Journal of Modern Applied Statistical Methods

Volume 15 | Issue 2 Article 11

11-1-2016

Preliminary Tests of Normality When Comparing Three Independent Samples

Björn Lantz
Chalmers University of Technology, bjorn.lantz@chalmers.se

Roy Andersson
Jönköping University

Peter Manfredsson
University of Borås

Journal of Modern Applied Statistical Methods
November 2016, Vol. 15, No. 2, 135-148.
doi: 10.22237/jmasm/1478002140

ISSN 1538 − 9472

Preliminary Tests of Normality When Comparing Three Independent Samples

Björn Lantz
Chalmers University of Technology
Göteborg, Sweden

Roy Andersson
Jönköping University
Jönköping, Sweden

Peter Manfredsson
University of Borås
Borås, Sweden

This paper uses simulation to explore the performance of a two-stage procedure where a preliminary Shapiro-Wilk test is used to choose between the ANOVA and Kruskal-Wallis tests as a three-sample location test. The results suggest that the two-stage procedure actually seems to be preferable when conducting such location tests.

Keywords: Normality, assumptions, preliminary tests, ANOVA, Kruskal-Wallis, Shapiro-Wilk

Introduction

It is common among applied researchers in psychology to conduct data analyses as two-stage procedures where one or more preliminary tests precede the test of interest (Keselman, Othman, & Wilcox, 2013). For example, when a researcher plans to compare two population means with Student's *t*-test, the underlying normality assumption is often checked with a preliminary goodness-of-fit test. If the null hypothesis of normality is rejected, the Mann-Whitney test (or some other non-parametric test) is used to analyze the data. If the null hypothesis of normality is not rejected, the underlying homoscedasticity assumption may be checked in a similar manner. If the null hypothesis of homoscedasticity is rejected, Welch's *t*-test (or some other robust test) is used. If data were neither significantly non-normal nor significantly heteroscedastic, Student's *t*-test is used to compare the two means.

The normality assumption issue is highly relevant for data analyses in psychological research. For example, in the empirical study of achievement and psychometric measures conducted by Micceri (1989), significant non-normality contaminations were found in all 440 measures, including tail weights from the uniform to the double exponential, exponential level asymmetry, and bimodality. Furthermore, recent research has shown that most real data samples are at least

Dr. Lantz is an Associate Professor of Industrial Engineering in the Department of Technology Management and Economics. Email him at: bjorn.lantz@chalmers.se.

slightly non-normal in terms of skewness and kurtosis (Blanca, Arnau, López-Montiel, Bono, & Bendayan, 2013) and that the variance heterogeneity assumption is violated in a nontrivial number of published studies (Ruscio & Roche, 2012). However, there are several conceptual reasons why the use of a two-stage test procedure with a preliminary test of normality and/or the homoscedasticity assumption may be problematic in practice (Wells & Hinze, 2006):

- The probability of a type I error as well as a type II error in the procedure may be heavily distorted. This is because the distribution of the location test statistic is not only related to the parental distribution(s), but also conditional on the preliminary test since both type I errors and type II errors are possible in the first stage. For example, even if a parental population is significantly contaminated from the exponential distribution, many samples will not look non-normal enough to fail the normality test. This is because of the random component of the sampling procedure. However, the samples that pass the normality test will often be significantly different from the other samples, not only in terms of shape but also in terms of mean and/or standard deviation.
- A preliminary test in which the null hypothesis of normality is not rejected does not constitute proof that the normality assumption holds. In fact, no null hypothesis is strictly ever true when empirical data are considered (Cohen, 1994). From this perspective, normality assumptions are always violated.
- Preliminary test procedures rely on assumptions themselves. This means that, strictly speaking, those assumptions also need to be tested. This would however also require new assumptions, and so on, and so forth.
- Even though a preliminary test correctly indicates that a normality assumption does not hold, a parametric test with higher power than the corresponding non-parametric test might still be valid because of high robustness against the current type of non-normality.

Recently, the performance of different two-stage procedures, where samples are checked with preliminary tests of normality before univariate or bivariate location tests, have been studied. For example, Rochon and Kieser (2011) examined the type I error rate of the one-sample Student's *t*-test with a preliminary normality test. They found an increase in the type I error rate for conditional samples compared to unconditional ones, especially when parental distributions were

skewed. Schucany and Ng (2006) found similar results for the one-sample Student's *t*-test with a preliminary normality test, concluding that graphical diagnostics are probably better in practice than formal pretests. Rochon, Gondan, and Kieser (2012) examined a two-stage procedure where a preliminary normality test was used to decide between the two-sample Student's *t*-test and the Mann-Whitney test in the second stage. They concluded that even though the two-stage procedure might be considered incorrect from a formal perspective, the procedure seemed to satisfactorily maintain the nominal significance level and had acceptable power properties in the investigated examples. Rasch, Kubinger, and Moder (2011), on the other hand, found that it is preferable to use Welch's *t*-test without pre-testing for normality rather than the two-stage procedure including Student's *t*-test as a standard test, and that the corresponding non-parametric test should not be used in the given context. Preliminary tests have also recently been discussed in related contexts by, for example, Lantz (2013), Zimmerman (2004, 2011, 2014), Shuster (2005, 2009), and Schoder, Himmelmann, and Wilhelm (2006).

Overall, there seems to be a general consensus in the literature that two-stage procedures including preliminary tests are unnecessary at best, or harmful at worst, in a one-sample or a two-sample location test context. However, there does not seem to exist any similar literature based on simulated two-stage location tests for three (or more) groups, even though both Othman, Keselman, and Wilcox (2015) and Keselman, Othman, and Wilcox (2014) analyze the two-stage procedure problem itself in a multi-group context based on simulations. The focus in both papers is on the normality screening rather on the two-stage procedure as a whole, though.

We thus seek to answer the following question in this paper: what are the properties of a two-stage procedure where a normality test at the first stage is used to decide between the omnibus one-way ANOVA and the Kruskal-Wallis test in the second stage? The purpose of this paper is to present the results from a simulation study designed to shed light on this issue. In the next section the methodology of the study is described. The results of the simulations are then presented and discussed in relation to previous research. Finally, the paper concludes with the implications of these results for use in statistical analysis in practice.

Methodology

In the simulations, random samples from three independent groups were drawn from four different distributions, in line with the typical contaminations found by

Micceri (1989) in his empirical study of achievement and psychometric measures. The distributions used are also the ones typically used in this type of study (e.g., Rochon et al., 2012), that is, the normal, the uniform, the exponential, and the Laplace distributions. The uniform distribution represents a decent approximation of the normal distribution, while the exponential and the Laplace distributions represent two different types of distinct non-normality in terms of skewness and kurtosis. The normal distribution is included for the purpose of comparability.

The probability density function of the normal distribution is given by

$$\mathrm{f}\left(x\right)=\frac{\mathrm{e}^{-\left(x-\mu\right)^{2}/2\sigma^{2}}}{\sqrt{2\pi\sigma^{2}}} \tag{1}$$

with mean μ and variance σ^2. It has no skewness and by definition no excess kurtosis.

The probability density function of the uniform distribution is given by

$$\mathrm{f}\left(x\right)=\begin{cases}\dfrac{1}{b-a}, & a\leq x\leq b\\ 0, & \text{otherwise}\end{cases} \tag{2}$$

The uniform distribution is symmetric, like the normal distribution, and slightly platykurtic.

The probability density function of the exponential distribution is given by

$$\mathrm{f}\left(x\right)=\begin{cases}\lambda\mathrm{e}^{-\lambda x}, & x\geq 0\\ 0, & x<0\end{cases} \tag{3}$$

where λ is the rate parameter. It represents a distinct form of non-normality due to its heavy skewness to the right and its strong leptokurtic form. In reality, it can often approximate, e.g., the time between events or the time of events.

The probability density function of the Laplace distribution, finally, is given by

$$\mathrm{f}\left(x\right)=\frac{\mathrm{e}^{-\left|x-\mu\right|/b}}{2b} \tag{4}$$

which means that it is symmetric and significantly leptokurtic with an excess kurtosis of 3. At first glance, one might think that the Laplace distribution resembles the normal distribution. The huge difference, however, is that outliers are much more common due to the fatter tails. Hence, it represents an important form of non-normality where wild randomness exists (for realistic examples of such cases, see e.g., Mandelbrot & Taleb, 2006).

In the simulations, the standard deviation was kept constant at 1 for all distributions in all cases while the mean values were varied to accomplish five different effect sizes in order to evaluate actual significance as well as actual power. Table 1 shows the manner in which the true mean values of the distributions were shifted to achieve a suitable range of effect sizes (see Cohen, 1992), ranging from no effect ($f = 0.00$) to a very large effect ($f = 0.65$).

The simulated data sets for the three groups were subject to individual normality screening at various significance levels based on the Shapiro-Wilk test with the Royston algorithm (Royston, 1992), that is, the default algorithm in SPSS and other statistical software. The Shapiro-Wilk test has recently been found to have the best power among the tests commonly used for normality screening (Marmolejo-Ramos & González-Burgos, 2013; Razali & Wah, 2011), even though other researchers recommend other tests, such as the Anderson-Darling test (see Keselman et al., 2013), for normality screening.

If the normality hypothesis for at least one group was rejected, a location test was performed with the Kruskal-Wallis test (Kruskal & Wallis, 1952) at the 0.05 significance level. If not, a location test was performed with the omnibus one-way ANOVA at the 0.05 significance level. This two-stage procedure was repeated 100,000 times for each combination of effect size and distribution, and for three different sample sizes ($n = 15$, $n = 30$, and $n = 60$ in each group). All 100,000 data sets were also analyzed with the ANOVA without a preliminary test, as well as with the Kruskal-Wallis test without a preliminary test. This was done in both cases for each combination of effect size, distribution, and sample size. All simulation procedures were conducted using Microsoft Excel 2010.

Table 1. The different combinations of mean values

Effect size *f*	μ_1	μ_2	μ_3
0.00	0.000	0.000	0.000
0.10	0.000	0.123	0.246
0.25	0.000	0.307	0.614
0.40	0.000	0.490	0.980
0.65	0.000	0.796	1.592

Results

Estimated Type I Error Probabilities

This section presents the results when all samples were drawn from distributions with the same mean value. Table 2 displays the frequencies of significant tests out of the 100,000 conducted tests for the different combinations of test procedure (the ANOVA without the preliminary test, the Kruskal-Wallis test without the preliminary test, or the two-stage procedure), sample size (n = 15, n = 30, or n = 60 in each group) and distribution (normal, uniform, exponential, or Laplace). For example, the two-stage procedure (TSP) where the preliminary Shapiro-Wilk test for normality was conducted on a significance level of 0.1 yielded 4,970 significant tests when n = 30. Hence, the estimated type I error probability for this specific combination of distribution, test procedure, and sample size was 4.97% when samples were taken from exponential distributions.

Table 2. Estimated type I error probabilities

Distribution	Method	n = 15	n = 30	n = 60
Normal	ANOVA	4.99%	4.99%	4.93%
	Kruskal-Wallis	4.79%	4.86%	4.86%
	TSP 0.1	5.13%	5.17%	5.06%
	TSP 0.05	5.14%	5.15%	5.05%
	TSP 0.01	5.06%	5.06%	4.97%
	TSP 0.005	5.03%	5.03%	4.96%
Uniform	ANOVA	5.12%	5.10%	5.08%
	Kruskal-Wallis	4.72%	4.90%	4.99%
	TSP 0.1	4.88%	4.93%	4.99%
	TSP 0.05	5.00%	4.96%	4.99%
	TSP 0.01	5.09%	5.08%	5.01%
	TSP 0.005	5.11%	5.11%	5.03%
Exp	ANOVA	4.46%	4.59%	4.74%
	Kruskal-Wallis	4.75%	4.83%	4.85%
	TSP 0.1	4.82%	4.83%	4.85%
	TSP 0.05	4.95%	4.83%	4.85%
	TSP 0.01	5.49%	4.84%	4.85%
	TSP 0.005	5.69%	4.87%	4.85%
Laplace	ANOVA	4.77%	4.93%	4.90%
	Kruskal-Wallis	4.78%	4.93%	4.92%
	TSP 0.1	4.92%	4.97%	4.92%
	TSP 0.05	4.97%	5.00%	4.92%
	TSP 0.01	5.00%	5.04%	4.97%
	TSP 0.005	4.96%	5.06%	4.98%

The overall picture seems to be that the two pure tests both perform in a similar way as the two stage process. The only, but rather minor, exception seems to be that that the two stage process generates slightly more type I errors when samples of a small size are drawn from an exponential distribution. This tendency is amplified when the preliminary test is conducted at a smaller significance level, but diminishes when the sample size becomes larger. The reason is probably that the normality screening of samples taken from exponential distributions favors samples with smaller standard deviations (Rochon & Kieser, 2011).

Estimated Power under Exponential Distribution

This section presents the results when all samples were drawn from exponential distributions with different mean values. Table 3 displays the frequencies of significant tests out of the 100,000 conducted tests for the different combinations of test procedure (the ANOVA without the preliminary test, the Kruskal-Wallis test without the preliminary test, or the two-stage procedure), sample size ($n = 15$, $n = 30$, or $n = 60$ in each group), and effect size ($f = 0.10$, $f = 0.25$, $f = 0.40$, or $f = 0.65$). For example, the ANOVA without the preliminary test yielded 12,740 significant tests when the effect size was $f = 0.10$ and when $n = 30$. Hence, the proportion of significant tests for this specific combination test procedure, sample

Table 3. Estimated power under the exponential distribution

Method	Sample size *n*	*f* = 0.10	*f* = 0.25	*f* = 0.40	*f* = 0.65
ANOVA	15	8.45%	32.05%	66.47%	95.54%
	30	12.74%	56.37%	91.57%	99.65%
	60	21.05%	85.11%	99.20%	99.42%
Kruskal-Wallis	15	13.34%	52.24%	85.71%	99.18%
	30	24.15%	84.81%	99.05%	99.69%
	60	45.35%	98.63%	99.42%	99.42%
TSP 0.1	15	13.40%	52.27%	85.70%	99.18%
	30	24.15%	84.81%	99.05%	99.69%
	60	45.35%	98.63%	99.42%	99.42%
TSP 0.05	15	13.53%	52.29%	85.64%	99.17%
	30	24.15%	84.81%	99.05%	99.69%
	60	45.35%	98.63%	99.42%	99.42%
TSP 0.01	15	13.82%	51.53%	84.66%	98.97%
	30	24.16%	84.80%	99.05%	99.69%
	60	45.35%	98.63%	99.42%	99.42%
TSP 0.005	15	13.71%	50.44%	83.43%	98.76%
	30	24.18%	84.78%	99.04%	99.69%
	60	45.35%	98.63%	99.42%	99.42%

size, and effect size was 12.74% when samples were taken from exponential distributions.

The two-stage procedure (regardless of the significance level of the Shapiro-Wilk test) and the Kruskal-Wallis test perform similarly at all combinations of effect size and sample size. However, the ANOVA has substantially less power than both other procedures. Furthermore, this pattern remains the same regardless of the sample size. The main reason is of course that the preliminary normality screening in the two-stage procedure in most cases favors the Kruskal-Wallis test at the second stage.

Estimated Power under Laplace Distribution

This section presents the results when all samples were drawn from Laplace distributions with different mean values. Table 4 displays the frequencies of significant tests out of the 100,000 conducted tests for the different combinations of test procedure, sample size, and effect size.

As when samples were drawn from exponential distributions, the ANOVA has a lower power than both the two-stage procedure (regardless of the significance level of the Shapiro-Wilk test) and the Kruskal-Wallis test when the samples come from Laplace distributions. The effect is somewhat smaller, however.

Table 4. Estimated power under the Laplace distribution

Method	Sample size *n*	*f* = 0.10	*f* = 0.25	*f* = 0.40	*f* = 0.65
ANOVA	15	8.32%	30.06%	64.72%	96.08%
	30	12.34%	54.84%	91.80%	99.41%
	60	20.70%	84.70%	98.73%	98.90%
Kruskal-Wallis	15	9.56%	37.45%	74.34%	98.09%
	30	15.70%	68.64%	96.86%	99.44%
	60	28.16%	93.86%	98.89%	98.90%
TSP 0.1	15	9.54%	36.47%	72.93%	97.90%
	30	15.50%	67.61%	96.51%	99.44%
	60	28.08%	93.72%	98.89%	98.90%
TSP 0.05	15	9.49%	35.60%	71.92%	97.72%
	30	15.29%	66.64%	96.18%	99.44%
	60	27.94%	93.55%	98.89%	98.90%
TSP 0.01	15	9.19%	33.53%	69.24%	97.20%
	30	14.65%	63.70%	95.20%	99.43%
	60	27.15%	92.65%	98.88%	98.90%
TSP 0.005	15	9.01%	32.74%	68.23%	96.97%
	30	14.27%	62.36%	94.72%	99.43%
	60	26.70%	92.09%	98.86%	98.90%

The Kruskal-Wallis test has slightly higher power than the two-stage procedure, and this effect is amplified when the preliminary test is conducted at a smaller significance level irrespective of the sample size. The reason is probably that normality screening at a lower significance level favors the ANOVA at the second stage because the Laplace distribution, despite its leptokurtic shape, resembles the normal distribution more than the exponential distribution does due to its symmetry and unimodality.

Estimated Power under Uniform Distribution

This section presents the results when all samples were drawn from uniform distributions with different mean values. Table 5 displays the frequencies of significant tests out of the 100,000 conducted tests for the different combinations of test procedure, sample size, and effect size.

In line with previous research (see Schmider, Ziegler, Danay, Beyer, & Buhner, 2010, for a review), the ANOVA shows slightly higher power than the Kruskal-Wallis test when samples are drawn from uniform distributions. The main reason is of course that the uniform distribution, in terms of skewness and/or kurtosis, does not impose an equally serious violation of normality as the Laplace and exponential distributions do.

Table 5. Estimated power under the uniform distribution

Method	Sample size *n*	*f* = 0.10	*f* = 0.25	*f* = 0.40	*f* = 0.65
ANOVA	15	8.25%	27.83%	62.92%	97.66%
	30	12.13%	53.63%	92.81%	100.00%
	60	20.35%	85.70%	99.91%	100.00%
Kruskal-Wallis	15	7.75%	25.05%	56.24%	94.61%
	30	11.53%	49.05%	88.43%	99.96%
	60	19.43%	81.37%	99.64%	100.00%
TSP 0.1	15	8.08%	26.69%	59.52%	95.37%
	30	11.62%	49.56%	88.70%	99.96%
	60	19.43%	81.37%	99.64%	100.00%
TSP 0.05	15	8.21%	27.41%	61.28%	96.16%
	30	11.77%	50.59%	89.37%	99.96%
	60	19.43%	81.38%	99.64%	100.00%
TSP 0.01	15	8.28%	27.84%	62.83%	97.40%
	30	12.13%	53.05%	91.67%	99.97%
	60	19.64%	81.88%	99.65%	100.00%
TSP 0.005	15	8.28%	27.85%	62.91%	97.57%
	30	12.16%	53.49%	92.30%	99.98%
	60	19.89%	82.69%	99.67%	100.00%

In general, the ANOVA performs somewhat better than the two-stage procedure while the Kruskal-Wallis test performs somewhat worse. As one might expect, the performance of the two-stage procedure approaches the performance of the ANOVA when the normality tests are conducted at a lower significance level as that favors the ANOVA at the second stage. However, the difference in performance between the Kruskal-Wallis test and the two-stage procedure also diminishes when the sample size is larger.

Estimated Power under Normal Distribution

This section presents the results when all samples were drawn from normal distributions with different mean values. Table 6 displays the frequencies of significant tests out of the 100,000 conducted tests for the different combinations of test procedure, sample size, and effect size.

As one would expect, the Kruskal-Wallis test performs somewhat worse than the ANOVA. The two-stage procedure on the other hand has a performance very similar to the ANOVA, which is easy to understand as the Shapiro-Wilk test only favors the Kruskal-Wallis test at the second stage in a few cases.

Table 6. Estimated power under the normal distribution

Method	Sample size *n*	*f* = 0.10	*f* = 0.25	*f* = 0.40	*f* = 0.65
ANOVA	15	8.29%	28.55%	63.53%	97.05%
	30	12.23%	54.16%	92.60%	99.99%
	60	20.50%	85.40%	99.87%	100.00%
Kruskal-Wallis	15	7.80%	26.63%	60.49%	96.13%
	30	11.65%	51.75%	91.21%	99.98%
	60	19.71%	83.61%	99.82%	100.00%
TSP 0.1	15	8.43%	28.48%	63.15%	96.90%
	30	12.38%	53.85%	92.30%	99.99%
	60	20.65%	85.00%	99.86%	100.00%
TSP 0.05	15	8.45%	28.69%	63.47%	97.02%
	30	12.39%	54.15%	92.51%	99.99%
	60	20.68%	85.26%	99.87%	100.00%
TSP 0.01	15	8.37%	28.67%	63.63%	97.10%
	30	12.31%	54.28%	92.65%	99.99%
	60	20.60%	85.45%	99.88%	100.00%
TSP 0.005	15	8.36%	28.64%	63.63%	97.10%
	30	12.29%	54.25%	92.66%	99.99%
	60	20.58%	85.46%	99.88%	100.00%

Conclusion

A preliminary test of normality before conducting a location test will yield one of four possible outcomes:

- Incorrectly rejecting H_0 (i.e. a type I error), resulting in the use of a location test with less power than necessary at the second stage.
- Correctly rejecting H_0, resulting in the (correct) use of a non-parametric location test at the second stage.
- Incorrectly 'accepting' H_0 (i.e. a type II error), resulting in the use of an invalid location test (i.e. with uncertain actual power and significance) at the second stage.
- Correctly 'accepting' H_0, resulting in the (correct) use of a parametric location test at the second stage.

Therefore, the probability of a type I error as well as of a type II error of the entire two-stage procedure may be heavily distorted, if it is at all possible to determine. In this study, we have used simulations in order to shed some light on this problem. While we have been unable to see any specific disturbance in the type I error probability of the two-stage procedure, the effect on power exhibits some interesting patterns in comparison to the 'pure' methods. The overall impression is that the two-stage procedure performs similarly to the ANOVA, but slightly better than the Kruskal-Wallis test when the parent distributions are 'relatively normally' distributed. On the other hand, the two-stage procedure performs similarly to the Kruskal-Wallis test, but substantially better than the ANOVA, when the parent distributions are characterized by a more distinct violation of normality. These observed patterns are also relatively insensitive to the sample sizes.

The choice of level of significance for the preliminary tests also requires some thought. If we, for example, want to compare six groups and choose to use $\alpha = 0.10$ during the normality screening, the overall probability of a type I error, leading us to use a less powerful non-parametric test to compare the means in the second stage, would be around 50%. On the other hand, since the ANOVA typically perform a lot worse than the Kruskal-Wallis when there is a more distinct violation of normality while the Kruskal-Wallis only perform slightly worse when normality actually holds, type II errors are potentially a lot more harmful than type I errors in the first stage of the two-stage procedure.

Hence, in contrast to previous similar research on bivariate situations, the two-stage procedure seems in general to be the preferable choice when conducting location tests for three samples as neither the ANOVA nor the Kruskal-Wallis test as one-stage procedures perform noticeably better than the two-stage procedure, while the two-stage procedure is substantially better than the ANOVA when data are distinctly non-normally distributed. This is especially so when the normality screening is conducted at a relatively high significance level. Hence, the two-stage procedure seems to have no practical shortcoming but an apparent practical advantage. The theoretical weakness, of course, is that the true probability of type I and type II errors may be unknown, which, in addition to the fact that the ANOVA is known to be relatively robust to non-normally distributed data when groups sizes are roughly equal, albeit more sensitive to non-normality when group sizes are unequal, should be borne in mind (Schmider et al., 2010; Wilcox, 2012; Field, 2013).

Future research should extend the design in this study, for example, by using different sample sizes in the groups, and/or by including other statistical distributions in order to evaluate other types of non-normality than those related to skewness and kurtosis. Further research in this field should also aim at comparing other types of parametric methods with their non-parametric counterparts as two-stage procedures, as well as comparing two-stage procedures with robust procedures in general such as bootstrapping. Screening for other types of violations, for example, heteroscedasticity, in the first stage would also be interesting to consider.

References

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology, 9*(2), 78-84. doi: 10.1027/1614-2241/a000057

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. doi: 10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997-1003. doi: 10.1037/0003-066X.49.12.997

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics*. London: SAGE Publications.

Keselman, H. J., Othman, A. R., & Wilcox, R. R. (2013). Preliminary testing for normality: Is this a good practice? *Journal of Modern Applied*

Statistical Methods, 12(2), 2-19. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol12/iss2/2/

Keselman, H. J., Othman, A. R., & Wilcox, R. R. (2014). Testing for normality in the multi-group problem: Is this a good practice? *Clinical Dermatology, 2*(1), 29-43. doi: 10.11138/cderm/2014.2.1.029

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

Lantz, B. (2013). The impact of sample non-normality on ANOVA and alternative methods. *British Journal of Mathematical and Statistical Psychology, 66*(2), 224-244. doi: 10.1111/j.2044-8317.2012.02047.x

Mandelbrot, B., & Taleb, N. (2006, March 24). A focus on the exceptions that prove the rule. *Financial Times*.

Marmolejo-Ramos, F., & González-Burgos, J. (2013). A power comparison of various tests of univariate normality on ex-Gaussian distributions. *Methodology, 9*(4), 137-149. doi: 10.1027/1614-2241/a000059

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

Othman, A. R., Keselman, H. J., & Wilcox, R. R. (2015). Assessing normality: Applications in multi-group designs. *Malaysian Journal of Mathematical Sciences, 9*(1), 53-65. Retrieved from http://einspem.upm.edu.my/journal/volume9.1.php

Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample *t* test: Pre-testing its assumptions does not pay off. *Statistical Papers, 52*(1), 219-231. doi: 10.1007/s00362-009-0224-x

Razali, N., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics, 2*(1), 21-33.

Rochon, J., Gondan, M., & Kieser, M. (2012). To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology, 12*(81). doi: 10.1186/1471-2288-12-81

Rochon, J., & Kieser, M. (2011). A closer look at the effect of preliminary goodness-of-fit testing for normality for the one-sample *t*-test. *British Journal of Mathematical and Statistical Psychology, 64*(3), 410-426. doi: 10.1348/2044-8317.002003

Royston, P. (1992). Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and Computing, 2*(3), 117-119. doi: 10.1007/BF01891203

Ruscio, J., & Roche, B (2012). Variance heterogeneity in published psychological research. *Methodology, 8*(1), 1-11. doi: 10.1027/1614-2241/a000034

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Buhner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology, 6*(4), 147-151. doi: 10.1027/1614-2241/a000016

Schoder, V., Himmelmann, A., & Wilhelm, K. P. (2006). Preliminary testing for normality: Some statistical aspects of a common concept. *Clinical and Experimental Dermatology, 31*(6), 757-761. doi: 10.1111/j.1365-2230.2006.02206.x

Schucany, W. R., & Ng, T. (2006). Preliminary goodness-of-fit tests for normality do not validate the one-sample Student *t*. *Communications in Statistics – Theory and Methods, 35*(12), 2275-2286. doi: 10.1080/03610920600853308

Shuster, J. J. (2005). Diagnostics for assumptions in moderate to large simple clinical trials: Do they really help? *Statistics in Medicine, 24*(16), 2431-2438. doi: 10.1002/sim.2175

Shuster, J. J. (2009). Student *t*-tests for potentially abnormal data. *Statistics in Medicine, 28*(16), 2170-2184. doi: 10.1002/sim.3581

Wells, C. S., & Hinze, J. M. (2006). Dealing with assumptions underlying statistical tests. *Psychology in the Schools, 44*(5), 495-502. doi: 10.1002/pits.20241

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. San Diego, CA: Academic Press.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology, 57*(1), 173-181. doi: 10.1348/000711004849222

Zimmerman, D. W. (2011). A simple and effective decision rule for choosing a significance test to protect against non-normality. *British Journal of Mathematical and Statistical Psychology, 64*(3), 388-409. doi: 10.1348/000711010X524739

Zimmerman, D. W. (2014). Consequences of choosing samples in hypothesis testing to ensure homogeneity of variance. *British Journal of Mathematical and Statistical Psychology, 67*(1), 1-29. doi: 10.1111/bmsp.12001