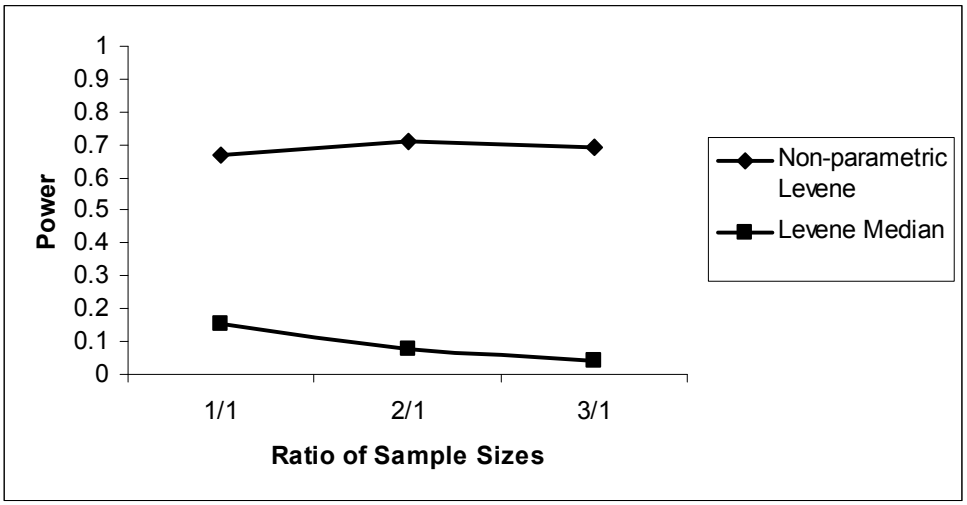

Figure 11. Power comparisons between the Levene median and the nonparametric Levene tests across simulated sample size ratios for direct pairings.

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 nonparametric 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 nonparametric 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 nonparametric 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 nonparametric Levene test for equality of variances.

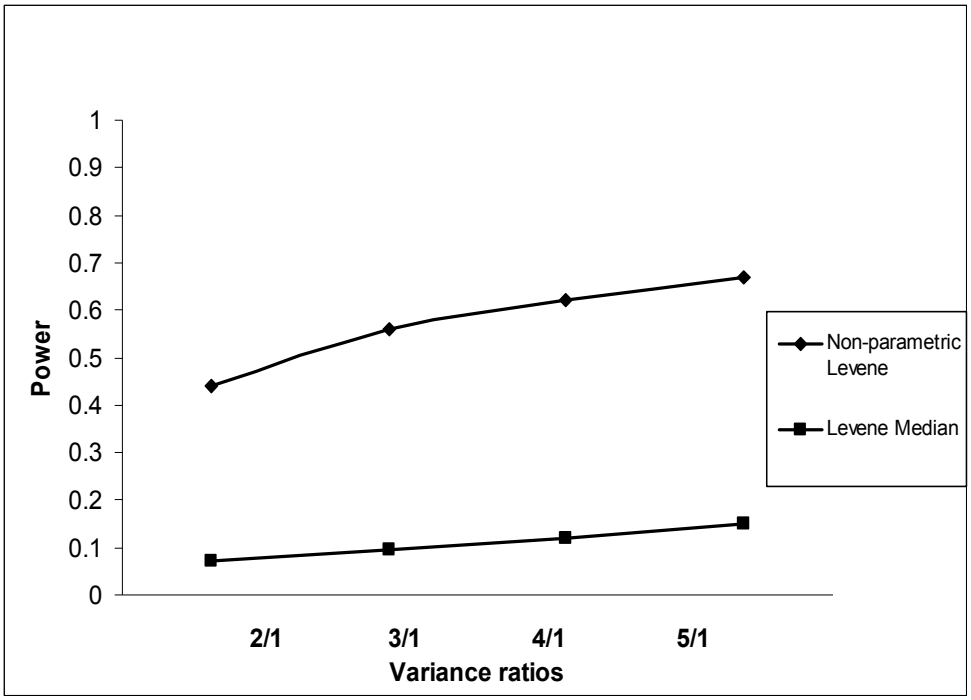

Figure 12. Power comparison between the Levene median and the nonparametric Levene test across simulated variance ratios when population distributions have skew=3.

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 nonparametric version of the Levene test for equal variances. In cases where samples were generated from population distributions with increasing skew, the nonparametric 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 nonparametric Levene test when there is evidence that data come from populations with skewed distributions. Future research will also expand the comparative study by Lim and Loh (1996) and investigate a possible bootstrap version of Nordstokke and Zumbo's nonparametric test.

REFERENCES

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

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*, 229-235.

Brown, M.B. & Forsythe, A.B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association, 69*, 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., & 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*, 351- 361.

Fox, J. (2009). *Package 'car'*. url: http://socserv.socsci.mcmaster.ca/jfox/

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32*, 675–701.

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics, 11*, 86–92.

Hui, W., Gel, Y.R., & Gastwirth, J.L. (2008). lawstat: An R Package for Law, Public Policy and Biostatistics. *Journal of Statistical Software, 28,* 1-26. url: http://www.jstatsoft.org/

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 ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 260*, 583-621.

Lim,T.-S., Loh, W.-Y. (1996). A comparison of tests of equality of variances. *Computational Statistical & Data Analysis, 22*, 287-301.

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*, 1-14.

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.

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

Tomarken, A.J. & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin, 99*, 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 Psychology, 57*, 173-181.

APPENDIX 1

Sample SPSS syntax for generating population distributions.

```
*  setting the seed, i.e., the starting value, for the pseudo-random number generator.

set seed=random.

*================== GROUP 1 ==============================

* Step #1: Generate the variables for your simulation.

INPUT PROGRAM.
LOOP #I = 1 to 60000.
   COMPUTE  group=1.
  COMPUTE dv = RV.CHISQ(1000).
END CASE.
END LOOP.
END FILE.
END INPUT PROGRAM.
EXECUTE.
compute dv = dv - 1000.
execute

* Step #2: sample from the population, with a particular sample size, and run the statistical
test.

COMPUTE draw=UNIFORM(1).
COMPUTE nobreak=1.
RANK  VARIABLES=draw  (A) BY nobreak  /NTILES (5000).
* Check that it works.
FREQ VAR=ndraw.
execute.

SORT CASES BY Ndraw (A) .

SAVE OUTFILE='x1.sav'
  /KEEP=all /COMPRESSED.
```

(Manuscript received: 3 July 2009; accepted: 6 November 2009)