British Journal of Mathematical and Statistical Psychology (2013), 66, 224–244

The British Psychological Society

www.wileyonlinelibrary.com

The impact of sample non-normality on ANOVA and alternative methods

Björn Lantz*

University of Borås, Sweden

In this journal, Zimmerman (2004, 2011) has discussed preliminary tests that researchers often use to choose an appropriate method for comparing locations when the assumption of normality is doubtful. The conceptual problem with this approach is that such a two-stage process makes both the power and the significance of the entire procedure uncertain, as type I and type II errors are possible at both stages. A type I error at the first stage, for example, will obviously increase the probability of a type II error at the second stage. Based on the idea of Schmider *et al.* (2010), which proposes that simulated sets of sample data be ranked with respect to their degree of normality, this paper investigates the relationship between population non-normality and sample non-normality with respect to the performance of the ANOVA, Brown–Forsythe test, Welch test, and Kruskal–Wallis test when used with different distributions, sample sizes, and effect sizes. The overall conclusion is that the Kruskal–Wallis test is considerably less sensitive to the degree of sample normality when populations are distinctly non-normal and should therefore be the primary tool used to compare locations when it is known that populations are not at least approximately normal.

1. Introduction

This paper focuses on the difference between sample non-normality and population non-normality with respect to statistical significance testing. Sample non-normality prevails when the sample data have a shape suggesting that their parent population might be non-normal. Population non-normality, on the other hand, prevails when a parent population actually is non-normal. Owing to the variations in samples, some sets of data will be characterized by sample non-normality which indicates population non-normality even though the parent populations are normal. On the other hand, for the same reason, other samples will seem relatively normal even though the parent populations are characterized by a distinct non-normality. The problem is that in both cases a researcher with no previous understanding of the distribution of the parent populations may use suboptimal methods to compare locations from these populations.

*Correspondence should be addressed to Björn Lantz, University of Borås, 50190 Borås, Sweden (e-mail: bjorn.lantz@hb.se).

DOI:10.1111/j.2044-8317.2012.02047.x

Most statistics textbooks recommend the one-way fixed effect analysis of variance (ANOVA) as the best method for comparing the means of several populations if they are approximately normally distributed and have similar variances. If these assumptions are not met, textbooks usually then recommend a robust alternative method such as the Brown–Forsythe test (Brown & Forsythe, 1974), the Welch test (Welch, 1951), or the non-parametric Kruskal–Wallis test (Kruskal & Wallis, 1952). The conceptual problem in this approach is that a two-stage process, in which a preliminary decision on whether an assumption should be regarded as valid or not is made first and the main comparison of locations is then conducted based on the preliminary findings, changes the probability of type I and type II errors (see Zimmerman, 2004, 2011). Hence, the actual levels of significance and power obtained by the overall procedure remain unclear.

Population non-normality appears to be quite common in psychological data. For example, in his empirical study of 440 achievement and psychometric measures, Micceri (1989) found significant non-normality contaminations of different types in all of them, including tail weights from the uniform to the double exponential, exponential level asymmetry, and bimodality. Four general types of measures were examined: general achievement/ability measures, criterion/mastery measures, psychometric measures, and gain scores (the difference between a before and after measure). Micceri (1989) classified all measures in terms of the degree of tail weight contamination and the degree of asymmetry. Interestingly, the four types of measures were rather different in terms of the degree of contamination they typically displayed. In Table 1, the location of the median observation for each measure is indicated. Gain scores were usually relatively symmetrically distributed even though they often had excess tail weight. Ability measures were often characterized by moderate asymmetry, but with less pronounced excess tail weight than was the case in gain scores. Psychometric measures tended to combine a more significant excess tail weight and a stronger asymmetry than ability measures, and criterion/mastery measures were even more extreme in both regards.

A great deal of prior research has examined and compared the degree of sensitivity of the ANOVA and its alternatives for different types of population non-normality. The results show that alternative methods perform worse than the ANOVA when populations are actually normally distributed with equal variances, but often perform better when this is not the case (Glass, Peckham, & Sanders, 1972; Tomarken & Serlin, 1986; Harwell,

Table 1. Median observations for the four measure types in Micceri (1989)

		Degree of asymmetry			
		Near symmetry	Moderate	High	Extreme
Tail weight	Light				
	Normal				
	Somewhat heavy		Ability measures		
	Very heavy	Gain scores		Psychometric measures	
	Extreme				Criterion/mastery measures

Rubinstein, Hayes, & Olds, 1992; Khan & Rayner, 2003). In this paper, we will evaluate the sensitivity of the ANOVA and its alternatives to different degrees of sample normality under different types of population non-normality. Recently, Schmider, Ziegler, Danay, Beyer, and Bühner (2010) presented evidence regarding the robustness of the one-way ANOVA against violations of the assumption of normally distributed response variable populations. They used the Monte Carlo simulation to create random samples from normal, uniform, and exponential distributions under different effect sizes and filtered the samples via a goodness-of-fit procedure, based on the Kolmogorov–Smirnov test. Even though they only tested one type of distinct population non-normality, one sample size, and no alternative methods, the authors interpret their results as offering 'strong support for the robustness of the ANOVA under application of non-normally distributed data' (Schmider *et al.*, 2010, p. 150).

The basic idea of using the Monte Carlo simulation to test statistical models in general is by no means new (e.g. Box & Muller, 1958), and different aspects of the ANOVA have been analysed using Monte Carlo methods for a long time (e.g. Glass *et al.*, 1972; Feir & Toothaker, 1974). However, the fact that the Monte Carlo based research on the performance of the ANOVA under different conditions is still in progress (e.g. Watthanacheewakul, 2011; Schmider *et al.*, 2010; Van Hecke, 2010) indicates that our knowledge about the performance of the ANOVA is still incomplete, either on its own, or in relation to its robust alternatives.

From a theoretical point of view, it is easy to see why the ANOVA fails when the assumptions of underlying normal distributions with equal variances are violated. When data are skewed, the central location may not be reflected by the mean. In addition, when populations have different variances, the noise from one population can overshadow the true signal from another. This is why alternative methods such as the Kruskal–Wallis test are recommended in textbooks when the assumptions of the ANOVA are violated. However, even though the Kruskal–Wallis test eliminates the issues of validity, such as when underlying distributions are skewed, other problems may emerge when it is used, since the precision of a non-parametric test may be reduced owing to the transformation into ranks (e.g. Edgington & Onghena, 2007), its power may be lower (e.g. Feir & Toothaker, 1974; Tanizaki, 1997), and concurrent violation of several assumptions may bias a non-parametric test more than its parametric counterpart (Zimmerman, 1998). Weighing the evidence on both sides, contemporary researchers have resigned themselves to the fact that, in many cases, they must simply put up with the possible problems regarding the validity of the ANOVA when using sample data to compare locations of three or more populations. This conclusion, however, ignores the previously discussed problem regarding the relationship between population non-normality and sample non-normality. It is this contradiction that forms the basis for the present study.

The idea of Schmider *et al.* (2010) to rank Monte Carlo simulated samples from different underlying populations based on the Kolmogorov–Smirnov goodness of fit to normality provides us with three different aspects of non-normality when sampling from one distinct non-normal distribution and one normal distribution. In Table 2, category B corresponds to random samples from an underlying normal distribution, but with a low degree of sample normality; category C to random samples from a distinct non-normal distribution, but with a high degree of sample normality; and category D to random samples from a distinct non-normal distribution, but with a low degree of sample normality; finally, category A corresponds to random samples from a normal distribution with a high degree of sample normality.

Table 2. Three different aspects of non-normality

		Sample normality	
		High	Low
Underlying distribution	Normal	A	B
	Non-normal	C	D

The aim of this study is to explore the differences between the performance of the ANOVA and its common alternatives, the Brown–Forsythe test, the Welch test, and the Kruskal–Wallis test, with respect to these different aspects of non-normality. In Section 2 the methodology of the study is described. Then, in Sections 3 and 4, the results of the simulations are presented and discussed. The paper concludes with the implications of these results for use in statistical analysis in practice.

2. Methodology

The one-way ANOVA is based on the idea that the true means in groups are more likely to be equal if the variation between the groups is small, as compared to the variation within the groups. Without going into the technicalities involved (see Tomarken & Serlin, 1986, for summaries of the mathematical definitions of these methods), the Brown–Forsythe and Welch tests are considered robust compared to the ANOVA because their definitions of variation within groups are based on the relationship between the different sample sizes in the different groups instead of a simple pooled variance estimate, which means that they become less sensitive to heteroscedasticity. The Kruskal–Wallis test is considered robust because it is based on ranks instead of actual values, which means that the underlying distribution does not matter as long as the observed values can be ranked.

The simulations in this study were based on random numbers from the following five different probability distributions based on the findings of Micceri (1989) regarding non-normality in psychological data: normal, uniform, exponential, U-quadratic, and Laplace (see Figure 1). The uniform distribution represents an extreme form of light tail weight contamination without asymmetry. It can also be seen as a fairly good approximation of the normal distribution owing to its mound shape and lack of outliers. The exponential distribution represents an extreme form of heavy tail weight contamination combined with extreme asymmetry. The U-quadratic distribution exhibits another common type of non-normal contamination found in the study of Micceri (1989), namely bimodality. Owing to its symmetry and platykurtic shape, the U-quadratic distribution can be seen as an extreme bimodal version of the uniform distribution. Hence, it represents an extreme form of light tail weight contamination combined with bimodality. The Laplace distribution represents an extreme form of heavy tail weight contamination without asymmetry. At first glance, the Laplace distribution might appear to resemble the normal

Figure 1. Normal, uniform, exponential, U-quadratic and Laplace distributions.

Table 3. The 20 different combinations of populations means and sample sizes tested for each combination of procedure and distribution

		$n_1/n_2/n_3$			
Effect size, f	$\mu_1/\mu_2/\mu_3$	25/25/25	5/5/5	5/5/25	5/25/25
0.00	0.000/0.000/0.000	1	6	11	16
0.10	0.000/0.123/0.246	2	7	12	17
0.25	0.000/0.307/0.614	3	8	13	18
0.40	0.000/0.490/0.980	4	9	14	19
0.65	0.000/0.797/1.594	5	10	15	20

distribution. The major difference, however, is that in the Laplace distribution, outliers are much more common owing to heavier tails. Hence, it represents an important form of non-normality in cases where an extreme degree of randomness exists (e.g. Mandelbrot & Taleb, 2006).

Our experimental design is conceptually similar to that of Schmider *et al.* (2010), even though it is considerably extended. A design with three populations ($k = 3$) and four different combinations of small (defined as $n = 5$) and large (defined as $n = 25$) sample sizes was used. For the purpose of comparison, parameter values were chosen to obtain random numbers with a mean of 0 and a variance of 1 from all distributions. Another reason for keeping the variance constant throughout the study was to isolate the effect of the different aspects of normality. Table 3 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 ($f = 0.65$). All mean values were calculated using G ∗ Power version 3.1.2 (Faul, Erdfleder, Lang, & Buchner, 2007). As shown in Table 3, each combination of test procedure and parent population distribution was evaluated for 20 different combinations of sample sizes and effect size. Thus, a total of 400 different combinations of test procedure, parent population distribution, sample size, and effect size were evaluated in this study. First, for each combination, 3 × 50,000 sets of random numbers were generated using the inverse transformation method (Devroye, 1986), and the null hypothesis that corresponds to no difference between the locations of the populations was challenged at an alpha level of .05, irrespective of the degree of sample normality. Second, for each combination, 3 × 100,000 sets of random numbers were generated using the inverse transformation method (Devroye, 1986), and the Kolmogorov–Smirnov goodness-of-fit statistic with respect to normal distribution was calculated based on the empirical mean and variance of the data in each set. The sets were then ranked based on their Kolmogorov–Smirnov statistics, and the top 10% and bottom 10% were identified and analysed in the same manner described above. All simulation procedures were conducted using Microsoft Excel 2010.

3. Results without regard to the degree of sample normality

In this section unfiltered simulation data are used to evaluate the performance of the four test procedures, each at five different effect sizes and with four different combinations of sample sizes. A considerable number of the results presented in this section have been previously reported in the literature. Some have not, however. For example, we have not

Table 4. Performance under normal distribution disregarding sample normality

Effect size, f	ANOVA	Kruskal–Wallis	Brown–Forsythe	Welch	p-value (chi-square)
			Panel A: 25/25/25		
0.0	0.0495	0.0474	0.0491	0.0498	0.3073
0.1	0.1077	0.1029	0.1069	0.1066	0.0906
0.25	0.4582	0.4335	0.4570	0.4490	0.0000
0.4	0.8651	0.8454	0.8643	0.8583	0.0023
0.65	0.9949	0.9945	0.9949	0.9948	0.9999
			Panel B: 5/5/5		
0.0	0.0490	0.0449	0.0394	0.0407	0.0000
0.1	0.0595	0.0528	0.0471	0.0468	0.0000
0.25	0.1100	0.0962	0.0909	0.0869	0.0000
0.4	0.2142	0.1868	0.1822	0.1706	0.0000
0.65	0.4977	0.4442	0.4479	0.4115	0.0000
			Panel C: 5/5/25		
0.0	0.0510	0.0449	0.0468	0.0550	0.0000
0.1	0.0707	0.0603	0.0614	0.0675	0.0000
0.25	0.1869	0.1648	0.1514	0.1495	0.0000
0.4	0.4199	0.3745	0.3286	0.3129	0.0000
0.65	0.8426	0.7942	0.7223	0.6907	0.0000
			Panel D: 5/25/25		
0.0	0.0515	0.0467	0.0528	0.0575	0.0000
0.1	0.0752	0.0694	0.0752	0.0775	0.0000
0.25	0.2263	0.2081	0.2104	0.1949	0.0000
0.4	0.5158	0.4796	0.4644	0.4317	0.0000
0.65	0.9183	0.8956	0.8633	0.8567	0.0000

been able to find any previous research in which the U-quadratic distribution was used in a simulation study conducted to compare the ANOVA with alternative procedures. Furthermore, in an experiment, it is generally essential to have a control group to assess the results from the test group correctly.

For a better understanding of the reliability of the statistics presented in this section, it should be noted that the standard error of a sample proportion at a sample size of 50,000 is about 0.002 when the proportion is 0.5, and decreases to about 0.001 when the proportion is 0.05 or 0.95.

3.1. Sampling from a normal distribution

In Table 4, all simulation results in which the parent populations are characterized by a normal distribution are shown.

As was expected, based upon both theory and previous research, the ANOVA performs slightly but significantly better than the Kruskal–Wallis test when the parent population is normal for all combinations of effect size and sample size, except when all sample sizes are large and the effect size is very large. There are no notable differences

Table 5. Performance under uniform distribution disregarding sample normality

Effect size, f	ANOVA	Kruskal–Wallis	Brown–Forsythe	Welch	p-value (chi-square)
Panel A: 25/25/25					
0.0	0.0506	0.0476	0.0504	0.0512	0.0482
0.1	0.1068	0.1019	0.1066	0.1049	0.0594
0.25	0.4556	0.4141	0.4550	0.4404	0.0000
0.4	0.8688	0.8125	0.8684	0.8594	0.0000
0.65	0.9995	0.9973	0.9995	0.9995	0.9800
Panel B: 5/5/5					
0.0	0.0542	0.0450	0.0439	0.0547	0.0000
0.1	0.0631	0.0520	0.0508	0.0600	0.0000
0.25	0.1072	0.0875	0.0901	0.0909	0.0000
0.4	0.1966	0.1612	0.1743	0.1558	0.0000
0.65	0.4700	0.3881	0.4355	0.3589	0.0000
Panel C: 5/5/25					
0.0	0.0495	0.0451	0.0531	0.0802	0.0000
0.1	0.0682	0.0606	0.0649	0.0886	0.0000
0.25	0.1836	0.1541	0.1424	0.1466	0.0000
0.4	0.4118	0.3361	0.3050	0.2704	0.0000
0.65	0.8450	0.7442	0.7274	0.6364	0.0000
Panel D: 5/25/25					
0.0	0.0506	0.0469	0.0623	0.0728	0.0000
0.1	0.0742	0.0688	0.0816	0.0865	0.0000
0.25	0.2239	0.1979	0.2007	0.1808	0.0000
0.4	0.5110	0.4451	0.4421	0.3901	0.0000
0.65	0.9239	0.8653	0.8787	0.8459	0.0000

between the ANOVA and the Brown–Forsythe or Welch tests for large sample sizes, but when at least one sample size is small, the ANOVA tends to perform better in almost all cases. Since this is the only case in which no ANOVA assumption is violated, there is of course no reason to choose an alternative method to analyse these types of data.

3.2. Sampling from a uniform distribution

In Table 5, all simulation results in which the parent populations are characterized by a uniform distribution are shown. The uniform distribution can be used as an example of a distribution that does not severely violate the assumption of a normal distribution owing to its platykurtic mound shape.

When all sample sizes are large, the ANOVA outperforms the Kruskal–Wallis tests even more markedly than it does under the normal distribution. There are no notable differences between the ANOVA and the Brown–Forsythe or the Welch tests with large sample sizes, but when all sample sizes are small, the ANOVA performs better for all effect sizes. However, when the sample sizes are unequal, the Welch test actually outperforms the ANOVA when the effect size is small, although the ANOVA again performs best

Table 6. Performance under exponential distribution disregarding sample normality

Effect size, f	ANOVA	Kruskal–Wallis	Brown–Forsythe	Welch	p-value (chi-square)
			Panel A: 25/25/25		
0.0	0.0471	0.0475	0.0455	0.0532	0.0000
0.1	0.1150	0.2057	0.1126	0.1229	0.0000
0.25	0.4893	0.7706	0.4862	0.4948	0.0000
0.4	0.8650	0.9799	0.8635	0.8711	0.0000
0.65	0.9952	0.9977	0.9950	0.9963	0.9711
			Panel B: 5/5/5		
0.0	0.0399	0.0450	0.0224	0.0279	0.0000
0.1	0.0532	0.0666	0.0310	0.0411	0.0000
0.25	0.1302	0.1680	0.0886	0.1148	0.0000
0.4	0.2763	0.3241	0.2137	0.2521	0.0000
0.65	0.5797	0.6020	0.5065	0.5401	0.0000
			Panel C: 5/5/25		
0.0	0.0512	0.0450	0.0562	0.0959	0.0000
0.1	0.0569	0.1268	0.1126	0.1776	0.0000
0.25	0.2041	0.3707	0.2567	0.3545	0.0000
0.4	0.4880	0.6233	0.4362	0.5418	0.0000
0.65	0.8569	0.8766	0.6995	0.7803	0.0000
			Panel D: 5/25/25		
0.0	0.0484	0.0468	0.0504	0.0972	0.0000
0.1	0.0665	0.1325	0.1046	0.1663	0.0000
0.25	0.2410	0.4612	0.3026	0.3397	0.0000
0.4	0.5619	0.7877	0.5496	0.5694	0.0000
0.65	0.9179	0.9790	0.8274	0.8804	0.0000

for larger effect sizes. On the other hand, it should be noted that the Welch displays a heightened probability of type I errors when the sample sizes are unequal. The Brown–Forsythe test generally performs better than the Welch when adjustments are made to account for this heightened probability.

3.3. Sampling from an exponential distribution

In Table 6, all simulation results in which the parent populations are characterized by an exponential distribution are shown. The exponential distribution can be used as an example of a distribution that severely violates an assumption of a normal distribution owing to its leptokurtic and skewed shape.

Similar to the case of uniform distribution, the Welch test is characterized by better performance than the ANOVA when the sample sizes are unequal and the effect size is small or medium; however, this is achieved with a higher probability of type I errors. The Brown–Forsythe test, on the other hand, is not characterized by a heightened probability of type I errors, even though it also outperforms the ANOVA when the sample sizes are unequal and the effect size is small or medium. When all the sample sizes are large,

Table 7. Performance under U-quadratic distribution disregarding sample normality

Effect size, *f*	ANOVA	Kruskal–Wallis	Brown–Forsythe	Welch	*p*-value (chi-square)
			Panel A: 25/25/25		
0.0	0.0508	0.0476	0.0507	0.0519	0.0155
0.1	0.1055	0.3054	0.1053	0.1036	0.0000
0.25	0.4522	0.8280	0.4518	0.4340	0.0000
0.4	0.8709	0.9449	0.8707	0.8598	0.0000
0.65	0.9996	0.9917	0.9996	0.9997	0.5023
			Panel B: 5/5/5		
0.0	0.0577	0.0450	0.0423	0.0599	0.0000
0.1	0.0641	0.0742	0.0480	0.0634	0.0000
0.25	0.1058	0.1731	0.0868	0.0914	0.0000
0.4	0.1908	0.2550	0.1705	0.1420	0.0000
0.65	0.4534	0.3598	0.4322	0.3206	0.0000
			Panel C: 5/5/25		
0.0	0.0501	0.0451	0.0506	0.1197	0.0000
0.1	0.0677	0.1225	0.062	0.1185	0.0000
0.25	0.1811	0.3430	0.1334	0.1404	0.0000
0.4	0.4058	0.4741	0.2941	0.2402	0.0000
0.65	0.8442	0.6425	0.7330	0.5976	0.0000
			Panel D: 5/25/25		
0.0	0.0504	0.0469	0.0697	0.0800	0.0000
0.1	0.0733	0.1593	0.0817	0.0908	0.0000
0.25	0.2209	0.5354	0.1882	0.1713	0.0000
0.4	0.5085	0.7597	0.4305	0.3676	0.0000
0.65	0.9241	0.8965	0.8898	0.8404	0.0000

there is no notable difference in the performance between the ANOVA and the Brown–Forsythe or Welch tests; however, at small and equal sample sizes, the ANOVA performs the best among the three parametric methods.

In our study results, though, the Kruskal–Wallis test outperforms all competitors for all combinations of effect size and sample sizes when the parent populations are exponentially distributed. It should be noted that this superiority is not accompanied by an increased probability of type I errors.

3.4. Sampling from an U-quadratic distribution

In Table 7, all simulation results in which the parent populations are characterized by a U-quadratic distribution are shown. The U-quadratic distribution can be used as an example of a distribution that severely violates the assumption of a normal distribution owing to its inverted mound shape.

As in the previous non-normal cases, the Welch test has an increased probability of type I errors when the sample sizes are unequal, even though it also performs better with small effect sizes than the ANOVA. With larger effect sizes, the ANOVA again outperforms

Table 8. Performance under Laplace distribution disregarding sample normality

Effect size, f	ANOVA	Kruskal–Wallis	Brown–Forsythe	Welch	p-value (chi-square)
			Panel A: 25/25/25		
0.0	0.0479	0.0474	0.0473	0.0464	0.7472
0.1	0.1091	0.1361	0.1076	0.1098	0.0000
0.25	0.4718	0.5980	0.4692	0.4746	0.0000
0.4	0.8618	0.9344	0.8602	0.8609	0.0000
0.65	0.9940	0.9955	0.9939	0.9935	0.9898
			Panel B: 5/5/5		
0.0	0.0416	0.0449	0.0294	0.0264	0.0000
0.1	0.0537	0.0579	0.0390	0.0351	0.0000
0.25	0.1194	0.1287	0.0920	0.0906	0.0000
0.4	0.2469	0.2600	0.2019	0.2085	0.0000
0.65	0.5458	0.5436	0.4785	0.4934	0.0000
			Panel C: 5/5/25		
0.0	0.0544	0.0449	0.0356	0.0351	0.0000
0.1	0.0744	0.0713	0.0552	0.0529	0.0000
0.25	0.1988	0.2322	0.1675	0.1698	0.0000
0.4	0.4427	0.5055	0.3726	0.3885	0.0000
0.65	0.8406	0.8618	0.7286	0.7505	0.0000
			Panel D: 5/25/25		
0.0	0.0521	0.0467	0.0447	0.0429	0.0000
0.1	0.0782	0.0821	0.0713	0.0705	0.0000
0.25	0.2359	0.3009	0.2262	0.2298	0.0000
0.4	0.5284	0.6432	0.4989	0.5024	0.0000
0.65	0.9133	0.9548	0.8571	0.8771	0.0000

both the Brown–Forsythe and the Welch tests when the sample sizes are unequal. When all the sample sizes are large, there are no notable differences between the ANOVA and the Brown–Forsythe test, but the Welch test performs slightly worse with medium and large effect sizes. Under small and equal sample sizes, the ANOVA generally performs slightly better than both the Brown–Forsythe and the Welch tests, particularly when the effect becomes larger.

As in the exponential case, however, the Kruskal–Wallis test dominates the competitors for all combinations of effect size and sample sizes, except for the specific case in which the sample sizes are large, the effect size is very large and, the parent populations are U-quadratic; this is achieved without a heightened probability of type I errors.

3.5. Sampling from Laplace distribution

In Table 8, all simulation results in which the parent populations are characterized by a Laplace distribution are shown. The Laplace distribution can be used as an example of a distribution that severely violates the assumption of a normal distribution owing to its leptokurtic shape, even though this shape is symmetric around a single peak.

Unlike the previous case of non-normal parent distributions, under a Laplace distribution when the sample sizes are unequal, the Welch test (as well as any other method) is not characterized by an increased probability of type I errors. It even seems that both the Brown–Forsythe and the Welch tests become too conservative when small sample sizes are included. In most cases, the ANOVA performs as well as or better than the Brown–Forsythe and Welch tests. This is particularly true with larger effect sizes when small samples are included, a case in which the ANOVA displays great superiority.

However, as in the other distinct non-normal cases, the Kruskal–Wallis test dominates in terms of performance. It has significantly more power than its competitors for almost all combinations of effect size and sample size.

4. Results with regard to the degree sample normality

In this section, simulation data filtered through the Kolmogorov–Smirnov procedure are used to evaluate the performance of the four test models, each with five different effect sizes and four different combinations of sample sizes. In each case, 100,000 sets of samples from three groups were simulated and sorted on the Kolmogorov–Smirnov statistic, using the sample mean and sample standard deviation to calculate the statistic in each case. Ten thousand sets of samples with the highest degree of sample normality (e.g. the lowest value of the Kolmogorov–Smirnov statistic) and 10,000 sets of samples with the lowest degree of sample normality were used to evaluate the sensitivity to the degree of sample normality of the four different test models.

To offer a better understanding of the reliability of the statistics presented in this section, it should be noted note that the standard error of a sample proportion at a sample size of 10,000 is about 0.005 when the proportion is 0.5, and decreases to about 0.002 when the proportion is 0.05 or 0.95.

4.1. Sampling from a normal distribution

In Table 9, all simulation results in which the parent populations are characterized by a normal distribution are shown.

When parent populations are normal, all test methods seem relatively insensitive to the degree of sample normality. There are no marked deviations from the unfiltered measures presented in Table 4 in the previous section. The Kruskal–Wallis test, however, is characterized by a slight increase in power as well as in the probability of type I errors when the degree of sample normality is low. In general, there are no notable differences in power between the ANOVA and the Brown–Forsythe or Welch tests when all sample sizes are large, but the ANOVA is superior to both of these when at least one sample size is small.

4.2. Sampling from a uniform distribution

In Table 10, all simulation results in which the parent populations are characterized by a uniform distribution are shown.

While the probability of a type I error was close to the nominal significance level in most cases in which unfiltered measures were used to evaluate the test methods under a uniform distribution, we can now see that this probability generally is significantly

Table 9. Performance under normal distribution with respect to sample normality

	10% 'most normal' samples					10% 'least normal' samples				
Effect size, f	ANOVA	Kruskal–Wallis	Brown–Forsythe	Welch	p-value (chi-square)	ANOVA	Kruskal–Wallis	Brown–Forsythe	Welch	p-value (chi-square)
Panel A: 25/25/25										
0.0	0.0490	0.0431	0.0483	0.0502	0.1032	0.0511	0.0724	0.0508	0.0498	0.0000
0.1	0.1098	0.0940	0.1094	0.1071	0.0012	0.1100	0.1344	0.1094	0.1089	0.0000
0.25	0.4594	0.4236	0.4579	0.4503	0.0003	0.4683	0.4802	0.4666	0.4575	0.1342
0.4	0.8684	0.8465	0.8674	0.8632	0.3078	0.8664	0.8605	0.8656	0.8597	0.9380
0.65	0.9991	0.9984	0.9991	0.9988	0.9999	0.9993	0.9992	0.9993	0.9992	1.0000
Panel B: 5/5/5										
0.0	0.0518	0.0405	0.0419	0.0416	0.0003	0.0542	0.0675	0.0433	0.0441	0.0000
0.1	0.0597	0.0470	0.0477	0.0490	0.0001	0.0648	0.0778	0.0523	0.0524	0.0000
0.25	0.1088	0.0908	0.0913	0.0895	0.0000	0.1173	0.1296	0.0980	0.0956	0.0000
0.4	0.2132	0.1753	0.1831	0.1700	0.0000	0.2230	0.2325	0.1915	0.1793	0.0000
0.65	0.4985	0.4350	0.4481	0.4128	0.0000	0.4997	0.4857	0.4503	0.4192	0.0000
Panel C: 5/5/25										
0.0	0.0515	0.0392	0.0471	0.0552	0.0000	0.0560	0.0665	0.0512	0.0576	0.0001
0.1	0.0700	0.0562	0.0621	0.0707	0.0001	0.0754	0.0846	0.0659	0.0703	0.0000
0.25	0.1866	0.1532	0.1499	0.1477	0.0000	0.1957	0.2046	0.1580	0.1545	0.0000
0.4	0.4221	0.3628	0.3277	0.3099	0.0000	0.4248	0.4249	0.3400	0.3191	0.0000
0.65	0.8449	0.7954	0.7259	0.6934	0.0000	0.8439	0.8200	0.7198	0.6889	0.0000
Panel D: 5/25/25										
0.0	0.0499	0.0419	0.0501	0.0549	0.0005	0.0516	0.0670	0.0540	0.0549	0.0000
0.1	0.0730	0.0596	0.0712	0.0740	0.0002	0.0752	0.0941	0.0784	0.0762	0.0000
0.25	0.2268	0.1952	0.2122	0.1943	0.0000	0.2341	0.2520	0.2173	0.2011	0.0000
0.4	0.5205	0.4741	0.4677	0.4386	0.0000	0.5137	0.5207	0.4641	0.4379	0.0000
0.65	0.9247	0.9021	0.8694	0.8618	0.0000	0.9193	0.9108	0.8696	0.8638	0.0000

Table 10. Performance under uniform distribution with respect to sample normality

	10% 'most normal' samples					10% 'least normal' samples				
Effect size, f	ANOVA	Kruskal–Wallis	Brown–Forsythe	Welch	p-value (chi-square)	ANOVA	Kruskal–Wallis	Brown–Forsythe	Welch	p-value (chi-square)
Panel A: 25/25/25										
0.0	0.0264	0.0250	0.0264	0.0287	0.4493	0.1201	0.1167	0.1194	0.1216	0.7878
0.1	0.0795	0.0708	0.0792	0.0774	0.0922	0.1839	0.1945	0.1832	0.1800	0.0928
0.25	0.4923	0.4231	0.4916	0.4713	0.0000	0.4668	0.4972	0.4661	0.4564	0.0002
0.4	0.9376	0.8884	0.9374	0.9287	0.0005	0.8011	0.7904	0.8004	0.7906	0.7235
0.65	0.9999	0.9999	0.9999	0.9999	1.0000	0.9964	0.9855	0.9963	0.9966	0.8248
Panel B: 5/5/5										
0.0	0.0500	0.0447	0.0393	0.0549	0.0000	0.1005	0.0860	0.0817	0.0769	0.0000
0.1	0.0600	0.0505	0.0473	0.0609	0.0000	0.1132	0.0966	0.0949	0.0862	0.0000
0.25	0.1084	0.0879	0.0908	0.0974	0.0000	0.1678	0.1542	0.1483	0.1279	0.0000
0.4	0.2047	0.1710	0.1794	0.1677	0.0000	0.2627	0.2427	0.2366	0.2034	0.0000
0.65	0.5138	0.4255	0.4808	0.3814	0.0000	0.4938	0.4436	0.4632	0.4020	0.0000
Panel C: 5/5/25										
0.0	0.0426	0.0426	0.0503	0.0809	0.0000	0.0970	0.0936	0.0856	0.0991	0.0104
0.1	0.0621	0.0594	0.0661	0.0938	0.0000	0.1195	0.1188	0.1061	0.1115	0.0135
0.25	0.1731	0.1459	0.1426	0.1515	0.0000	0.2403	0.2501	0.1954	0.1779	0.0000
0.4	0.4376	0.3500	0.3213	0.2792	0.0000	0.4391	0.4224	0.3551	0.3122	0.0000
0.65	0.9164	0.8428	0.8074	0.6922	0.0000	0.7791	0.6995	0.6871	0.6264	0.0000
Panel D: 5/25/25										
0.0	0.0371	0.0359	0.0527	0.0691	0.0000	0.1059	0.1039	0.1070	0.1054	0.9252
0.1	0.0589	0.0557	0.0713	0.0807	0.0000	0.1410	0.1409	0.1375	0.1278	0.0371
0.25	0.2122	0.1782	0.1910	0.1737	0.0000	0.2904	0.3047	0.2639	0.2327	0.0000
0.4	0.5642	0.4726	0.4750	0.4151	0.0000	0.5166	0.5209	0.4680	0.4186	0.0000
0.65	0.9707	0.9399	0.9475	0.9213	0.0041	0.8577	0.8193	0.8212	0.7890	0.0000

affected by the degree of sample normality. When all sample sizes are large, all methods are too conservative when the degree of sample normality is high and too liberal when the degree of sample normality is low. A low degree of sample normality generally seems to correspond to an increased probability of a type I error for all methods with all combinations of sample sizes. The Welch test, however, also has an increased probability of a type I error with unequal sample sizes under a high degree of sample normality. Owing to the high level of conservatism they exhibit when samples have a high degree of normality, all methods are more powerful with small effect sizes when the degree of sample normality is low. However, they all become more powerful with larger effect sizes when the degree of sample normality is high. This tendency can be observed for all combinations of sample sizes.

In general, there are no notable differences in power between the ANOVA and the Brown–Forsythe or Welch tests when all sample sizes are large, but the ANOVA is superior to both of these when at least one sample size is small. Furthermore, as in the case in which unfiltered data were used, the Kruskal–Wallis test is less powerful than the ANOVA, and the Brown–Forsythe test is generally more powerful than the Welch in all situations when the parent population is uniform.

4.3. Sampling from an exponential distribution

In Table 11, all simulation results in which the parent populations are characterized by an exponential distribution are shown.

When unfiltered data from exponential distributions were used in the previous section, it should be noted that the probability of a type I error was close to the significance level for all test methods. However, when the samples with the highest degree of normality were compared with the samples with the lowest degree of normality, there are significant differences. The ANOVA as well as the Brown–Forsythe and the Welch tests are all too conservative when the degree of sample normality is low and too liberal when the degree of sample normality is high, for all combinations of sample sizes. Hence, under an exponential distribution, all three methods are highly sensitive to the degree of sample normality. The Kruskal–Wallis test also has an increased probability of a type I error when the degree of sample normality is high, but this is less pronounced than in the other methods in all cases, except when all the sample sizes are small. When the degree of sample normality is low, the Kruskal–Wallis test seems to have a neutral probability of a type I error, except in the case in which all the sample sizes are small.

Another significant difference between the Kruskal–Wallis test and the other three models was revealed by studying the differences in their power. The Kruskal–Wallis test is much more powerful when the degree of sample normality is low, partly because of the high degree of conservatism that characterizes the other models in that case. With small effect sizes and a high degree of sample normality, the Kruskal–Wallis test seems less powerful, particularly for large sample sizes, but the reason for this is the high degree of liberalism that characterizes the other models. For larger effect sizes, the Kruskal–Wallis test again dominates when the effect of this liberalism diminishes.

The most notable difference between the Brown–Forsythe test and the Welch test is that the Welch test is generally more liberal. This effect is particularly obvious when the sample sizes are unequal, and the performance of the ANOVA then lies between that of the Brown–Forsythe and Welch tests in many cases.

Table 11. Performance under exponential distribution with respect to sample normality

	10% 'most normal' samples					10% 'least normal' samples				
Effect size, f	ANOVA	Kruskal–Wallis	Brown–Forsythe	Welch	p-value (chi-square)	ANOVA	Kruskal–Wallis	Brown–Forsythe	Welch	p-value (chi-square)
Panel A: 25/25/25										
0.0	0.1389	0.0745	0.1364	0.1366	0.0000	0.0091	0.0565	0.0085	0.0113	0.0000
0.1	0.2610	0.2229	0.2590	0.2565	0.0000	0.0330	0.2695	0.0317	0.0355	0.0000
0.25	0.6948	0.7339	0.6933	0.6952	0.0008	0.2678	0.9006	0.2643	0.2756	0.0000
0.4	0.9520	0.9707	0.9515	0.9573	0.4744	0.7095	0.9980	0.7065	0.7084	0.0000
0.65	0.9997	1.0000	0.9996	1.0000	1.0000	0.9885	1.0000	0.9876	0.9915	0.8095
Panel B: 5/5/5										
0.0	0.1048	0.0703	0.0580	0.0677	0.0000	0.0050	0.0267	0.0030	0.0053	0.0000
0.1	0.1342	0.1015	0.0795	0.0954	0.0000	0.0073	0.0348	0.0039	0.0075	0.0000
0.25	0.2833	0.2480	0.2135	0.2259	0.0000	0.0220	0.0876	0.0133	0.0230	0.0000
0.4	0.4936	0.4528	0.4283	0.4197	0.0000	0.0680	0.1772	0.0446	0.0666	0.0000
0.65	0.7796	0.7564	0.7382	0.7373	0.0013	0.2558	0.3648	0.1925	0.2357	0.0000
Panel C: 5/5/25										
0.0	0.1215	0.0677	0.1150	0.1797	0.0000	0.0114	0.0470	0.0174	0.0272	0.0000
0.1	0.1627	0.1662	0.2005	0.2904	0.0000	0.0084	0.1962	0.0424	0.0665	0.0000
0.25	0.4286	0.4383	0.3946	0.4919	0.0000	0.0480	0.5696	0.1216	0.1801	0.0000
0.4	0.7168	0.6971	0.5969	0.6737	0.0000	0.2241	0.8196	0.2316	0.3295	0.0000
0.65	0.9471	0.9251	0.8446	0.8796	0.0000	0.6744	0.9600	0.4693	0.5793	0.0000
Panel D: 5/25/25										
0.0	0.1307	0.0740	0.1397	0.1669	0.0000	0.0122	0.0501	0.0106	0.0367	0.0000
0.1	0.1831	0.1664	0.2337	0.2624	0.0000	0.0130	0.1876	0.0263	0.0782	0.0000
0.25	0.4677	0.4844	0.4739	0.4864	0.1791	0.0810	0.6491	0.1368	0.1805	0.0000
0.4	0.7723	0.7906	0.7097	0.7352	0.0000	0.3056	0.9323	0.3333	0.3454	0.0000
0.65	0.9753	0.9795	0.9213	0.9578	0.0001	0.7909	0.9983	0.6512	0.7043	0.0000

4.4. Sampling from a U-quadratic distribution

In Table 12, all simulation results in which the parent populations are characterized by a U-quadratic distribution are shown.

With large sample sizes, the probabilities of type I errors are reversed in comparison to the exponential distribution. When the parent populations are U-quadratic and all samples are large, a high degree of sample normality generates a high degree of conservatism in all methods, while the probability of a type I error is large when the degree of sample normality is low. On the other hand, when at least one sample size is low, the probability of a type I error is large or very large in the case of a high degree of sample normality, as well in the opposite case. Interestingly, when at least one sample size is low, the Welch test is the least liberal model in all cases of a low degree of normality, but the most liberal model in all cases of a high degree of normality. On average, the Brown–Forsythe test seems to be the model least affected by the degree of sample normality when at least one sample size is small. When all the sample sizes are large, on the other hand, the Kruskal–Wallis seems to be the least sensitive model from this perspective. The reason for this is probably that the Kruskal–Wallis test generally is much more powerful than the other three models when the parent populations are U-quadratic, without having the downside of a higher probability of a type I error, as was previously shown in Table 7.

4.5. Sampling from a Laplace distribution

In Table 13, all simulation results in which the parent populations are characterized by a Laplace distribution are shown.

The Laplace distribution seems to resemble a normal distribution, but it has a significant excess kurtosis. Hence, outliers are much more common in the Laplace distribution. Thus, it is easy to imagine the main difference between a sample from a Laplace distribution with a high degree of normality and a sample with a low degree of normality. As one would expect, the pattern for samples with a high degree of normality closely follows the pattern found in Table 9 for the normal distribution: the ANOVA has a slightly higher power than the other three models for all combinations of effect size and sample size. The Kruskal–Wallis test and the ANOVA, however, both dominate the Brown–Forsythe and Welch tests in the case of larger effect sizes when the sample sizes are unequal.

When the degree of normality is low, on the other hand, we observe that the Kruskal–Wallis test becomes somewhat too liberal for all sample size combinations, whereas the other three tests are too conservative when at least one sample size is small, and the Kruskal–Wallis test is slightly too liberal when all sample sizes are large. While the ANOVA has greater power than the Brown–Forsythe and Welch tests when at least one sample size is small, the Kruskal–Wallis test outperforms all three of the other test methods.

5. Discussion

The main difference between this study and most previous research that compares the omnibus one-way ANOVA with alternative test methods is that we have not used only the overall performance measures with a certain effect size and/or a certain combination of sample sizes as the basis for a final verdict on their relative performance. Instead, we have taken the analysis a step further by investigating the sensitivity of the different

Table 12. Performance under U-quadratic distribution with respect to sample normality

	10% 'most normal' samples					10% 'least normal' samples				
Effect size, *f*	ANOVA	Kruskal–Wallis	Brown–Forsythe	Welch	*p*-value (chi-square)	ANOVA	Kruskal–Wallis	Brown–Forsythe	Welch	*p*-value (chi-square)
Panel A: 25/25/25										
0.0	0.0035	0.0034	0.0035	0.0037	0.9874	0.3569	0.2824	0.3565	0.3488	0.0000
0.1	0.0208	0.0951	0.0206	0.0207	0.0000	0.3981	0.5584	0.3977	0.3893	0.0000
0.25	0.3930	0.7779	0.3926	0.3669	0.0000	0.5658	0.8202	0.5654	0.5547	0.0000
0.4	0.9550	0.9748	0.9546	0.9484	0.2488	0.7905	0.8905	0.7903	0.7807	0.0000
0.65	1.0000	1.0000	1.0000	1.0000	1.0000	0.9962	0.9414	0.9962	0.9963	0.0000
Panel B: 5/5/5										
0.0	0.2239	0.1947	0.1699	0.2130	0.0000	0.1584	0.1671	0.1365	0.0685	0.0000
0.1	0.2474	0.2367	0.1900	0.2376	0.0000	0.2227	0.2783	0.2028	0.0885	0.0000
0.25	0.3242	0.3255	0.2731	0.3301	0.0000	0.3536	0.4388	0.3409	0.2232	0.0000
0.4	0.3878	0.3798	0.3531	0.3779	0.0004	0.4444	0.4997	0.4313	0.3355	0.0000
0.65	0.5732	0.5002	0.5459	0.4850	0.0000	0.6029	0.5429	0.5909	0.5185	0.0000
Panel C: 5/5/25										
0.0	0.1685	0.1390	0.2036	0.4383	0.0000	0.2167	0.1956	0.1434	0.0859	0.0000
0.1	0.1984	0.1933	0.2283	0.4200	0.0000	0.2579	0.4124	0.1932	0.1229	0.0000
0.25	0.2779	0.2829	0.2899	0.3660	0.0000	0.4106	0.6451	0.3556	0.2572	0.0000
0.4	0.4513	0.4180	0.3885	0.3870	0.0000	0.6093	0.6803	0.5595	0.4331	0.0000
0.65	0.8817	0.7411	0.8508	0.7613	0.0000	0.7691	0.7016	0.7334	0.6760	0.0000
Panel D: 5/25/25										
0.0	0.0727	0.0623	0.2229	0.2808	0.0000	0.2826	0.2316	0.2563	0.1982	0.0000
0.1	0.1133	0.1298	0.2067	0.2800	0.0000	0.3231	0.4429	0.2977	0.2351	0.0000
0.25	0.2083	0.3480	0.2237	0.2754	0.0000	0.4414	0.7473	0.4261	0.3744	0.0000
0.4	0.5023	0.6958	0.4609	0.3794	0.0000	0.6354	0.8310	0.6126	0.5463	0.0000
0.65	0.9662	0.9492	0.9770	0.9439	0.0630	0.8436	0.8545	0.8361	0.7996	0.0001

Table 13. Performance under Laplace distribution with respect to sample normality

	10% 'most normal' samples					10% 'least normal' samples				
Effect size, f	ANOVA	Kruskal–Wallis	Brown–Forsythe	Welch	p-value (chi-square)	ANOVA	Kruskal–Wallis	Brown–Forsythe	Welch	p-value (chi-square)
Panel A: 25/25/25										
0.0	0.0448	0.0394	0.0443	0.0444	0.2083	0.0570	0.0680	0.0557	0.0538	0.0001
0.1	0.1184	0.1107	0.1173	0.1181	0.3293	0.1118	0.2019	0.1103	0.1099	0.0000
0.25	0.5440	0.5419	0.5419	0.5487	0.9040	0.3985	0.7301	0.3961	0.4020	0.0000
0.4	0.9241	0.9218	0.9232	0.9188	0.9815	0.7703	0.9712	0.7682	0.7660	0.0000
0.65	0.9994	0.9996	0.9994	0.9992	1.0000	0.9891	0.9999	0.9890	0.9885	0.8199
Panel B: 5/5/5										
0.0	0.0525	0.0361	0.0382	0.0316	0.0000	0.0178	0.0740	0.0130	0.0117	0.0000
0.1	0.0640	0.0473	0.0485	0.0431	0.0000	0.0258	0.0950	0.0188	0.0164	0.0000
0.25	0.1323	0.1101	0.1078	0.1069	0.0000	0.0609	0.1856	0.0442	0.0454	0.0000
0.4	0.2672	0.2372	0.2220	0.2365	0.0000	0.1546	0.3184	0.1167	0.1289	0.0000
0.65	0.5927	0.5484	0.5252	0.5357	0.0000	0.4201	0.5402	0.3512	0.3673	0.0000
Panel C: 5/5/25										
0.0	0.0588	0.0367	0.0452	0.0414	0.0000	0.0325	0.0669	0.0198	0.0185	0.0000
0.1	0.0817	0.0579	0.0668	0.0650	0.0000	0.0483	0.1104	0.0311	0.0303	0.0000
0.25	0.2293	0.1959	0.1853	0.1899	0.0000	0.1516	0.3161	0.1070	0.1043	0.0000
0.4	0.5041	0.4786	0.4153	0.4309	0.0000	0.3513	0.5814	0.2624	0.2773	0.0000
0.65	0.8999	0.8776	0.7773	0.8039	0.0000	0.7435	0.8645	0.6045	0.6317	0.0000
Panel D: 5/25/25										
0.0	0.0550	0.0386	0.0471	0.0484	0.0000	0.0458	0.0665	0.0363	0.0283	0.0000
0.1	0.0847	0.0665	0.0772	0.0784	0.0001	0.0698	0.1232	0.0554	0.0484	0.0000
0.25	0.2786	0.2619	0.2607	0.2529	0.0040	0.1980	0.4094	0.1706	0.1670	0.0000
0.4	0.6079	0.5991	0.5588	0.5592	0.0000	0.4411	0.7506	0.3893	0.3984	0.0000
0.65	0.9557	0.9513	0.8978	0.9234	0.0000	0.8397	0.9747	0.7711	0.7921	0.0000

methods to different degrees of sample normality. As a result, we have been able to show that there are big differences between the sensitivities of these methods. It is therefore important to take these differences into account when choosing a method with which to compare population means in an empirical study. By first testing data for normality, and then applying the ANOVA or another parametric method if the assumption of population normality could not be rejected owing to insufficient sample non-normality, the researcher accepts a familywise effect that changes the probability of error in the overall test procedure. For example, samples from distinct non-normal parent distributions are sometimes characterized by a degree of normality that is high enough not to reject the null hypothesis of population normality, which means that a test method with a lower actual power can be chosen. The results in this study can be used to quantify the impact of this problem.

The results in this study also highlight the importance of pilot studies and/or the analysis of previous research to obtain a preliminary estimate of the actual effect size, as well as to provide a better understanding of the true distribution of the parent population. To simply collect data and 'open-mindedly' test them for normality is generally an inadequate approach, since the wrong choice of test method may lead to a significantly heightened risk of either a type I or type II error. Suppose, for example, that there is no difference between the true means of three populations and the parent populations in reality are exponentially distributed, but the degree of sample normality is not low enough not to reject the null hypothesis of normal distribution in any parent population. In this case, the choice of the ANOVA, the Brown–Forsythe test, or the Welch test instead of the Kruskal–Wallis test to analyse these data could entail almost twice the risk of a type I error, depending on the sample sizes. Now suppose that the true difference between the means in three populations characterized by U-quadratic distributions corresponds to a medium effect size, the sample sizes are 25 in each group, and the degree of sample normality is not low enough not to reject the null hypothesis of normal distribution in any parent population. In this case, the choice of the ANOVA, the Brown–Forsythe test, or the Welch test instead of the Kruskal–Wallis test to analyse these data could entail an almost tripled risk of a type II error.

6. Conclusion

In many cases in which true parent distributions are unknown or where there may be reason to believe that they are characterized by a distinct non-normality, even though the sample data are not significantly non-normal, the results from this study suggest that the Kruskal–Wallis test is the best choice to test for differences in location. The main argument for this is that the difference between the performance of approximately normal parent populations and distinctly non-normal parent populations from which 'rather normal' samples have been collected is much larger for the ANOVA and its parametric counterparts. Furthermore, since the Kruskal–Wallis test does not rely on a preliminary normality test, the researcher avoids the familywise error problem for which the significance level must be corrected when performing an ANOVA after a normality test, even though such correction is rare in practice.

Hence, the general advice suggested by this study would be for researchers to use the ANOVA, the Brown–Forsythe test, or the Welch test, depending on the degree of heteroscedasticity, to test for differences between locations in populations that are known for a fact to be normal or approximately normal. However, even if the sample data

have a relatively good normal fit, the Kruskal–Wallis test should be used if the underlying populations might be distinctly non-normal or if their true shapes are unknown.

One limitation of this study is that, by design, we have not examined cases involving heteroscedasticity. The reason for this, of course, is that we have focused exclusively on the importance of sample normality under different types of population non-normality. Future research should aim to extend this kind of analysis in order to include violations of the assumption of homoscedasticity.

References

Box, G. E. P., & Muller, M. E. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, *29*(2), 610–611. doi: 10.2307/2237361

Brown, M. B., & Forsythe, A. B. (1974). The ANOVA and multiple comparisons for data with heterogeneous variances. *Biometrics*, *30*, 719–724. doi: 10.2307/2529238

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

Devroye, L. (1986). *Non-uniform random variate generation*. New York: Springer.

Edgington, E. S., & Onghena, P. (2007). *Randomization tests*. Boca Raton, FL: Taylor & Francis.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G * Power 3: A flexible statistical power analysis for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. doi: 10.3758/BF03193146

Feir, B. J., & Toothaker, L. E. (1974). *The ANOVA F-test versus the Kruskal-Wallis test: A robustness study*. Paper presented to the Annual Meeting of the American Educational Research Association, Chicago.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237–288. doi: 10.2307/1169991

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational and Behavioral Statistics*, *17*, 315–339. doi: 10.2307/1165127

Khan, A., & Rayner, G. D. (2003). Robustness to non-normality of common tests for the many-sample location problem. *Journal of Applied Mathematics and Decision Sciences*, *7*(4), 187–206. doi: 10.1155/S1173912603000178

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

Mandelbrot, B., & Taleb, N. (2006). A focus on the exceptions that prove the rule. *Financial Times*, 23 March.

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

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *6*, 147–151. doi: 10.1027/1614-2241/a000016

Tanizaki, H. (1997). Power comparison of non-parametric tests: Small-sample properties from Monte Carlo experiments. *Journal of Applied Statistics*, *24*, 603–632. doi: 10.1080/02664769723576

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

Van Hecke, T. (2010). Power study of anova versus Kruskal-Wallis test. *In Practice*, 1–6. Retrieved from http://interstat.statjournals.net/YEAR/2010/articles/1011002.pdf

Watthanacheewakul, L. (2011). *Comparisons of power of parametric and nonparametric test for testing means of several Weibull populations*. Hong Kong: IMECS.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*, 330–336. doi: 10.2307/2332579

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education*, *67*(1), 55–68. doi: 10.1080/00220979809598344

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, *57*, 173–181. doi: 10.1348/000711010X524739

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*, 388–409. doi: 10.1348/000711010X524739

Received 10 December 2011; revised version received 2 February 2012