Journal of Modern Applied Statistical Methods

Volume 16 | Issue 1 Article 12

5-1-2017

Robustness and Power Comparison of the Mood-Westenberg and Siegel-Tukey Tests

Linda C. Lowenstein
Wasthington University in St. Louis, lclowenstein@gmail.com

Shlomo S. Sawilowsky
Wayne State University, shlomo@wayne.edu

Recommended Citation
Lowenstein, L. C. & Sawilowsky, S. S. (2017). Robustness and power comparison of the Mood-Westenberg and Siegel-Tukey tests. Journal of Modern Applied Statistical Methods, 16(1), 195-232. doi: 10.22237/jmasm/1493597460

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 195-232.
doi: 10.22237/jmasm/1493597460

ISSN 1538 − 9472

Robustness and Power Comparison of the Mood-Westenberg and Siegel-Tukey Tests

Linda C. Lowenstein
Washington University in St. Louis
St. Louis, MO

Shlomo S. Sawilowsky
Wayne State University
Detroit, MI

The Mood-Westenberg and Siegel-Tukey tests were examined to determine their robustness with respect to Type-I error for detecting variance changes when their assumptions of equal means were slightly violated, a condition that approaches the Behrens-Fisher problem. Monte Carlo methods were used via 34,606 variations of sample sizes, α levels, distributions/data sets, treatments modeled as a change in scale, and treatments modeled as a shift in means. The Siegel-Tukey was the more robust, and was able to handle a more diverse set of conditions.

Keywords: Behrens-Fisher, Mood-Westenberg, Siegel-Tukey

Introduction

"Heteroscedasticity, refers to situations where two or more of the variances are unequal" (Wilcox, 1996, p. 174). The applied statistical literature is vast on how poorly the t and F tests perform under this condition. For instance, it has been demonstrated that small sample sizes, unequal sample sizes, and one-tailed tests can be problematic for the t-test with respect to heteroscedasticity and non-normal data (Sawilowsky & Blair, 1992; Wilcox, 1996; Sawilowsky, 2002). With respect to the Analysis of Variance (ANOVA) F test, the problem is even worse (Brown & Forsythe, 1974; Rogan & Keselman, 1977; Tomarken & Serlin, 1986). Wilcox (1996) stated "our hope is that any problem associated with unequal variances might diminish when there are more than two groups, but the reverse seems to be true" (p. 180). Referring to the ratio (R) of standard deviation between groups in a survey of educational studies, Wilcox (1996) "found that that estimates of R are

Linda C. Lowenstein is a Postdoctoral Researcher in Applied Mathematics and the corresponding author. Email her at: lclowenstein@gmail.com. Dr. Sawilowsky is a Professor in the College of Education and the founding editor of this journal. Email him at professorshlomo@gmail.com.

often higher than 4” (p. 180; see Wilcox, 1989), noting R’s as large as 11 were observed in real world data applications.

Keppel and Wickens (2004) noted “the actual significance level could appreciably exceed the nominal α level when the group variances were unequal. Under these circumstances, we need a way to adjust or modify our analysis” (p. 152). Hence, inflated Type-I errors lead to pronouncements of the statistical significance of nonsense treatments.

Under the truth of the null hypothesis, the counter-argument is having equal means with unequal variance is unrealistic (see, e.g., Sawilowsky, 2002). “That is, this situation will never arise in practice because if the variances are unequal, surely the means are unequal, in which case a Type-I error is not an issue” (Wilcox, 1996, p. 180). The condition of unequal variances between groups is known as the Behrens-Fisher problem, named after the work of W. V. Behrens (1929) and Sir Ronald A. Fisher (1935, 1939) who developed the first expression and approximate solution. Sawilowsky (2002) noted the Behrens-Fisher problem “arises in testing the difference between two means with a t test when the ratio of variances of the two populations from which the data were sampled is not equal to one” (p. 461), and of course expands to layouts with more than two groups.

When the null hypothesis is false, another problem with heteroscedasticity is the t, F, and other parametric tests’ concomitant lack of comparative statistical power. Wilcox (1996) mentioned “there is evidence that problems with Type-I errors with unequal variances reflect undesirable power properties even under normality (Wilcox, Charlin, & Thompson, 1986; Wilcox, 1995)” (p. 180), noting “the power curve might be unusually flat in a region near the null hypothesis (Wilcox, 1995)” especially when the data are skewed (Wilcox, 1996, p. 181). There are situations where the null hypothesis is false, yet the probability of rejecting the null hypothesis is less than α. In this case, small but possibly important treatment effects might be missed.

Sawilowsky and Fahoome (2003) noted non-homogeneity renders most rank-based non-parametric tests even more so ineffective. For example, the Wilcoxon Rank Sum test (Wilcoxon, 1945), which is three to four times more powerful than the t test under common conditions of non-normality due to skew, fares even worse when the treatment impacts scale. Similarly, Sawilowsky (2002) noted “for the case of $K > 2$, Feir-Walsh and Toothaker (1974) and Keselman, Rogan, and Feir-Walsh (1977) found the Kruskal-Wallis test (Kruskal & Wallis, 1952) and expected normal scores test (McSweeney & Penfield, 1969) to be ‘substantially affected by inhomogeneity of variance’” (p. 463).

Change in Scale

There are no exact solutions to the Behrens-Fisher problem. According to Wilcox (1996) and Sawilowsky (2002), the non-parametric Yuen solution (Yuen, 1974), with various modifications, is considered as one of the best approximate solutions. Moreover, methods designed for the purpose of detecting scale or variance changes between sample groups with regard to the level of heteroscedasticity necessary to invoke the Behrens-Fisher problem have been generally overlooked in the applied statistical literature. With respect to the often-cited classical Hartley's (1950) *F*-statistic for determining dispersion (variance) differences as a preliminary test, for example, Sawilowsky (2002) noted the deleterious nature of sequential testing that increases the Type-I error rate. Keppel and Wickens (2004) noted the additional problem of non-normality can greatly impact that *F*-statistic for variance difference detection:

> Unfortunately, in spite of its simplicity and of the fact that it is provided by many packaged computer programs, the *F* max statistic is unsatisfactory. Its sampling distribution, as reflected in the Pearson-Hartley tables, is extremely sensitive to the assumption that the scores have a normal distribution. (p. 150)

According to Neave and Worthington (1988), there were no satisfactory nonparametric tests that could determine the potential of unequal variances irrespective of whether there was also a shift in location. They noted the Mood-Westenberg dispersion test (Westenberg, 1948; Mood, 1950), a non-parametric test based on quartile location and Fisher exact probabilities, determined differences in variances under the assumption that the means of two samples are equal, but stopped short of recommending it as a preliminary test for detecting the Behrens-Fisher condition.

Similarly, Neave and Worthington (1988) noted the Siegel-Tukey test (Siegel & Tukey, 1960), another ordinal non-parametric test based on rankings and Mann-Whitney-*U* probabilities, assumes roughly equal means/medians for detecting variance differences between groups. They bemoaned the absence of detection methods for this condition:

> Several attempts have been made to solve the problem, but all resulting tests suffer from being rather un-powerful or not truly distribution-free or both….It is particularly unfortunate that there appears to be no good distribution-free solution to this problem since

> several researchers have shown that non-normality can upset the behavior of the F-statistic to a very considerable extent. (p.135)

The question arises, therefore, if there are no tests that can detect the occurrence of different variances irrespective of means, then how can it be known if heteroscedasticity or the Behrens-Fisher problem arises so as to be alerted to the need to subsequently apply any of the myriad approximate solutions?

Purpose of the Study

There are no early warning or detection systems indicating the Behrens-Fisher condition exists. The Mood-Westenberg and Siegel-Tukey tests appear promising to fill that need in the statistical repertoire in applied data analysis. In the two group layout, both tests assume equal means (or medians) and $\mu_1 = \mu_2$ (or $\theta_1 = \theta_2$). The null hypothesis (H_0) is the variances are equal. The alternative hypothesis (H_A) is that the variances are not equal. The purpose of this study, therefore, is to examine via Monte Carlo methods their Type-I error rates and comparative statistical power properties as the treatment condition approaches the Behrens-Fisher problem, in order to determine if either test can be used as an early warning.

Methodology

Monte Carlo Methods

An Absoft Pro Fortran (version 14.0.4) program with the IMSL Fortran Numerical Library (version 7.0) was coded to randomly select and assign values to simulated control and treatment groups through sampling with replacement. Rangen 2.0 subroutine (Fahoome, 2002), a 90/95 update to the Fortran 77 version (Blair, 1987), was used to generate pseudo-random numbers from the normal and theoretical distributions. Realpops subroutine 2.0 (Sawilowsky, Blair, & Micceri, 1990) was used to generate pseudo-random samples obtained from real education and psychology populations.

For the Mood-Westenberg code, duplicates found in the control (A) and treatment groups (B) were coded to layout the groups as ABABABAB until all duplicates were accounted for; this method was selected as reasonable because this pattern appears to be unbiased for both groups (the pattern could favor either A or B in the extreme quarters depending upon the random variates sampled). Algorithm AS 62 (Dinneen & Blakesley, 1973) was used to calculate the Mann-

Whitney exact probabilities for the Siegel-Tukey test.[1] When sorting was required, the Recursive Fortran 95 quicksort routine that sorts real numbers into ascending numerical order was used.[2]

There were 34,606 combinations of study parameter conditions employed, based on 11 sample sizes, two α levels (0.05, 0.01) (four levels, including 0.025 and 0.005 were calculated and reviewed in preliminary testing), 11 mathematical distributions and real world data sets, 11 variance changes and 13 small means shifts. Independent sample sizes included $(n_1, n_2) = (5, 5)$; (5, 15); (10, 10); (10, 30); (15, 45); (20, 20); (30, 30); (30, 90); (45, 45); (65, 65); (90, 90). They were generated from three theoretical distributions (normal, exponential, uniform), and eight real world education and psychology data sets identified by Micceri (1986, 1989). The data sets were described as smooth symmetric, extreme asymmetric (growth), extreme asymmetric (decline), extreme bimodality, multimodality and lumpy, discrete mass at zero, discrete mass at zero with gap, and digit preference (see Sawilowsky & Blair, 1992). The use of real data sets in addition to data generated from mathematical models was deemed important in rigorous systematic studies by Bradley (1978) and many others.

Next, the means and variances were modified, beginning with no treatment effect via equal means to establish baseline results. Then, treatment effects of location shifts were gradually increased in small magnitudes, thus increasingly violating the statistical assumption of both tests. Type-I (identifying a variance change when none occurred) and Type-II (not finding a true variance change) error rates under the violations were compared to the counterfactual conditions of equal means.

Type-I and -II Errors

In order to determine robustness measures with respect to Type-I and -II errors, the long-run average rejection rates were calculated after executing 100,000 iterations for each study condition. A counter was incremented for statistically significant iterations. The counter totals were reported as rejection percentages (counter total/100,000). Thus, the long-run averages for the p rejection rate, β rejection rate, and power levels $(1 - \beta)$ were determined.

[1] Additional code was provided by Miller, retrieved from http://lib.stat.cmu.edu/apstat/62
[2] Quicksort routine algorithm provided by Rew with additions from Brainard, retrieved from http://www.fortran.com/qsort_c.f95

Robustness Results

A robust test maintains Type-I and -II error rates in light of assumption violations. Bradley's (1978) liberal limits for Type-I errors of $0.5\alpha \leq$ Type-I error $\leq 1.5\alpha$ was adopted.

Asymptotic and exact probabilities were invoked for each test during preliminary testing. For the Mood-Westenberg test, the Chi-squared (asymptotic) and Fisher exact probabilities were selected. For the Siegel-Tukey test, *Z*-scores (asymptotic) and Mann-Whitney (exact) probabilities were selected. Based on the results for the primary testing, only the asymptotic probabilities were reported because the two probabilities for each statistic were found to track closely to each other. Two α levels, 0.05 and 0.01, were reported during the primary testing (four levels, including 0.025 and 0.005, were calculated and reviewed in preliminary testing).

Simulating Location Shifts and Scale Changes

A treatment was modeled as a shift in location, by multiplying a constant $c = 0.01\text{-}0.12$ (0.01) by the distribution's σ. For example, the standard deviation of the smooth symmetric data set was 4.91. Therefore, a treatment effect of $0.1\sigma = 0.491$ was added to the treatment variates. Cohen (1988) suggested $0.2(\sigma)$ represents a small treatment effect, $0.5(\sigma)$ a moderate treatment effect, and $0.8(\sigma)$ a large treatment effect. On the basis of personal communications with Cohen, Sawilowsky (2009) updated Cohen's de facto standards to also define $d(0.01)$ = very small, $d(1.2)$ = very large, and $d(2.0)$ = huge. The focus of this study, based on Sawilowsky's (2009) standard, was to review only small shifts ($c << 0.2$), and therefore the effect sizes of shift in location selected were 0-0.12σ (0.01), $d = 0$ representing the baseline.

A treatment was modeled as a change in scale by multiplying a constant scale shift of $K = 1 – 3.5$ (0.25) by the random variates of the treatment group after they were centered around zero for both groups by subtracting the distribution mean from the variates; this sets the standard deviation of the control group, over the long run, to approach a normal curve having a variance of 1. Heteroscedasticity is simulated when R, representing the variance ratio difference between the treatment group and the control group, is not equal to 1. K^2, the new simulated variance of the treatment group, is the ratio difference, R, between the post-test treatment and control groups.

It was expected that with ratio variance differences from 1.56 ($K = 1.25$) to 12.25 ($K = 3.5$) (with K increments of 0.25 for K), the alternative hypothesis (H_1)

would be accepted. When the ratio of the variances between the treatment and control groups was equal to 1 ($K = 1$), the condition of equal variances, then the null hypothesis (H_0) was expected to be retained (i.e., fail to reject). These variance ratio differences are consistent with Brown and Forsythe (1974), who reported standard deviation ratio differences of 3 and found concomitant unacceptably high Type-I error rates, and Wilcox (1989), who surveyed the literature and found estimates of standard deviation ratio differences are often higher than 4, and sometimes even as large as 11.

Results

Simulating No Research Treatment Effects with Equal Means Assumption in Place

Demonstration of Adequacy of Algorithms used in this Simulation: Type-I error for Normal Distribution, Means and Variances are Equal

To demonstrate the adequacy of the algorithms used in this simulation, preliminary testing with data sampled from the Gaussian distribution, with equal mean and variances, was performed for all of sample sizes (Table 1). The minimum and maximum asymptotic upper tail rejection rates for α set at 0.05, 0.025, 0.01, and 0.005 for Mood-Westenberg (Chi-squared) were 0.022-0.080, 0.008-0.033, 0.004-0.033, and 0.000-0.016 respectively. For the Siegel-Tukey (Z-scores) they were 0.044-0.058, 0.016-0.027, 0.004-0.010, and 0.000-0.005, respectively. The exact rates tracked close to the associated asymptotic probabilities for both statistics. Exact rates for Mood-Westenberg (Fisher exact) were 0.016-0.072; 0.008-0.033; 0.000-0.020; and 0.000-0.008, and for Siegel-Tukey (Mann-Whitney-U) were 0.044-0.050; 0.016-0.025; 0.008-0.010; and 0.004-0.005. The rejection range was larger for Mood-Westenberg. Additional testing for all equal sample sizes (n_1, n_2) = (5, 5) to (200, 200) yielded robust rates for both statistics (Table 2).

For all sample sizes and α levels, Siegel-Tukey's rejection rates for asymptotic and exact probabilities tracked closer to nominal α as compared with the performance of the Mood-Westenberg Chi-squared and Fisher exact probabilities. It appeared that the latter test's Type-I error rates were dependent on the sample size, and it tracked in an unusual and repeating saw-tooth-like pattern as equal sample sizes were increased by 1 from (5, 5) to (200, 200) at 10,000 iterations (Figures 1 and 2).

The Mood-Westenberg Type-I Fisher exact error rates were occasionally nearly as high as 10% when nominal α was 5%, and 2.4% when nominal α was 1%. Thus, the Mood-Westenberg was observed as an inconsistent test because it did not fit the expected pattern for the Type-I rejection rates to approach nominal α level and remain steadfast as the sample size increased. Instead, it moved in and out of threshold defining robustness as the sample sizes increased. This may be due to the instability of the sampling distribution of the median. See Figures 3 and 4 for Siegel-Tukey results.

Type-I Error: All Distributions/Data Sets, Means and Variances are Equal

At large and equal sample sizes ((45, 45) and above), both statistical tests generally demonstrated robust Type-I rates for the distributions and data sets. Conservative non-robust rate exceptions were noted for discrete mass zero with gap, extreme asymmetric decay, and extreme bimodal data sets (Table 3). However, these conservative non-robust rates suggested unlikely pronouncements of false positives when determining variance change in research settings; hence, at this initial stage, each statistic remained viable candidates to provide robust and powerful heteroscedasticity detection with large and equal sample sizes.

With respect to smaller and unequal sample sizes, Mood-Westenberg demonstrated both liberal and conservative non-robust rates for the distributions/data sets while Siegel-Tukey maintained the same robust rates (and conservatively non-robust for the three data sets mentioned above in Table 3) at all sample sizes except for the smallest sample size of (5, 5) where a few more non-robust conservative rates surfaced for other distributions/data sets at α below 5%. At this point, Siegel-Tukey appeared a more consistent statistic for small and unequal sample sizes with respect to Type-I rates.

Type-II Error: All Distributions/Data Sets, Means are Equal and Variances Change (Classical Behrens-Fisher)

For this phase of testing, in order to provide more stability for Mood-Westenberg, the testing occurred only with the large sample size (90, 90) to observe effects of variance changes simulated with the constant $K = 1.25\text{-}3.5\ (0.25)$. Both statistics were powerful (73-100%) for data sampled from the conservatively non-robust data sets discrete mass zero/gap, extreme asymptotic decay, and extreme bimodal, starting with the smallest variance change when $K = 1.25$ (Table 4; grey shaded area = 100% power). As to be expected, each statistic demonstrated increases in power as the α levels and variance ratio increased. Strong power for these data sets, with conservative Type-I rates, continued to affirm both statistics as potential detection tools; these statistics did not lack for power with these data sets. Siegel-Tukey demonstrated consistent power for these data sets at or above 99% while Mood-Westenberg recorded the same and lower power rates for extreme bimodal (73-90%) when $K = 1.25$.

For the other data sets and distribution at sample size (90, 90) (previously all shown to demonstrate robust Type-I error rates), power was lower as compared to the conservatively non-robust data sets mentioned above, yet still good, for both test statistics, particularly for $K = 1.5$ and above. For Mood-Westenberg, power increased dramatically and quickly, doubling or tripling as variance changed from $K = 1.25$-1.5 (Table 5) for these other data sets/distributions. For Siegel-Tukey, the power also increased quickly, but not as dramatically as Mood-Westenberg because the Siegel-Tukey power rates started off higher at lower K constants.

In general, both statistics demonstrated power approaching 40% or higher early on ($K = 1.25$-1.5, larger α). Siegel-Tukey demonstrated power levels equal to or greater than Mood-Westenberg, sometimes 20-40% higher than Mood-Westenberg with smaller variance changes, as demonstrated in Table 4. For instance, at the smallest change of $K = 1.25$, $\alpha = 0.05$, Siegel-Tukey's power rate for smooth symmetric asymptotic was 0.550 compared to Mood-Westenberg at .165. When α equaled 0.01, Siegel-Tukey's rate was 0.288 as compared to Mood-Westenberg's rate at 0.061. When the variance change level was $K = 1.5$ (Table 5), most α levels yielded power of 40-100%, generally, for all distributions and data sets, for both statistics.

The Siegel Tukey asymptotic and exact probabilities (at $\alpha = 0.05$, 0.025, 0.01, and 0.005) consistently demonstrated equal or greater power rates than the Mood-Westenberg probabilities at every comparison point (α and K's) with all distributions/data sets. Both probability measures for Siegel-Tukey quickly approached 100% power, generally arriving with $K = 2$-2.25 (Table 6); Mood-Westenberg arrived at near 100% with $K = 2.75$-3.0. Siegel-Tukey reached power of nearly 90% and above at all α levels at $K = 1.75$, whereas Mood-Westenberg did not reach these levels until $K = 2.25$ (Table 6). As to be expected, power increased for both statistics as variance change and α levels increased, and therefore these preliminary tests demonstrated that each statistic is robust and powerful, in general, when their mutual assumptions of equal means/medians in place. However, Siegel-Tukey generally appeared more powerful than Mood-Westenberg after this testing phase.

Simulating Research Treatment Effects by Violating the Assumption of Equal Means

At this point, attention was turned to the primary focus of the study: would the Mood-Westenberg and the Siegel-Tukey tests remain robust with respect to Type-I and Type-II rejection rates under conditions of simulated treatment effects (i.e., the means began to shift slightly, violating the statistical assumptions). Preliminary testing results of 10,000 means shifts from 0.00001 to 0.1 (0.00001) suggested an appropriate mean shift range, useful for testing, would be 0.01-0.12 (0.01).

To determine the properties for each statistic after sampling from the thousands of combination of populations, sample sizes, means shifts, variance change, and α levels, it would be necessary to review all output, particularly with respect to the smaller and unequal sample sizes. However, general conclusions are made and presented here for both statistics, with respect to whether the mathematical distributions and real-world data sets could be characterized as a normal type distribution (e.g., unimodal shape, asymptotic light tails, symmetric about the means) or not. Normal type distributions are discussed as a group and include: normal, digit preference, discrete mass zero, smooth symmetric, and uniform. Non-normal type distributions, discussed as a group, include: extreme asymmetric growth, extreme asymmetric decay, extreme bimodal, and discrete mass zero with gap. Having demonstrated unique outcomes, exponential and multi-modal lumpy are discussed separately.

With minor exceptions for the exponential and multi-modal lumpy, general conclusions for the distributions and data sets were not greatly affected by the range of the tested means shift levels 0.01-0.12 (0.01); therefore, conclusions for particular distributions and data sets will generally hold for all of the tested means shift levels, especially for larger sample sizes and α levels of 0.05. When robustness was present, larger α levels (0.05), larger and equal sample sizes and larger variance change levels rendered testing measurements more robust and powerful for each distribution and data set.

Type-I Rejection Rates: For All Distributions/Data Sets, Variances are Equal

The statistics were first tested with slight means shifts, $0.01(\sigma)$-$0.12(\sigma)$ (0.01), when simulating post-test equal variance outcomes. Typical results are noted in Table 7 for sample size (90, 90) and mean shift at $c = 0.06$. The expectation was that nominal α rejection rates would hold when the means began to shift. Mood-Westenberg, for most normal type distributions (e.g., digit preference, normal, smooth symmetric, uni), particularly for large sample sizes (i.e., (20, 20);

(30, 30); (30, 90)), maintained generally robust (and conservative non-robust) rejection rates at all of the tested means shifts with some slightly liberal rate exceptions at some small and small/unequal sample sizes or sometimes at 1% α. As noted with sample size (90, 90), in Table 7, the normal type discrete mass zero, sometimes demonstrated small liberal, non-robust rates but robust rejection rates were noted for many other sample sizes, particularly when nominal α was 5%. However, analyzing non-normal distributions (asymmetric growth, discrete mass zero with gap, extreme asymmetric decay, extreme bimodal), Mood-Westenberg, for both asymptotic and exact probabilities at the large sample size (90, 90), calculated many extremely liberally non-robust rejection rates even at the smallest incremental level of 0.01. The test results from data sampled from multi-modal lumpy demonstrated liberal non-robust rejection rates generally at and above means shift $c = 0.09$ for some sample sizes, such as (90, 90), and was robust for many other sample sizes. Results from data sampled from the exponential distribution demonstrated robust rates up to means shifts of 0.06 when, for instance, for sample size (65, 65) or (90, 90) (Table 7), for nominal α below 2.5%, the rejection rates started to trend above nominal α levels in the liberal direction, increasing in slight liberalness with each increase in means shift. Starting with mean shift $c = 0.07$ and above, under Mood-Westenberg, the test results demonstrated that the exponential distribution was liberally non-robust at all α levels for sample size (90, 90). Other sample sizes for exponential also reflected this pattern. Generally, the non-robust Mood-Westenberg results for the exponential distribution were in the liberal direction.

With respect to the Siegel-Tukey statistic, at sample size (90, 90) and mean shift $c = 0.06$, (Table 7), for both asymptotic and exact probability measures and for all other means shifts, testing revealed robust rates for the data sampled from all of the normal type distributions (digit preference, discrete mass zero, normal, smooth symmetric, and uniform). This robust rejection rate pattern was also demonstrated at most small and small/unequal sample sizes, unlike Mood-Westenberg. Similar to the Mood-Westenberg, as the means shifted, non-robust results were detected for the data sampled from most non-normal type distributions (including asymmetric growth, discrete mass zero with gap, extreme asymmetric decay); however, unlike Mood-Westenberg, all indicators of these non-robust measures were in the conservative direction except the liberal rates found with the test results from asymmetric growth.

A particularly strong and unique outcome for Siegel-Tukey was noted for the non-normal extreme bimodal data set. At sample size (90, 90), Siegel-Tukey, unlike Mood-Westenberg, demonstrated robust measures at virtually all means shifts for extreme bimodal (slight liberal exceptions were noted at 0.5% α level when means shift was at $c = 0.02$, 0.03, and 0.1). This strong robust rejection pattern for all means shifts was also noted in the data sampled from the extreme bimodal data for all equal sample sizes and for unequal sample sizes when α was 0.05.

Results demonstrated that the data sampled from the multi-modal lumpy data set was robust at lower means shifts but began to show conservative non-robust measures at means shifts generally at and above 0.09 for sample size (90, 90). However, many other sample sizes were robust at all means shifts. Results for data sampled from the exponential distribution became conservatively non-robust at means shift of $c = 0.03$ at sample size (90, 90). This was a general pattern for other large and equal sample sizes, although some smaller and unequal sample sizes maintained robust rates at higher mean shifts.

Siegel-Tukey's conservative non-robust rate exceptions, for non-normal distributions, multi-modal lumpy, and exponential, were deemed positive outcomes because this condition would obviate large pronouncements of nonsense variance changes. It did not demonstrate sample size instability that seemed pervasive throughout the study for Mood-Westenberg. At this point, after demonstrating large liberal rejection rates as the means shifted slightly with the non-normal type distributions, the Mood-Westenberg necessarily dropped out of consideration as a method to detect variance changes with respect to these distributions/data sets (though it maintained viability for exponential distributions and multi-modal lumpy data sets at lower means shift levels); however with the exception of the asymmetric growth data set, which measured liberal rejection rates, Siegel-Tukey demonstrated robust and conservatively robust rejection rates and thus continued as a viable instrument to detect heteroscedasticity for all other distributions/data sets provided power could be demonstrated next as the variance began to change.

Type-II Rejection Rates: For All Distributions/Data Sets, Variances are Unequal

During the final phase of the primary study, as assumptions were violated and variance changes simulated, the investigation focused upon reporting Mood-Westenberg and Siegel-Tukey asymptotic probabilities (Chi-squared and *Z*-scores, respectively) with nominal α of 0.05 and 0.01. The expectation was that power levels of at least 40% would be generally demonstrated.

With respect to the normal type distributions, both statistics generally demonstrated at least 40% power for all means shifts and variance changes for large samples sizes (i.e., (30, 30) and (30, 90)), especially for $\alpha = 0.05$. Power (at sample size (30, 30) and above) approached 40% generally around variance change with $K = 1.75$-2 for α 0.05 and 0.01. For these normal type distributions, Siegel-Tukey typically demonstrated 40% power starting at smaller sample sizes (sample size (20, 20); Table 8) and often at lower levels of K changes ($K = 1.5$; Table 9) as compared to Mood-Westenberg (see also sample size (20, 20), uniform, for Siegel-Tukey's superior power; Table 10). Power for each statistic was shown to increase as α, variance, and sample size increased as demonstrated when the uniform sample size increased from (20, 20) (Table 10) to (45, 45) (Table 11) to (65, 65) (Table 12). While there were power improvements for both statistics as these parameters increased, Siegel-Tukey always demonstrated greater (or equal) power as compared to Mood-Westenberg at each point of comparison, sometimes yielding 20-40% more power at lower variance change levels.

For data sampled from non-normal distributions, both statistics reported much larger rejection rates as compared to the normal types when the variance changed and means shifted. This high rejection rate, starting from the smallest constant $K = 1.25$-3.5 (0.25), is reported for the representative data set, discrete mass zero with gap at sample sizes (45, 45) (Table 13). However, these large power rate results for the data sampled from non-normal distributions under Mood-Westenberg were meaningless due to the large liberal rejection rates noted for these when the variances were equal at $K = 1$ (see also large rate rejections 0.991-1 for discrete mass zero with gap and asymmetric decay in Table 7, at sample size (90, 90) when variances were equal).

However, given the conservative Type-I rejection rates (0.000) demonstrated when variances were equal for Siegel-Tukey, the large power it reported as variances changed is meaningful and impressive. For both small (e.g., (10, 10); Table 14) and large (e.g., (45, 45); Table 13) sample sizes, the Siegel-Tukey results for non-normal distributions, with the exception of asymmetric growth with many liberal Type-I rejection rates, had significant power that quickly approaching 99% at even the lowest levels of variance change (see also extreme bimodal; Table 15). For these non-normal power rates, a desired more gradual increase in power for Siegel-Tukey might have been demonstrated at lower levels of variance change between $K = 1$ and 1.25, but these levels were not tested. An impressive power finding was noted for the extreme bimodal data set under the Siegel-Tukey statistic, wherein the Type-I rejection rates were generally robust (instead of conservatively non-robust as Siegel-Tukey demonstrated with other non-normal distributions), particularly when sample sizes were equal (Table

7) and for unequal samples sizes when $\alpha = 0.05$. These robust findings, together with the high power noted in Table 15, renders the Siegel-Tukey test particularly useful in research settings where extreme bimodal data sets are common.

Finally, the results for both statistics with the data sampled from multi-modal lumpy and exponential demonstrated at least 40% power with large sample sizes (generally (30, 30), and above, including (30, 90)), especially when $\alpha = 0.05$. For Mood-Westenberg these results were attained typically at $K = 1.5$; for Siegel-Tukey at the lower $K = 1.25$. For the multi-modal lumpy data set with $\alpha = 0.05$ and the smallest variance change $K = 1.25$, 40% power was generally attained when sample size was (65, 65) for Mood-Westenberg and (30, 30) for Siegel-Tukey (Table 16, 17). For the exponential distribution (Table 18, 19), when $\alpha = 0.05$, 40% power was generally attained when $K = 1.5$ at sample size (30, 30) and (20, 20), respectively. Once again, Siegel-Tukey demonstrated greater or equal power at all comparison points than Mood-Westenberg for both of these distributions/data sets. For Mood-Westenberg, stable power was generally best when means shifts were below $c = 0.09$ for multi-modal lumpy and $c = 0.06$ for exponential due to some liberal non-robust Type-I rates at larger means shift levels. Siegel-Tukey was most powerful for these with lower means shifts ($c = 0.01$-0.08 for multi-modal lumpy and $c = 0.01$-0.03 for exponential) due to some conservative non-robust null rejections at larger mean shift levels.

Conclusion

Methods for Behrens-Fisher detection have been overlooked in statistical literature and, up to now, there have been no early warning or detective systems indicating the Behrens-Fisher condition exists. Siegel-Tukey appears promising as a method that might fill this void. Invoking the Siegel-Tukey statistic for the purpose of detecting variance changes could provide an effective precursor to the discovery of small yet important treatment effects in many research settings approaching Behrens-Fisher.

The Mood-Westenberg statistic also identified variance changes accompanied by slight mean shifts for normal type distributions, particularly with large sample sizes at or above $n = 30, 30$ (and at some smaller mean shifts for the multi-modal lumpy data set and the exponential distribution). However, Mood-Westenberg could not approach the levels of superior power demonstrated by Siegel-Tukey with these data sets/distributions and could not consistently demonstrate Siegel-Tukey's robust Type-I rejection rates at small sample sizes, especially when α was at 0.01.

Another significant comparative advantage demonstrated by the Siegel-Tukey statistic was its robust (or conservatively non-robust) and powerful results for non-normal distributions while Mood-Westenberg could not withstand the same means shift assumption violations for these types, demonstrating large liberal Type-I rejection rates. Therefore, as a detection tool for determining outcomes approaching Behrens-Fisher, the Mood-Westenberg statistic would be limited to research settings utilizing only normal type data distributions (best with larger sample sizes), the multi-modal lumpy data set, and the exponential distribution. Additionally, it is believed that the inability to stabilize Type-I rejection rates to approach nominal α level as sample sizes increased would render the Mood-Westenberg statistic generally less reliable in research settings.

Therefore, the Siegel-Tukey statistic might reasonably be promoted as the current statistic of choice in many scientific, educational and psychological research environments to detect heteroscedasticity whenever conditions approaching Behrens-Fisher arise with the concomitant problem of determining the existence of small means shift around zero. Siegel-Tukey demonstrated particularly strong measures for the extreme bimodal data set, often found within educational settings, when samples sizes were equal (or unequal at $\alpha = 0.05$). Siegel-Tukey's robust and powerful measures in detecting variance changes with all but one (asymmetric growth) of the 11 tested distributions/data sets demonstrated that it could be an important new instrument in the researcher's repertoire for data analysis. It has the potential to operate within a broad range of testing conditions to alert the researcher to the necessity of choosing an appropriate test statistic which could ultimately lead to the discovery of small treatments that might otherwise go unnoticed. The Siegel-Tukey statistic demonstrated its ability to be an effective precursor that would make known the need to replace testing statistics dependent on the equal variance assumptions, such as Student's-*t*, and to choose instead to apply any of the myriad of approximate Behrens-Fisher solutions, such as the Yuen's solution.

References

Behrens, W. V. (1929). Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen. *Landwirtschaftliche Jahrbücher, 68*, 807-837.

Blair, R. C. (1987). Rangen (Version 1.0) [Software]. Boca Raton, FL: IBM.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics, 16*(1), 129-132. doi: 10.2307/1267501

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dinneen, L. C., & Blakesley, B. C. (1973). Algorithm AS 62: A generator for the sampling distribution of the Mann-Whitney *U* statistic. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 22*(2), 269-273. doi: 10.2307/2346934

Fahoome, G. F. (2002). JMASM1: *RANGEN* 2.0 (*Fortran* 90/95). *Journal of Modern Applied Statistical Methods, 1*(1), 182-190. doi: 10.22237/jmasm/1020255960

Feir-Walsh, B. J., & Toothaker, L. E. (1974). An empirical comparison of the ANOVA F-test, normal scores test and Kruskal-Wallis test under violation of assumptions. *Educational and Psychological Measurement, 34*(4), 789-799. doi: 10.1177/001316447403400406

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics, 6*(4), 391-398. doi: 10.1111/j.1469-1809.1935.tb02120.x

Fisher, R. A. (1939). The comparison of samples with possibly unequal variances. *Annals of Eugenics, 9*(2), 174-180. doi: 10.1111/j.1469-1809.1939.tb02205.x

Hartley, H. O. (1950). The maximum *F*-ratio as a short-cut test for heterogeneity of variance. *Biometrika, 37*(3/4), 308-312. doi: 10.2307/2332383

Keppel, G. & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook*. Upper Saddle River, NJ: Pearson Prentice Hall.

Keselman, H. J., Rogan, J. C., & Feir-Walsh, B. J. (1977). An evaluation of some non-parametric and parametric tests for location equality. *British Journal of Mathematical and Statistical Psychology, 30*(2), 213-221. doi: 10.1111/j.2044-8317.1977.tb00742.x

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one criterion variance analysis. *Journal of the American Statistical Association, 47*(260), 583-621. doi: 10.2307/2280779

McSweeney, M., & Penfield, D. (1969). The normal scores test for the *c*-sample problem. *British Journal of Mathematical and Statistical Psychology, 22*(2), 177-192. doi: 10.1111/j.2044-8317.1969.tb00429.x

Micceri, T. (1986, November). *A futile search for the statistical chimera of normality*. Paper presented at the 31st Annual Convention of the Florida Educational Research Association, Tampa, FL.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166. doi: 10.1037//0033-2909.105.1.156

Mood, A. M. (1950). *Introduction to the theory of statistics*. New York, NY: McGraw Hill.

Neave, H. R., & Worthington, P. L. (1988). *Distribution-free tests*. London, UK: Unwin Hyman Ltd.

Rogan, J. C., & Keselman, H. J. (1977). Is the ANOVA *F*-test robust to variance heterogeneity when sample sizes are equal? An investigation via a coefficient of variation. *American Educational Research Journal, 14*(4), 493-498. doi: 10.2307/1162346

Sawilowsky, S. S. (2002). Fermat, Schubert, Einstein, and Behrens-Fisher: The probable difference between two means when $\sigma_1^2 \neq \sigma_2^2$. *Journal of Modern Applied Statistical Methods, 1*(2), 461-472. doi: 10.22237/jmasm/1036109940

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods, 8*(2), 597-599.

Sawilowsky, S.S., & Blair, R.C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin, 111*(2), 353-360. doi: 10.1037/0033-2909.111.2.352

Sawilowsky, S. S., Blair, R. C., & Micceri, T. (1990). A PC FORTRAN subroutine library of psychology and educational data sets. *Psychometrika, 55*(4), 729.

Sawilowsky, S. S., & Fahoome, G. F. (2003). *Statistics through Monte Carlo simulation with Fortran*. Oak Park, MI: JMASM.

Siegel, S. &. Tukey, J. W. (1960). A non-parametric sum of ranks procedure for relative spread in unpaired samples. *Journal of the American Statistical Association, 55*(291), 429-445. doi: 10.2307/2281906

Tomarken, A. J., & Serlin R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin, 99*(1), 90-99. doi: 10.1037/0033-2909.99.1.90

Westenberg, J. (1948). Significance test for median and interquartile range in samples from continuous population of any form. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, 51*, 252-261.

Wilcox, R. R. (1989). Adjusting for unequal variances when comparing means in one-way and two-way fixed effects ANOVA models. *Journal of Educational Statistics, 14*(3), 269-278. doi: 10.2307/1165019

Wilcox, R. R. (1995). ANOVA: The practical importance of heteroscedastic methods, using trimmed means versus means, and designing simulation. *British Journal of Mathematical and Statistical Psychology*, *48*(1), 99-114. doi: 10.1111/j.2044-8317.1995.tb01052.x

Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.

Wilcox, R. R., Charlin, V., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA *F*, *W*, and *F** statistics. *Communications in Statistics – Simulation and Computation, 15*(4), 933-944. doi: 10.1080/03610918608812553

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*(6), 80-83. doi: 10.2307/3001968

Yuen, K. K. (1974). The two-sample trimmed *t* for unequal population variances. *Biometrika*, *61*(1), 165-170. doi: 10.2307/2334299

Appendix A: Figures

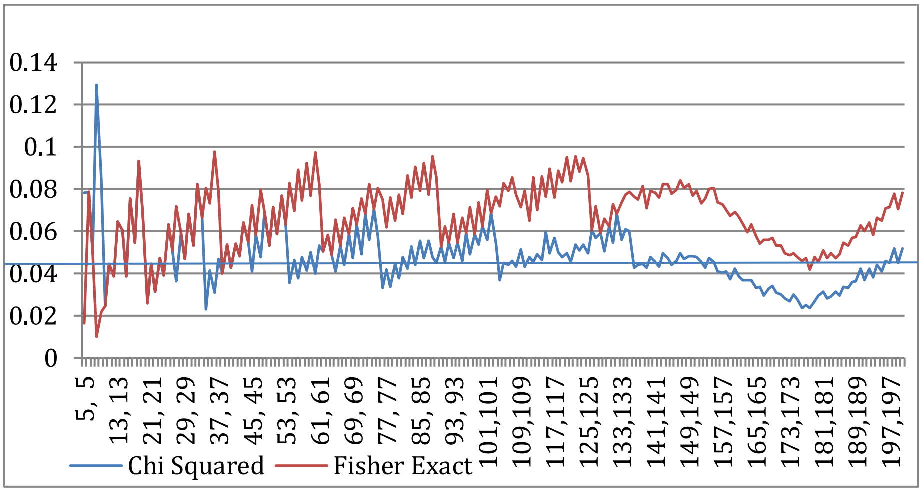

Figure 1. Mood-Westenberg Type-I error rate, comparisons between Chi Squared (blue) and Fisher Exact (red) for all equal sample sizes from (5, 5) to (200, 200), for Normal distribution, 0.05 *α*, 10,000 repetitions

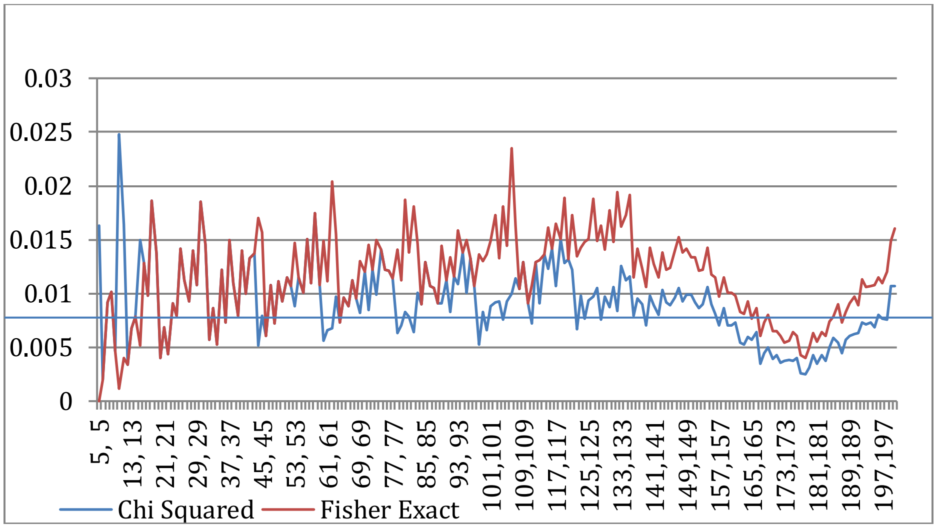

Figure 2. Mood-Westenberg Type-I error rate, comparisons between Chi Squared (blue) and Fisher Exact (red) for all equal sample sizes from (5, 5) to (200, 200), for Normal distribution, 0.01 *α*, 10,000 repetitions

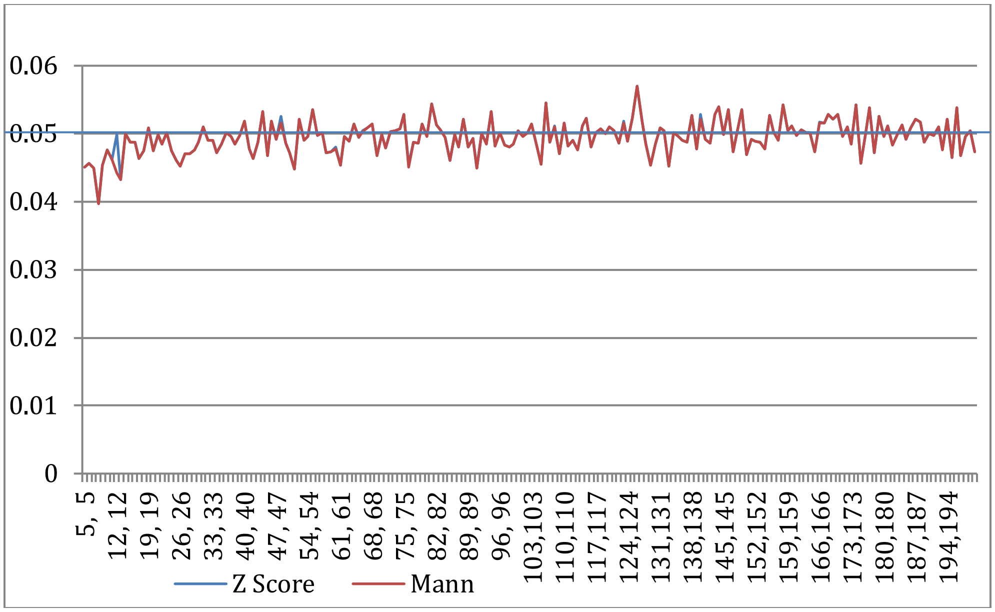

Figure 3. Siegel-Tukey Type-I error rate, comparisons between *Z* Scores (blue) and Mann-Whitney (red) for all equal sample sizes from (5, 5) to (200, 200), for Normal distribution, 0.05 *α*, 10,000 repetitions

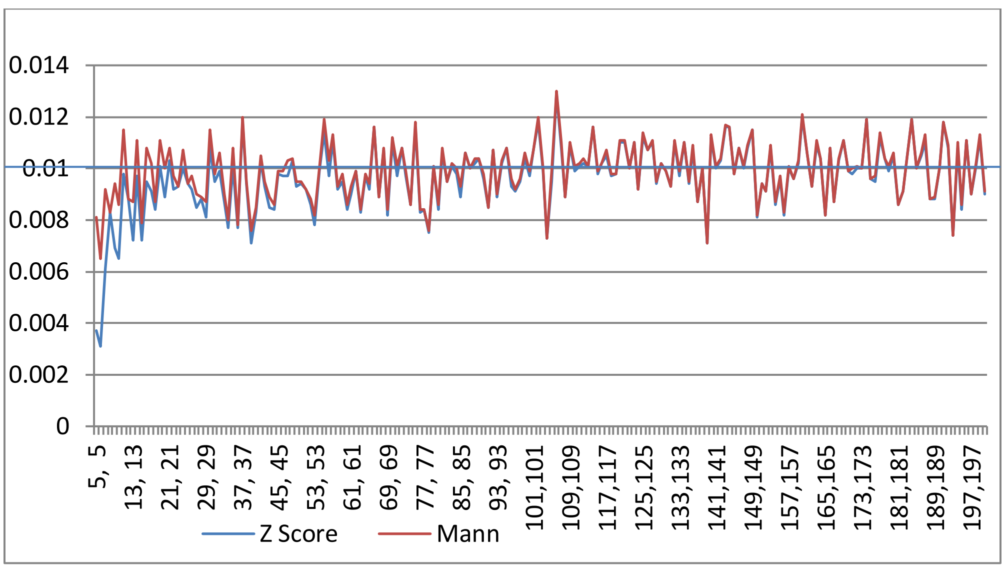

Figure 4. Siegel-Tukey Type-I error rate, comparisons between *Z* Scores (blue) and Mann-Whitney (red) for all equal sample sizes from (5, 5) to (200, 200), for Normal distribution, 0.01 *α*, 10,000 repetitions

Appendix B: Tables

Table 1. Type-I error rates for Mood-Westenberg and Siegel-Tukey, one-tailed directional rest, for various sample sizes and *α* levels when sampling is from the normal distribution, 100,000 repetitions, variances are equal and means are equal

	α							
	0.050		**0.025**		**0.010**		**0.005**	
Sample Size	**A**	**E**	**A**	**E**	**A**	**E**	**A**	**E**
Mood-Westenberg								
5, 5	0.080	0.016	0.016	0.016	0.016	0.000	0.016	0.000
5, 15	0.033	0.033	0.033	0.033	0.033	0.000	0.000	0.000
10, 10	0.022	0.022	0.022	0.022	0.022	0.001	0.001	0.001
10, 30	0.066	0.066	0.008	0.008	0.008	0.008	0.008	0.008
15, 45	0.072	0.072	0.016	0.016	0.016	0.016	0.002	0.002
20, 20	0.026	0.026	0.026	0.026	0.004	0.004	0.004	0.004
30, 30	0.068	0.068	0.019	0.019	0.019	0.019	0.004	0.004
30, 90	0.056	0.056	0.020	0.020	0.006	0.020	0.006	0.006
45, 45	0.043	0.070	0.025	0.025	0.007	0.014	0.004	0.004
65, 65	0.041	0.063	0.026	0.026	0.010	0.010	0.006	0.006
90, 90	0.052	0.052	0.025	0.025	0.011	0.011	0.004	0.004
Siegel-Tukey								
5, 5	0.047	0.047	0.016	0.016	0.004	0.008	0.000	0.004
5, 15	0.058	0.048	0.025	0.021	0.010	0.010	0.004	0.004
10, 10	0.044	0.044	0.021	0.021	0.007	0.009	0.003	0.004
10, 30	0.051	0.047	0.024	0.024	0.010	0.010	0.004	0.004
15, 45	0.051	0.050	0.027	0.025	0.010	0.010	0.005	0.005
20, 20	0.048	0.048	0.025	0.025	0.010	0.010	0.004	0.005
30, 30	0.050	0.050	0.023	0.024	0.009	0.010	0.005	0.005
30, 90	0.050	0.049	0.025	0.024	0.009	0.010	0.005	0.005
45, 45	0.049	0.049	0.024	0.024	0.010	0.010	0.005	0.005
65, 65	0.049	0.049	0.024	0.024	0.010	0.010	0.005	0.005
90, 90	0.050	0.050	0.025	0.025	0.010	0.010	0.005	0.005

Note: For Mood-Westenberg, A = asymptotic Chi-squared probability, E = Fisher exact probability; for Siegel-Tukey, A = asymptotic *Z*-score probability, E = Mann-Whitney-U exact probability

Table 2. Type-I error rate averages for all sample sizes (5, 5) to (200, 200) for 10,000 repetitions, Normal distribution

Mood-Westenberg

α							
0.050		**0.025**		**0.010**		**0.005**	
A	**E**	**A**	**E**	**A**	**E**	**A**	**E**
0.048	0.067	0.024	0.031	0.009	0.012	0.005	0.005

Table 2, continued.

Siegel-Tukey							
α							
0.050		0.025		0.010		0.005	
A	E	A	E	A	E	A	E
0.049	0.049	0.024	0.025	0.010	0.010	0.005	0.005

Note: For Mood-Westenberg, A = asymptotic Chi-squared probability, E = Fisher exact probability; for Siegel-Tukey, A = asymptotic *Z*-score probability, E = Mann-Whitney-U exact probability

Table 3. Type-I error rates for Mood-Westenberg and Siegel-Tukey, one-tailed directional test, for sample size (45, 45) and *α* levels when sampling is from all distributions/data sets, 100,000 repetitions, variances are equal, and means are equal

	α							
	0.050		0.025		0.010		0.005	
Distribution	A	E	A	E	A	E	A	E
Mood-Westenberg								
Asym Growth	0.040	0.067	0.024	0.024	0.007	0.013	0.003	0.003
Digit pref	0.042	0.069	0.024	0.024	0.007	0.014	0.004	0.004
Disc mass zero	0.040	0.066	0.023	0.023	0.007	0.012	0.003	0.003
Disc mass zero gap	0.004	0.008	0.002	0.002	0.000	0.001	0.000	0.000
Exponential	0.043	0.071	0.025	0.025	0.007	0.014	0.004	0.004
Extrm asym decay	0.021	0.039	0.011	0.011	0.002	0.005	0.001	0.001
Extrm bimodal	0.022	0.041	0.011	0.011	0.002	0.005	0.001	0.001
Multi-modal lumpy	0.042	0.069	0.024	0.024	0.007	0.014	0.004	0.004
Normal	0.043	0.070	0.025	0.025	0.007	0.014	0.004	0.004
Smooth sym	0.040	0.066	0.023	0.023	0.007	0.013	0.003	0.003
Uni	0.043	0.070	0.025	0.025	0.008	0.015	0.004	0.004
Siegel-Tukey								
Asym Growth	0.046	0.047	0.022	0.022	0.008	0.009	0.004	0.004
Digit pref	0.049	0.050	0.024	0.025	0.009	0.010	0.005	0.005
Disc mass zero	0.047	0.048	0.023	0.024	0.009	0.009	0.004	0.005
Disc mass zero gap	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000
Exponential	0.050	0.050	0.026	0.026	0.010	0.010	0.005	0.005
Extrm asym decay	0.011	0.011	0.003	0.003	0.001	0.001	0.000	0.000
Extrm bimodal	0.023	0.024	0.009	0.009	0.003	0.003	0.001	0.001
Multi-modal lumpy	0.049	0.050	0.024	0.025	0.009	0.010	0.005	0.005
Normal	0.049	0.049	0.024	0.024	0.010	0.010	0.005	0.005
Smooth sym	0.048	0.048	0.023	0.024	0.009	0.009	0.004	0.004
Uni	0.049	0.049	0.025	0.025	0.009	0.009	0.005	0.005

Note: For Mood-Westenberg, A = asymptotic Chi-squared probability, E = Fisher exact probability; for Siegel-Tukey, A = asymptotic *Z*-score probability, E = Mann-Whitney-U exact probability

Table 4. Type-II errors/power rates for Mood-Westenberg and Siegel-Tukey, one-tailed directional test, for various *α* levels and sample size of (90, 90) when sampling is from all distributions/data sets, 100,000 repetitions, means are equal, and variance change is 1.25

	α							
	0.050		0.025		0.010		0.005	
Distribution	**A**	**E**	**A**	**E**	**A**	**E**	**A**	**E**
Mood-Westenberg								
Asym Growth	0.457	0.457	0.369	0.369	0.289	0.289	0.219	0.219
Digit pref	0.265	0.265	0.179	0.179	0.114	0.114	0.068	0.068
Disc mass zero	0.197	0.197	0.128	0.128	0.078	0.078	0.044	0.044
Disc mass zero gap			0.999	0.999	0.996	0.996	0.991	0.991
Exponential	0.478	0.478	0.360	0.360	0.256	0.256	0.170	0.170
Extrm asym decay					0.999	0.999	0.999	0.999
Extrm bimodal	0.897	0.897	0.852	0.852	0.795	0.795	0.726	0.726
Multi-modal lumpy	0.668	0.668	0.559	0.559	0.446	0.446	0.334	0.334
Normal	0.257	0.257	0.169	0.169	0.102	0.102	0.058	0.058
Smooth sym	0.165	0.165	0.104	0.104	0.061	0.061	0.034	0.034
Uni	0.330	0.330	0.230	0.230	0.150	0.150	0.090	0.090
Siegel-Tukey								
Asym Growth	0.886	0.886	0.815	0.816	0.703	0.706	0.614	0.616
Digit pref	0.512	0.513	0.389	0.389	0.258	0.261	0.184	0.186
Disc mass zero	0.568	0.569	0.446	0.447	0.308	0.310	0.225	0.227
Disc mass zero gap								
Exponential	0.830	0.830	0.735	0.735	0.603	0.605	0.502	0.504
Extrm asym decay					0.999	0.999	0.999	0.999
Extrm bimodal							0.999	0.999
Multi-modal lumpy	0.846	0.846	0.758	0.758	0.630	0.632	0.531	0.533
Normal	0.495	0.495	0.370	0.370	0.240	0.242	0.169	0.170
Smooth sym	0.550	0.550	0.425	0.426	0.288	0.290	0.210	0.212
Uni	0.750	0.750	0.639	0.639	0.494	0.496	0.394	0.397

Note: For Mood-Westenberg, A = asymptotic Chi-squared probability, E = Fisher exact probability; for Siegel-Tukey, A = asymptotic *Z*-score probability, E = Mann-Whitney-U exact probability

Table 5. Type-II errors/power rates for Mood-Westenberg and Siegel-Tukey, one-tailed directional test, for various *α* levels and sample size of (90, 90) when sampling is from all distributions/data sets, 100,000 repetitions, means are equal, and variance change is 1.5

	α							
	0.050		**0.025**		**0.010**		**0.005**	
Distribution	**A**	**E**	**A**	**E**	**A**	**E**	**A**	**E**
Mood-Westenberg								
Asym Growth	0.888	0.888	0.827	0.827	0.746	0.746	0.651	0.651
Digit pref	0.570	0.570	0.458	0.458	0.349	0.349	0.250	0.250
Disc mass zero	0.615	0.615	0.515	0.515	0.416	0.416	0.322	0.322
Disc mass zero gap			0.999	0.999	0.997	0.997	0.991	0.991
Exponential	0.916	0.916	0.861	0.861	0.787	0.787	0.692	0.692
Extrm asym decay								
Extrm bimodal	0.897	0.897	0.851	0.851	0.794	0.794	0.726	0.726
Multi-modal lumpy	0.971	0.971	0.946	0.946	0.906	0.906	0.849	0.849
Normal	0.643	0.643	0.527	0.527	0.407	0.407	0.293	0.293
Smooth sym	0.651	0.651	0.543	0.543	0.433	0.433	0.328	0.328
Uni	0.776	0.776	0.678	0.678	0.567	0.567	0.449	0.449
Siegel-Tukey								
Asym Growth	0.997	0.997	0.994	0.994	0.983	0.983	0.969	0.970
Digit pref	0.896	0.896	0.829	0.830	0.720	0.722	0.630	0.633
Disc mass zero	0.894	0.894	0.826	0.826	0.715	0.717	0.625	0.628
Disc mass zero gap								
Exponential	0.995	0.995	0.988	0.988	0.970	0.970	0.948	0.949
Extrm asym decay								
Extrm bimodal							0.999	0.999
Multi-modal lumpy	0.998	0.998	0.996	0.996	0.987	0.988	0.977	0.978
Normal	0.899	0.899	0.831	0.831	0.721	0.722	0.629	0.631
Smooth sym	0.902	0.902	0.835	0.836	0.729	0.732	0.641	0.644
Uni	0.988	0.988	0.974	0.974	0.942	0.943	0.907	0.908

Note: For Mood-Westenberg, A = asymptotic Chi-squared probability, E = Fisher exact probability; for Siegel-Tukey, A = asymptotic *Z*-score probability, E = Mann-Whitney-U exact probability

Table 6. Type-II errors/power rates for Mood-Westenberg and Siegel-Tukey, one-tailed directional test, for various *α* levels and sample size of (90, 90) when sampling is from all distributions/data sets, 100,000 repetitions, means are equal, and variance change is 2.25

	α							
	0.050		0.025		0.010		0.005	
Distribution	**A**	**E**	**A**	**E**	**A**	**E**	**A**	**E**
Mood-Westenberg								
Asym Growth								
Digit pref	0.985	0.985	0.971	0.971	0.948	0.948	0.913	0.913
Disc mass zero	0.990	0.990	0.981	0.981	0.965	0.965	0.940	0.940
Disc mass zero gap			0.999	0.999	0.996	0.996	0.990	0.990
Exponential							0.999	0.999
Extrm asym decay								
Extrm bimodal								
Multi-modal lumpy								
Normal	0.995	0.995	0.988	0.988	0.976	0.976	0.953	0.953
Smooth sym	0.985	0.985	0.970	0.970	0.946	0.946	0.909	0.909
Uni	0.999	0.999	0.998	0.998	0.996	0.996	0.990	0.990
Siegel-Tukey								
Asym Growth								
Digit pref					0.999	0.999	0.997	0.997
Disc mass zero					0.999	0.999	0.999	0.999
Disc mass zero gap								
Exponential								
Extrm asym decay								
Extrm bimodal								
Multi-modal lumpy								
Normal					0.999	0.999	0.999	0.999
Smooth sym							0.999	0.999
Uni								

Note: For Mood-Westenberg, A = asymptotic Chi-squared probability, E = Fisher exact probability; for Siegel-Tukey, A = asymptotic *Z*-score probability, E = Mann-Whitney-U exact probability

Table 7. Type-II errors/power rates for Mood-Westenberg and Siegel-Tukey, one-tailed directional test, for various α levels and sample size of (90, 90) when sampling is from all distributions/data sets, 100,000 repetitions, variances are equal, and means shift is 0.06

Mood-Westenberg

	α							
	0.050		**0.025**		**0.010**		**0.005**	
Distribution	**A**	**E**	**A**	**E**	**A**	**E**	**A**	**E**
Asym Growth	0.240	0.240	0.163	0.163	0.105	0.105	0.063	0.063
Digit pref	0.063	0.063	0.031	0.031	0.014	0.014	0.006	0.006
Disc mass zero	0.073	0.073	0.039	0.039	0.019	0.019	0.009	0.009
Disc mass zero gap			0.999	0.999	0.996	0.996	0.991	0.991
Exponential	0.071	0.071	0.037	0.037	0.018	0.018	0.008	0.008
Extrm asym decay			0.999	0.999	0.998	0.998	0.997	0.997
Extrm bimodal	0.537	0.537	0.459	0.459	0.383	0.383	0.310	0.310
Multi-modal lumpy	0.060	0.060	0.030	0.030	0.014	0.014	0.006	0.006
Normal	0.053	0.053	0.025	0.025	0.011	0.011	0.005	0.005
Smooth sym	0.065	0.065	0.033	0.033	0.015	0.015	0.007	0.007
Uni	0.052	0.052	0.025	0.025	0.010	0.010	0.004	0.004

Siegel-Tukey

Distribution	0.050 A	0.050 E	0.025 A	0.025 E	0.010 A	0.010 E	0.005 A	0.005 E
Asym Growth	0.298	0.298	0.198	0.198	0.111	0.112	0.071	0.072
Digit pref	0.050	0.050	0.025	0.026	0.010	0.011	0.005	0.005
Disc mass zero	0.040	0.040	0.020	0.020	0.008	0.008	0.004	0.004
Disc mass zero gap	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Exponential	0.011	0.011	0.005	0.005	0.001	0.001	0.001	0.001
Extrm asym decay	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Extrm bimodal	0.056	0.056	0.031	0.031	0.014	0.014	0.007	0.007
Multi-modal lumpy	0.038	0.038	0.018	0.018	0.007	0.007	0.003	0.003
Normal	0.050	0.050	0.025	0.025	0.010	0.010	0.005	0.005
Smooth sym	0.050	0.050	0.025	0.025	0.010	0.010	0.005	0.005
Uni	0.048	0.048	0.024	0.024	0.010	0.010	0.005	0.005

Note: For Mood-Westenberg, A = asymptotic Chi-squared probability, E = Fisher exact probability; for Siegel-Tukey, A = asymptotic *Z*-score probability, E = Mann-Whitney-U exact probability

Table 8. Power rates for one-tailed directional test for digit preference data set, various means shifts and variance changes for sample size (20, 20), 100,000 repetitions, $\alpha = 0.05$

Mood-Westenberg Chi-squared

Means shift	Variance change 1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
0.00	0.024	0.053	0.119	0.195	0.307	0.416	0.495	0.578	0.676	0.728	0.768
0.01	0.028	0.054	0.117	0.205	0.307	0.414	0.495	0.579	0.675	0.727	0.766
0.02	0.027	0.049	0.119	0.203	0.305	0.413	0.499	0.578	0.676	0.731	0.766
0.03	0.027	0.051	0.113	0.205	0.307	0.415	0.501	0.570	0.675	0.732	0.768
0.04	0.026	0.051	0.114	0.202	0.307	0.408	0.504	0.569	0.676	0.727	0.769
0.05	0.027	0.055	0.112	0.201	0.306	0.408	0.505	0.568	0.675	0.728	0.769
0.06	0.027	0.055	0.112	0.199	0.302	0.409	0.503	0.620	0.676	0.722	0.766
0.07	0.026	0.055	0.112	0.200	0.301	0.402	0.501	0.620	0.674	0.726	0.773
0.08	0.027	0.057	0.113	0.196	0.302	0.401	0.499	0.620	0.674	0.724	0.773
0.09	0.027	0.057	0.115	0.197	0.301	0.404	0.499	0.621	0.674	0.720	0.771
0.10	0.027	0.057	0.117	0.198	0.301	0.427	0.500	0.622	0.675	0.723	0.771
0.11	0.027	0.058	0.119	0.200	0.302	0.429	0.498	0.623	0.678	0.721	0.774
0.12	0.026	0.057	0.119	0.199	0.303	0.429	0.498	0.622	0.679	0.717	0.773

Siegel-Tukey *Z*-score

Means shift	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
0.00	0.048	0.177	0.366	0.535	0.687	0.789	0.849	0.897	0.933	0.954	0.963
0.01	0.050	0.179	0.362	0.543	0.687	0.788	0.849	0.897	0.932	0.953	0.963
0.02	0.050	0.168	0.363	0.540	0.686	0.788	0.853	0.896	0.933	0.948	0.963
0.03	0.050	0.168	0.354	0.543	0.688	0.789	0.853	0.897	0.933	0.949	0.964
0.04	0.049	0.169	0.355	0.524	0.690	0.794	0.853	0.897	0.934	0.947	0.964
0.05	0.049	0.179	0.352	0.527	0.685	0.792	0.855	0.897	0.932	0.947	0.964
0.06	0.049	0.177	0.352	0.525	0.673	0.793	0.855	0.906	0.933	0.947	0.964
0.07	0.049	0.178	0.351	0.521	0.671	0.774	0.855	0.904	0.932	0.949	0.966
0.08	0.050	0.185	0.354	0.525	0.669	0.774	0.843	0.907	0.932	0.948	0.965
0.09	0.050	0.186	0.356	0.526	0.672	0.773	0.842	0.906	0.931	0.954	0.965
0.10	0.050	0.185	0.357	0.528	0.670	0.779	0.843	0.895	0.933	0.954	0.964
0.11	0.050	0.186	0.361	0.535	0.671	0.780	0.844	0.893	0.929	0.954	0.965
0.12	0.050	0.184	0.362	0.534	0.670	0.782	0.841	0.896	0.931	0.948	0.965

Table 9. Power rates for one-tailed directional test for digit preference data set, various means shifts and variance changes for sample size (30, 30), 100,000 repetitions, $\alpha = 0.05$

	Mood-Westenberg Chi-squared										
Means	Variance change										
shift	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
0.00	0.065	0.145	0.292	0.441	0.613	0.743	0.817	0.874	0.930	0.951	0.965
0.01	0.073	0.142	0.291	0.457	0.611	0.744	0.813	0.873	0.930	0.951	0.965
0.02	0.073	0.135	0.291	0.458	0.615	0.742	0.821	0.875	0.930	0.952	0.964
0.03	0.073	0.134	0.279	0.456	0.612	0.743	0.820	0.867	0.931	0.953	0.965
0.04	0.072	0.133	0.281	0.454	0.612	0.730	0.821	0.866	0.929	0.953	0.965
0.05	0.072	0.141	0.276	0.452	0.614	0.730	0.820	0.866	0.927	0.954	0.965
0.06	0.073	0.143	0.278	0.451	0.611	0.730	0.821	0.904	0.931	0.949	0.964
0.07	0.073	0.142	0.278	0.454	0.607	0.727	0.819	0.905	0.930	0.950	0.966
0.08	0.073	0.151	0.281	0.445	0.611	0.727	0.818	0.902	0.930	0.949	0.967
0.09	0.075	0.150	0.280	0.444	0.613	0.727	0.819	0.903	0.931	0.948	0.967
0.10	0.074	0.150	0.292	0.443	0.610	0.761	0.819	0.908	0.930	0.948	0.967
0.11	0.073	0.148	0.292	0.447	0.611	0.760	0.818	0.907	0.933	0.948	0.967
0.12	0.073	0.153	0.292	0.443	0.610	0.762	0.818	0.908	0.932	0.948	0.967
	Siegel-Tukey *Z*-score										
0.00	0.047	0.237	0.498	0.706	0.849	0.922	0.957	0.977	0.989	0.994	0.996
0.01	0.051	0.234	0.498	0.711	0.849	0.922	0.957	0.978	0.989	0.994	0.996
0.02	0.051	0.219	0.498	0.714	0.849	0.922	0.958	0.976	0.988	0.992	0.996
0.03	0.053	0.217	0.482	0.710	0.849	0.922	0.960	0.978	0.989	0.993	0.996
0.04	0.051	0.218	0.482	0.690	0.849	0.926	0.959	0.976	0.989	0.992	0.996
0.05	0.051	0.230	0.479	0.690	0.849	0.925	0.958	0.977	0.988	0.993	0.996
0.06	0.050	0.229	0.482	0.692	0.835	0.926	0.959	0.980	0.989	0.993	0.996
0.07	0.052	0.232	0.479	0.691	0.832	0.913	0.959	0.980	0.988	0.992	0.997
0.08	0.052	0.247	0.483	0.694	0.834	0.912	0.952	0.980	0.988	0.992	0.996
0.09	0.051	0.244	0.481	0.695	0.835	0.912	0.952	0.980	0.988	0.994	0.997
0.10	0.052	0.246	0.489	0.691	0.835	0.915	0.951	0.975	0.989	0.994	0.996
0.11	0.053	0.245	0.488	0.703	0.834	0.915	0.951	0.975	0.989	0.994	0.997
0.12	0.051	0.242	0.490	0.699	0.833	0.915	0.951	0.976	0.988	0.992	0.996

Table 10. Power rates for one-tailed directional test for uniform distribution, various means shifts and variance changes for sample size (20, 20), 100,000 repetitions, $\alpha = 0.05$

Mood-Westenberg Chi-squared

Means shift	Variance change 1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
0.00	0.026	0.067	0.179	0.331	0.480	0.610	0.714	0.791	0.846	0.885	0.914
0.01	0.024	0.068	0.182	0.330	0.484	0.610	0.715	0.790	0.845	0.887	0.915
0.02	0.025	0.069	0.180	0.333	0.484	0.610	0.712	0.789	0.846	0.884	0.914
0.03	0.025	0.067	0.181	0.330	0.485	0.613	0.714	0.790	0.845	0.884	0.913
0.04	0.026	0.068	0.180	0.331	0.484	0.612	0.715	0.791	0.846	0.885	0.913
0.05	0.026	0.068	0.180	0.331	0.481	0.609	0.712	0.791	0.845	0.885	0.915
0.06	0.026	0.067	0.182	0.330	0.482	0.613	0.712	0.791	0.843	0.886	0.914
0.07	0.025	0.069	0.179	0.331	0.481	0.612	0.717	0.791	0.845	0.885	0.914
0.08	0.026	0.068	0.182	0.330	0.483	0.611	0.714	0.790	0.846	0.886	0.914
0.09	0.026	0.069	0.179	0.329	0.482	0.611	0.713	0.790	0.844	0.883	0.914
0.10	0.026	0.069	0.178	0.332	0.482	0.612	0.711	0.789	0.844	0.885	0.914
0.11	0.026	0.068	0.182	0.332	0.484	0.613	0.714	0.789	0.844	0.887	0.914
0.12	0.025	0.068	0.179	0.332	0.481	0.611	0.715	0.788	0.844	0.884	0.916

Siegel-Tukey *Z*-score

Means shift	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
0.00	0.048	0.272	0.548	0.745	0.859	0.922	0.955	0.973	0.984	0.989	0.994
0.01	0.046	0.272	0.548	0.744	0.860	0.922	0.955	0.973	0.984	0.990	0.994
0.02	0.048	0.273	0.548	0.746	0.861	0.921	0.955	0.974	0.984	0.989	0.993
0.03	0.047	0.269	0.549	0.745	0.861	0.922	0.955	0.974	0.985	0.990	0.993
0.04	0.048	0.272	0.547	0.746	0.860	0.921	0.956	0.974	0.984	0.990	0.993
0.05	0.048	0.272	0.549	0.745	0.859	0.922	0.955	0.973	0.984	0.990	0.994
0.06	0.049	0.270	0.547	0.743	0.858	0.922	0.956	0.974	0.984	0.990	0.993
0.07	0.048	0.269	0.545	0.745	0.860	0.923	0.955	0.974	0.985	0.990	0.993
0.08	0.047	0.273	0.547	0.745	0.859	0.920	0.955	0.974	0.983	0.990	0.993
0.09	0.048	0.271	0.546	0.743	0.859	0.921	0.955	0.973	0.983	0.990	0.994
0.10	0.046	0.269	0.545	0.745	0.859	0.922	0.954	0.973	0.983	0.990	0.993
0.11	0.047	0.266	0.545	0.743	0.859	0.922	0.956	0.974	0.984	0.990	0.993
0.12	0.047	0.267	0.545	0.744	0.857	0.923	0.955	0.973	0.983	0.990	0.994

Table 11. Power rates for one-tailed directional test for uniform distribution, various means shifts and variance changes for sample size (45, 45), 100,000 repetitions, $\alpha = 0.05$

Mood-Westenberg Chi-squared

Means shift	Variance change 1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
0.00	0.043	0.174	0.468	0.730	0.882	0.952	0.982	0.993	0.997	0.999	
0.01	0.044	0.176	0.468	0.730	0.882	0.953	0.983	0.993	0.997	0.999	0.999
0.02	0.044	0.175	0.468	0.731	0.884	0.952	0.981	0.993	0.997	0.999	0.999
0.03	0.044	0.174	0.465	0.731	0.883	0.953	0.981	0.993	0.997	0.999	
0.04	0.043	0.174	0.470	0.733	0.881	0.952	0.981	0.993	0.997	0.999	0.999
0.05	0.044	0.172	0.469	0.731	0.882	0.952	0.981	0.993	0.997	0.999	0.999
0.06	0.044	0.173	0.468	0.731	0.883	0.953	0.981	0.993	0.997	0.999	
0.07	0.043	0.175	0.465	0.731	0.883	0.953	0.981	0.993	0.997	0.999	
0.08	0.044	0.176	0.467	0.732	0.883	0.953	0.982	0.993	0.997	0.999	
0.09	0.042	0.174	0.469	0.730	0.883	0.952	0.981	0.992	0.997	0.999	
0.10	0.044	0.174	0.467	0.732	0.883	0.952	0.981	0.993	0.997	0.999	0.999
0.11	0.044	0.175	0.468	0.730	0.882	0.953	0.981	0.993	0.997	0.999	0.999
0.12	0.045	0.171	0.466	0.729	0.881	0.953	0.982	0.993	0.997	0.999	

Siegel-Tukey *Z*-score

Means shift	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
0.00	0.049	0.493	0.865	0.972	0.995	0.999					
0.01	0.050	0.493	0.863	0.973	0.995	0.999					
0.02	0.050	0.492	0.865	0.972	0.995	0.999					
0.03	0.050	0.492	0.862	0.973	0.995	0.999					
0.04	0.050	0.493	0.864	0.973	0.994	0.999					
0.05	0.050	0.491	0.866	0.972	0.995	0.999					
0.06	0.050	0.491	0.862	0.972	0.995	0.999					
0.07	0.049	0.491	0.862	0.972	0.994	0.999					
0.08	0.050	0.491	0.863	0.972	0.995	0.999					
0.09	0.048	0.489	0.863	0.973	0.995	0.999					
0.10	0.050	0.488	0.862	0.971	0.995	0.999					
0.11	0.049	0.491	0.862	0.973	0.995	0.999					
0.12	0.049	0.486	0.861	0.972	0.995	0.999					

Table 12. Power rates for one-tailed directional test for uniform distribution, various means shifts and variance changes for sample size (65, 65), 100,000 repetitions, $\alpha = 0.05$

Mood-Westenberg Chi-squared

Means	Variance change										
shift	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
0.00	0.043	0.230	0.609	0.867	0.964	0.992	0.998				
0.01	0.042	0.229	0.611	0.868	0.964	0.991	0.998	0.999			
0.02	0.043	0.230	0.612	0.865	0.964	0.991	0.998	0.999			
0.03	0.044	0.232	0.610	0.867	0.963	0.991	0.998	0.999			
0.04	0.042	0.230	0.612	0.869	0.964	0.991	0.998				
0.05	0.043	0.232	0.611	0.867	0.965	0.991	0.998				
0.06	0.042	0.232	0.610	0.867	0.964	0.991	0.998				
0.07	0.041	0.229	0.611	0.867	0.965	0.992	0.998	0.999			
0.08	0.043	0.229	0.613	0.868	0.965	0.991	0.998				
0.09	0.043	0.230	0.613	0.867	0.965	0.991	0.998				
0.10	0.042	0.232	0.613	0.866	0.964	0.992	0.998				
0.11	0.043	0.228	0.612	0.867	0.964	0.991	0.998				
0.12	0.041	0.229	0.611	0.867	0.965	0.992	0.998	0.999			

Siegel-Tukey *Z*-score

0.00	0.050	0.623	0.951	0.996							
0.01	0.048	0.623	0.952	0.996							
0.02	0.050	0.626	0.952	0.996							
0.03	0.050	0.627	0.951	0.996							
0.04	0.049	0.626	0.953	0.996							
0.05	0.049	0.625	0.952	0.996							
0.06	0.050	0.623	0.951	0.996							
0.07	0.048	0.622	0.951	0.996							
0.08	0.049	0.625	0.951	0.996							
0.09	0.049	0.623	0.952	0.996							
0.10	0.049	0.623	0.951	0.996							
0.11	0.049	0.620	0.951	0.996							
0.12	0.050	0.620	0.950	0.996							

Table 13. Power rates for one-tailed directional test for discrete mass zero with gap data set, various means shifts and variance changes for sample size (45, 45), 100,000 repetitions, $\alpha = 0.05$

Mood-Westenberg Chi-squared

Means shift	Variance change										
	1.00	**1.25**	**1.50**	**1.75**	**2.00**	**2.25**	**2.50**	**2.75**	**3.00**	**3.25**	**3.50**
0.00	0.004	0.959	0.960	0.960	0.961	0.957	0.957	0.956	0.955	0.955	0.957
0.01	0.960	0.960	0.961	0.960	0.961	0.956	0.956	0.957	0.956	0.956	0.957
0.02	0.961	0.960	0.961	0.959	0.961	0.961	0.956	0.956	0.957	0.957	0.955
0.03	0.961	0.961	0.960	0.961	0.960	0.960	0.957	0.956	0.957	0.957	0.956
0.04	0.960	0.961	0.960	0.960	0.959	0.960	0.957	0.955	0.955	0.956	0.957
0.05	0.961	0.960	0.960	0.961	0.960	0.959	0.956	0.957	0.956	0.956	0.957
0.06	0.960	0.960	0.961	0.961	0.961	0.960	0.956	0.956	0.957	0.955	0.956
0.07	0.960	0.960	0.960	0.961	0.961	0.960	0.956	0.956	0.956	0.956	0.956
0.08	0.961	0.961	0.960	0.960	0.961	0.959	0.961	0.955	0.957	0.955	0.956
0.09	0.960	0.961	0.960	0.959	0.960	0.961	0.961	0.956	0.955	0.957	0.956
0.10	0.961	0.960	0.960	0.961	0.961	0.961	0.960	0.956	0.955	0.956	0.956
0.11	0.960	0.961	0.960	0.960	0.960	0.960	0.961	0.955	0.957	0.957	0.956
0.12	0.961	0.961	0.961	0.960	0.961	0.960	0.960	0.957	0.957	0.957	0.956

Siegel-Tukey *Z*-score

Means shift	**1.00**	**1.25**	**1.50**	**1.75**	**2.00**	**2.25**	**2.50**	**2.75**	**3.00**	**3.25**	**3.50**
0.00	0.001	0.997	0.996	0.996	0.997	0.996	0.996	0.996	0.996	0.996	0.996
0.01	0.000	0.997	0.997	0.997	0.997	0.996	0.996	0.996	0.996	0.996	0.996
0.02	0.000	0.997	0.996	0.996	0.997	0.997	0.996	0.996	0.996	0.996	0.996
0.03	0.000	0.997	0.996	0.997	0.996	0.996	0.996	0.996	0.996	0.996	0.996
0.04	0.000	0.997	0.997	0.997	0.997	0.996	0.996	0.996	0.996	0.996	0.996
0.05	0.000	0.997	0.997	0.997	0.997	0.996	0.996	0.996	0.996	0.996	0.996
0.06	0.000	0.997	0.997	0.997	0.997	0.997	0.996	0.996	0.997	0.996	0.996
0.07	0.000	0.996	0.997	0.997	0.997	0.997	0.996	0.996	0.996	0.996	0.996
0.08	0.000	0.997	0.997	0.997	0.997	0.997	0.996	0.996	0.996	0.996	0.996
0.09	0.000	0.996	0.997	0.996	0.997	0.997	0.997	0.996	0.996	0.996	0.996
0.10	0.000	0.996	0.997	0.997	0.997	0.996	0.997	0.996	0.996	0.996	0.996
0.11	0.000	0.997	0.996	0.997	0.997	0.996	0.997	0.996	0.996	0.996	0.996
0.12	0.000	0.997	0.997	0.997	0.997	0.996	0.996	0.996	0.996	0.996	0.996

Table 14. Power rates for one-tailed directional test for discrete mass zero with gap data set, various means shifts and variance changes for sample size (10, 10), 100,000 repetitions, $\alpha = 0.05$

	Mood-Westenberg Chi-squared										
Means	**Variance change**										
shift	**1.00**	**1.25**	**1.50**	**1.75**	**2.00**	**2.25**	**2.50**	**2.75**	**3.00**	**3.25**	**3.50**
0.00	0.005	0.310	0.316	0.314	0.315	0.308	0.307	0.309	0.308	0.307	0.308
0.01	0.248	0.308	0.314	0.315	0.314	0.310	0.309	0.308	0.308	0.310	0.305
0.02	0.247	0.310	0.316	0.318	0.314	0.313	0.307	0.308	0.308	0.308	0.309
0.03	0.249	0.309	0.313	0.315	0.316	0.315	0.309	0.311	0.311	0.311	0.310
0.04	0.246	0.310	0.316	0.316	0.315	0.316	0.309	0.309	0.309	0.311	0.310
0.05	0.246	0.310	0.317	0.312	0.314	0.315	0.308	0.310	0.308	0.309	0.307
0.06	0.248	0.311	0.315	0.317	0.312	0.316	0.310	0.308	0.306	0.306	0.309
0.07	0.246	0.313	0.316	0.317	0.315	0.313	0.308	0.309	0.305	0.309	0.309
0.08	0.245	0.311	0.314	0.314	0.317	0.314	0.315	0.309	0.306	0.306	0.308
0.09	0.249	0.312	0.315	0.314	0.315	0.312	0.315	0.308	0.309	0.309	0.311
0.10	0.244	0.313	0.316	0.315	0.316	0.315	0.315	0.310	0.308	0.311	0.309
0.11	0.247	0.311	0.315	0.314	0.314	0.317	0.313	0.307	0.311	0.310	0.311
0.12	0.247	0.308	0.314	0.315	0.314	0.315	0.314	0.310	0.312	0.310	0.308
	Siegel-Tukey *Z*-score										
0.00	0.000	0.619	0.620	0.619	0.619	0.612	0.610	0.611	0.612	0.610	0.611
0.01	0.000	0.617	0.620	0.624	0.621	0.614	0.612	0.613	0.611	0.612	0.609
0.02	0.000	0.617	0.623	0.622	0.621	0.623	0.612	0.610	0.611	0.610	0.611
0.03	0.000	0.619	0.621	0.624	0.623	0.620	0.612	0.613	0.613	0.615	0.615
0.04	0.000	0.619	0.623	0.621	0.620	0.622	0.613	0.613	0.611	0.611	0.610
0.05	0.000	0.619	0.623	0.621	0.619	0.620	0.610	0.612	0.612	0.613	0.612
0.06	0.000	0.621	0.622	0.623	0.621	0.624	0.613	0.611	0.612	0.610	0.611
0.07	0.000	0.622	0.622	0.623	0.620	0.620	0.613	0.612	0.609	0.610	0.612
0.08	0.000	0.619	0.623	0.622	0.622	0.620	0.620	0.613	0.612	0.610	0.609
0.09	0.000	0.620	0.620	0.622	0.619	0.621	0.623	0.612	0.615	0.613	0.614
0.10	0.000	0.623	0.621	0.621	0.622	0.624	0.623	0.612	0.613	0.614	0.612
0.11	0.000	0.621	0.622	0.621	0.622	0.621	0.618	0.608	0.614	0.615	0.613
0.12	0.000	0.618	0.622	0.620	0.621	0.621	0.622	0.613	0.614	0.614	0.613

Table 15. Power rates for one-tailed directional test for discrete mass zero with gap data set, various means shifts and variance changes for sample size (10, 10), 100,000 repetitions, α = 0.05

Mood-Westenberg Chi-squared

Means shift	Variance change 1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
0.00	0.022	0.701	0.705	0.700							
0.01	0.347	0.699	0.701	0.701							
0.02	0.347	0.699	0.702	0.701							
0.03	0.349	0.703	0.701	0.704							
0.04	0.349	0.701	0.701	0.699							
0.05	0.349	0.700	0.701	0.701							
0.06	0.345	0.700	0.701	0.702							
0.07	0.346	0.701	0.703	0.703							
0.08	0.348	0.700	0.703	0.701							
0.09	0.347	0.702	0.702	0.700							
0.10	0.349	0.699	0.701	0.702							
0.11	0.350	0.702	0.702	0.702							
0.12	0.346	0.702	0.702	0.702							

Siegel-Tukey *Z*-score

Means shift	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
0.00	0.023	0.991	0.991	0.992							
0.01	0.055	0.991	0.991	0.992							
0.02	0.054	0.991	0.992	0.992							
0.03	0.055	0.991	0.992	0.992							
0.04	0.054	0.991	0.991	0.992							
0.05	0.054	0.991	0.991	0.992							
0.06	0.055	0.991	0.991	0.992							
0.07	0.054	0.991	0.991	0.992							
0.08	0.054	0.991	0.991	0.992							
0.09	0.054	0.992	0.991	0.992							
0.10	0.053	0.992	0.991	0.992							
0.11	0.053	0.991	0.991	0.992							
0.12	0.052	0.992	0.991	0.992							

Table 16. Power rates for one-tailed directional test for multi-modal lumpy data set, various means shifts and variance changes for sample size (30, 30), 100,000 repetitions, $\alpha = 0.05$

Mood-Westenberg Chi-squared

Means shift	Variance change 1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
0.00	0.068	0.303	0.652	0.846	0.936	0.971	0.988	0.995	0.998	0.999	0.999
0.01	0.074	0.272	0.624	0.840	0.935	0.971	0.988	0.995	0.998	0.999	
0.02	0.072	0.273	0.623	0.841	0.924	0.969	0.988	0.995	0.998	0.999	0.999
0.03	0.072	0.266	0.623	0.840	0.923	0.970	0.988	0.995	0.998	0.999	0.999
0.04	0.073	0.266	0.625	0.823	0.922	0.969	0.988	0.995	0.998	0.999	0.999
0.05	0.073	0.261	0.590	0.823	0.925	0.968	0.988	0.994	0.998	0.999	0.999
0.06	0.074	0.263	0.591	0.817	0.923	0.967	0.987	0.994	0.998	0.999	0.999
0.07	0.071	0.258	0.590	0.818	0.925	0.968	0.985	0.994	0.997	0.999	0.999
0.08	0.074	0.258	0.590	0.817	0.924	0.968	0.985	0.994	0.997	0.998	0.999
0.09	0.080	0.247	0.592	0.814	0.923	0.968	0.985	0.994	0.998	0.999	0.999
0.10	0.078	0.249	0.587	0.805	0.914	0.966	0.985	0.994	0.998	0.999	0.999
0.11	0.079	0.221	0.589	0.804	0.915	0.966	0.984	0.993	0.997	0.999	0.999
0.12	0.077	0.221	0.586	0.798	0.914	0.965	0.984	0.994	0.997	0.999	0.999

Siegel-Tukey *Z*-score

Means shift	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
0.00	0.049	0.444	0.831	0.961	0.992	0.998	0.999				
0.01	0.043	0.430	0.812	0.956	0.992	0.998					
0.02	0.043	0.431	0.811	0.958	0.989	0.998	0.999				
0.03	0.043	0.418	0.812	0.957	0.989	0.998	0.999				
0.04	0.043	0.417	0.814	0.952	0.989	0.997	0.999				
0.05	0.044	0.399	0.788	0.953	0.989	0.997	0.999				
0.06	0.043	0.399	0.790	0.948	0.989	0.997	0.999				
0.07	0.042	0.388	0.789	0.949	0.989	0.997	0.999				
0.08	0.044	0.388	0.788	0.945	0.989	0.997	0.999				
0.09	0.031	0.376	0.792	0.943	0.989	0.997	0.999				
0.10	0.032	0.378	0.773	0.939	0.985	0.997	0.999				
0.11	0.032	0.357	0.774	0.940	0.985	0.997	0.999				
0.12	0.032	0.357	0.772	0.940	0.985	0.997	0.999				

Table 17. Power rates for one-tailed directional test for multi-modal lumpy data set, various means shifts and variance changes for sample size (65, 65), 100,000 repetitions, α = 0.05

Mood-Westenberg Chi-squared

Means shift	Variance change 1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
0.00	0.041	0.486	0.894	0.985	0.998						
0.01	0.047	0.408	0.866	0.985	0.998						
0.02	0.047	0.406	0.866	0.983	0.997						
0.03	0.047	0.389	0.865	0.983	0.997						
0.04	0.047	0.392	0.868	0.975	0.997						
0.05	0.047	0.404	0.839	0.975	0.997						
0.06	0.047	0.404	0.838	0.975	0.997						
0.07	0.048	0.409	0.839	0.975	0.997	0.999					
0.08	0.046	0.413	0.839	0.976	0.997						
0.09	0.058	0.376	0.836	0.977	0.997						
0.10	0.057	0.375	0.833	0.971	0.996						
0.11	0.057	0.302	0.833	0.971	0.996						
0.12	0.058	0.302	0.831	0.966	0.996						

Siegel-Tukey *Z*-score

Means shift	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
0.00	0.050	0.727	0.988								
0.01	0.039	0.712	0.984								
0.02	0.039	0.711	0.984								
0.03	0.039	0.698	0.984								
0.04	0.040	0.695	0.983								
0.05	0.040	0.663	0.979	0.999							
0.06	0.040	0.664	0.978	0.999							
0.07	0.040	0.649	0.978	0.999							
0.08	0.038	0.651	0.978	0.999							
0.09	0.024	0.634	0.978	0.999							
0.10	0.024	0.634	0.973	0.999							
0.11	0.025	0.602	0.974	0.999							
0.12	0.024	0.600	0.973	0.999							

Table 18. Power rates for one-tailed directional test for exponential distribution, various means shifts and variance changes for sample size (20, 20), 100,000 repetitions, $\alpha = 0.05$

Mood-Westenberg Chi-squared

Means shift	Variance change 1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
0.00	0.026	0.094	0.273	0.464	0.617	0.725	0.794	0.845	0.879	0.899	0.916
0.01	0.026	0.092	0.264	0.458	0.609	0.721	0.794	0.844	0.879	0.899	0.916
0.02	0.026	0.085	0.258	0.448	0.606	0.716	0.793	0.843	0.875	0.901	0.914
0.03	0.027	0.081	0.248	0.439	0.602	0.713	0.791	0.841	0.875	0.902	0.915
0.04	0.025	0.077	0.240	0.435	0.596	0.710	0.789	0.840	0.876	0.899	0.916
0.05	0.028	0.072	0.234	0.426	0.592	0.707	0.786	0.842	0.874	0.901	0.917
0.06	0.029	0.069	0.226	0.421	0.584	0.704	0.783	0.839	0.873	0.899	0.916
0.07	0.029	0.066	0.220	0.411	0.578	0.698	0.781	0.835	0.874	0.900	0.918
0.08	0.030	0.063	0.212	0.404	0.569	0.693	0.779	0.835	0.873	0.899	0.915
0.09	0.032	0.059	0.204	0.400	0.565	0.693	0.778	0.835	0.873	0.897	0.917
0.10	0.034	0.055	0.197	0.392	0.562	0.685	0.774	0.831	0.871	0.899	0.915
0.11	0.035	0.053	0.191	0.382	0.555	0.683	0.771	0.830	0.869	0.897	0.914
0.12	0.037	0.051	0.186	0.375	0.550	0.677	0.769	0.828	0.869	0.900	0.915

Siegel-Tukey *Z*-score

Means shift	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
0.00	0.049	0.312	0.601	0.777	0.875	0.929	0.956	0.974	0.983	0.988	0.991
0.01	0.042	0.305	0.591	0.774	0.872	0.925	0.957	0.973	0.983	0.988	0.991
0.02	0.040	0.294	0.581	0.768	0.872	0.927	0.955	0.972	0.982	0.987	0.991
0.03	0.035	0.283	0.573	0.763	0.871	0.924	0.957	0.972	0.981	0.988	0.991
0.04	0.030	0.270	0.568	0.761	0.866	0.924	0.956	0.972	0.982	0.988	0.991
0.05	0.029	0.257	0.559	0.754	0.868	0.923	0.955	0.973	0.982	0.988	0.991
0.06	0.025	0.249	0.549	0.749	0.863	0.922	0.953	0.971	0.981	0.987	0.991
0.07	0.022	0.238	0.542	0.746	0.860	0.921	0.953	0.972	0.983	0.987	0.991
0.08	0.020	0.226	0.531	0.741	0.855	0.921	0.953	0.971	0.981	0.987	0.991
0.09	0.018	0.217	0.520	0.735	0.853	0.919	0.952	0.971	0.980	0.987	0.991
0.10	0.016	0.207	0.512	0.730	0.853	0.915	0.951	0.970	0.982	0.988	0.991
0.11	0.014	0.198	0.504	0.727	0.847	0.914	0.950	0.969	0.981	0.987	0.991
0.12	0.013	0.189	0.494	0.718	0.847	0.913	0.949	0.969	0.981	0.987	0.991

Table 19. Power rates for one-tailed directional test for exponential distribution, various means shifts and variance changes for sample size (30, 30), 100,000 repetitions, $\alpha = 0.05$

Means shift	Variance change 1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
Mood-Westenberg Chi-squared											
0.00	0.069	0.241	0.553	0.782	0.899	0.951	0.976	0.988	0.994	0.996	0.997
0.01	0.069	0.232	0.543	0.772	0.896	0.951	0.975	0.988	0.993	0.996	0.997
0.02	0.071	0.222	0.532	0.765	0.890	0.949	0.975	0.987	0.993	0.996	0.997
0.03	0.071	0.212	0.518	0.760	0.887	0.947	0.974	0.987	0.993	0.996	0.997
0.04	0.073	0.204	0.506	0.752	0.885	0.944	0.973	0.986	0.993	0.996	0.997
0.05	0.075	0.195	0.496	0.745	0.879	0.944	0.972	0.986	0.992	0.996	0.997
0.06	0.078	0.183	0.484	0.736	0.876	0.941	0.972	0.986	0.992	0.996	0.997
0.07	0.081	0.176	0.475	0.729	0.872	0.938	0.970	0.985	0.992	0.996	0.997
0.08	0.084	0.166	0.459	0.720	0.866	0.938	0.969	0.984	0.992	0.996	0.997
0.09	0.087	0.158	0.451	0.713	0.865	0.935	0.969	0.984	0.991	0.995	0.997
0.10	0.092	0.150	0.440	0.705	0.858	0.932	0.967	0.985	0.991	0.995	0.997
0.11	0.097	0.143	0.428	0.697	0.852	0.933	0.968	0.983	0.991	0.995	0.997
0.12	0.102	0.137	0.417	0.687	0.850	0.929	0.965	0.983	0.991	0.995	0.997
Siegel-Tukey *Z*-score											
0.00	0.049	0.428	0.768	0.917	0.970	0.989	0.995	0.998	0.999	0.999	
0.01	0.043	0.415	0.761	0.914	0.970	0.988	0.995	0.998	0.999		
0.02	0.038	0.398	0.755	0.910	0.968	0.988	0.995	0.998	0.999		
0.03	0.032	0.382	0.743	0.909	0.968	0.988	0.995	0.998	0.999		
0.04	0.029	0.369	0.734	0.904	0.966	0.987	0.995	0.998	0.999	0.999	
0.05	0.024	0.356	0.724	0.902	0.964	0.988	0.994	0.998	0.999		
0.06	0.021	0.336	0.720	0.897	0.963	0.986	0.995	0.997	0.999	0.999	
0.07	0.018	0.324	0.710	0.893	0.963	0.986	0.995	0.998	0.999		
0.08	0.016	0.306	0.700	0.891	0.961	0.986	0.995	0.998	0.999		
0.09	0.014	0.291	0.690	0.887	0.961	0.985	0.994	0.998	0.999	0.999	
0.10	0.012	0.275	0.678	0.882	0.958	0.985	0.994	0.998	0.999	0.999	
0.11	0.011	0.262	0.667	0.879	0.956	0.985	0.994	0.997	0.999		
0.12	0.009	0.249	0.658	0.875	0.955	0.984	0.994	0.998	0.999	0.999	