Psicothema

Psicothema 2017, Vol. 29, No. 4, 552-557
doi: 10.7334/psicothema2016.383

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Non-normal data: Is ANOVA still a valid option?

María J. Blanca[1], Rafael Alarcón[1], Jaume Arnau[2], Roser Bono[2] and Rebecca Bendayan[1,3]
[1] Universidad de Málaga, [2] Universidad de Barcelona and [3] MRC Unit for Lifelong Health and Ageing, University College London

Abstract

Background: The robustness of F-test to non-normality has been studied from the 1930s through to the present day. However, this extensive body of research has yielded contradictory results, there being evidence both for and against its robustness. This study provides a systematic examination of F-test robustness to violations of normality in terms of Type I error, considering a wide variety of distributions commonly found in the health and social sciences. **Method:** We conducted a Monte Carlo simulation study involving a design with three groups and several known and unknown distributions. The manipulated variables were: Equal and unequal group sample sizes; group sample size and total sample size; coefficient of sample size variation; shape of the distribution and equal or unequal shapes of the group distributions; and pairing of group size with the degree of contamination in the distribution. **Results:** The results showed that in terms of Type I error the F-test was robust in 100% of the cases studied, independently of the manipulated conditions.

Keywords: *F*-test, ANOVA, robustness, skewness, kurtosis.

Resumen

Datos no normales: ¿es el ANOVA una opción válida? **Antecedentes:** las consecuencias de la violación de la normalidad sobre la robustez del estadístico F han sido estudiadas desde 1930 y siguen siendo de interés en la actualidad. Sin embargo, aunque la investigación ha sido extensa, los resultados son contradictorios, encontrándose evidencia a favor y en contra de su robustez. El presente estudio presenta un análisis sistemático de la robustez del estadístico F en términos de error de Tipo I ante violaciones de la normalidad, considerando una amplia variedad de distribuciones frecuentemente encontradas en ciencias sociales y de la salud. **Método:** se ha realizado un estudio de simulación Monte Carlo considerando un diseño de tres grupos y diferentes distribuciones conocidas y no conocidas. Las variables manipuladas han sido: igualdad o desigualdad del tamaño de los grupos, tamaño muestral total y de los grupos; coeficiente de variación del tamaño muestral; forma de la distribución e igualdad o desigualdad de la forma en los grupos; y emparejamiento entre el tamaño muestral con el grado de contaminación en la distribución. **Resultados:** los resultados muestran que el estadístico F es robusto en términos de error de Tipo I en el 100% de los casos estudiados, independientemente de las condiciones manipuladas.

Palabras clave: estadístico *F*, ANOVA, robustez, asimetría, curtosis.

One-way analysis of variance (ANOVA) or F-test is one of the most common statistical techniques in educational and psychological research (Keselman et al., 1998; Kieffer, Reese, & Thompson, 2001). The *F*-test assumes that the outcome variable is normally and independently distributed with equal variances among groups. However, real data are often not normally distributed and variances are not always equal. With regard to normality, Micceri (1989) analyzed 440 distributions from ability and psychometric measures and found that most of them were contaminated, including different types of tail weight (uniform to double exponential) and different classes of asymmetry. Blanca, Arnau, López-Montiel, Bono, and Bendayan (2013) analyzed 693 real datasets from psychological variables and found that 80% of them presented values of skewness and kurtosis ranging between -1.25 and 1.25, with extreme departures from the normal distribution being infrequent. These results were consistent with other studies with real data (e.g., Harvey & Siddique, 2000; Kobayashi, 2005; Van Der Linder, 2006).

The effect of non-normality on *F*-test robustness has, since the 1930s, been extensively studied under a wide variety of conditions. As our aim is to examine the independent effect of non-normality the literature review focuses on studies that assumed variance homogeneity. Monte Carlo studies have considered unknown and known distributions such as mixed non-normal, lognormal, Poisson, exponential, uniform, chi-square, double exponential, Student's t, binomial, gamma, Cauchy, and beta (Black, Ard, Smith, & Schibik, 2010; Bünning, 1997; Clinch & Kesselman, 1982; Feir-Walsh & Thoothaker, 1974; Gamage & Weerahandi, 1998; Lix, Keselman, & Keselman, 1996; Patrick, 2007; Schmider, Ziegler, Danay, Beyer, & Bühner, 2010).

One of the first studies on this topic was carried out by Pearson (1931), who found that *F*-test was valid provided that the deviation from normality was not extreme and the number of degrees of freedom apportioned to the residual variation was not too small. Norton (1951, cit. Lindquist, 1953) analyzed the effect of distribution shape on robustness (considering either that the distributions had the same shape in all the groups or a different shape in each group)

Received: December 14, 2016 • Accepted: June 20, 2017
Corresponding author: María J. Blanca
Facultad de Psicología
Universidad de Málaga
29071 Málaga (Spain)
e-mail: blamen@uma.es

and found that, in general, *F*-test was quite robust, the effect being negligible. Likewise, Tiku (1964) stated that distributions with skewness values in a different direction had a greater effect than did those with values in the same direction unless the degrees of freedom for error were fairly large. However, Glass, Peckham, and Sanders (1972) summarized these early studies and concluded that the procedure was affected by kurtosis, whereas skewness had very little effect. Conversely, Harwell, Rubinstein, Hayes, and Olds (1992), using meta-analytic techniques, found that skewness had more effect than kurtosis. A subsequent meta-analytic study by Lix et al. (1996) concluded that Type I error performance did not appear to be affected by non-normality.

These inconsistencies may be attributable to the fact that a standard criterion has not been used to assess robustness, thus leading to different interpretations of the Type I error rate. The use of a single and standard criterion such as that proposed by Bradley (1978) would be helpful in this context. According to Bradley's (1978) liberal criterion a statistical test is considered robust if the empirical Type I error rate is between .025 and .075 for a nominal alpha level of .05. In fact, had Bradley's criterion of robustness been adopted in the abovementioned studies, many of their results would have been interpreted differently, leading to different conclusions. Furthermore, when this criterion is considered, more recent studies provide empirical evidence for the robustness of *F*-test under non-normality with homogeneity of variances (Black et al., 2010; Clinch & Keselman, 1982; Feir-Walsh & Thoothaker, 1974; Gamage & Weerahandi, 1998; Kanji, 1976; Lantz, 2013; Patrick, 2007; Schmider et al., 2010; Zijlstra, 2004).

Based on most early studies, many classical handbooks on research methods in education and psychology draw the following conclusions: Moderate departures from normality are of little concern in the fixed-effects analysis of variance (Montgomery, 1991); violations of normality do not constitute a serious problem, unless the violations are especially severe (Keppel, 1982); *F*-test is robust to moderate departures from normality when sample sizes are reasonably large and are equal (Winer, Brown, & Michels, 1991); and researchers do not need to be concerned about moderate departures from normality provided that the populations are homogeneous in form (Kirk, 2013). To summarize, *F*-test is robust to departures from normality when: a) the departure is moderate; b) the populations have the same distributional shape; and c) the sample sizes are large and equal. However, these conclusions are broad and ambiguous, and they are not helpful when it comes to deciding whether or not *F*-test can be used. The main problem is that expressions such as "moderate", "severe" and "reasonably large sample size" are subject to different interpretations and, consequently, they do not constitute a standard guideline that helps applied researchers decide whether they can trust their *F*-test results under non-normality.

Given this situation, the main goals of the present study are to provide a systematic examination of *F*-test robustness, in terms of Type I error, to violations of normality under homogeneity using a standard criterion such as that proposed by Bradley (1978). Specifically, we aim to answer the following questions: Is *F*-test robust to slight and moderate departures from normality? Is it robust to severe departures from normality? Is it sensitive to differences in shape among the groups? Does its robustness depend on the sample sizes? Is its robustness associated with equal or unequal sample sizes?

To this end, we designed a Monte Carlo simulation study to examine the effect of a wide variety of distributions commonly found in the health and social sciences on the robustness of *F*-test. Distributions with a slight and moderate degree of contamination (Blanca et al., 2013) were simulated by generating distributions with values of skewness and kurtosis ranging between -1 and 1. Distributions with a severe degree of contamination (Micceri, 1989) were represented by exponential, double exponential, and chi-square with 8 degrees of freedom. In both cases, a wide range of sample sizes were considered with balanced and unbalanced designs and with equal and unequal distributions in groups. With unequal sample size and unequal shape in the groups, the pairing of group sample size with the degree of contamination in the distribution was also investigated.

Method

Instruments

We conducted a Monte Carlo simulation study with non-normal data using SAS 9.4. (SAS Institute, 2013). Non-normal distributions were generated using the procedure proposed by Fleishman (1978), which uses a polynomial transformation to generate data with specific values of skewness and kurtosis.

Procedure

In order to examine the effect of non-normality on *F*-test robustness, a one-way design with 3 groups and homogeneity of variance was considered. The group effect was set to zero in the population model. The following variables were manipulated:

1. Equal and unequal group sample sizes. Unbalanced designs are more common than balanced designs in studies involving one-way and factorial ANOVA (Golinski & Cribbie, 2009; Keselman et al., 1998). Both were considered in order to extend our results to different research situations.
2. Group sample size and total sample size. A wide range of group sample sizes were considered, enabling us to study small, medium, and large sample sizes. With balanced designs the group sizes were set to 5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, and 100, with total sample size ranging from 15 to 300. With unbalanced designs, group sizes were set between 5 and 160, with a mean group size of between 10 and 100 and total sample size ranging from 15 to 300.
3. Coefficient of sample size variation *(*Δn*)*, which represents the amount of inequality in group sizes. This was computed by dividing the standard deviation of the group sample size by its mean. Different degrees of variation were considered and were grouped as low, medium, and high. A low Δn was fixed at approximately 0.16 (0.141 - 0.178), a medium coefficient at 0.33 (0.316 - 0.334), and a high value at 0.50 (0.491 - 0.521). Keselman et al. (1998) showed that the ratio of the largest to the smallest group size was greater than 3 in 43.5% of cases. With $\Delta n = 0.16$ this ratio was equal to 1.5, with $\Delta n = 0.33$ it was equal to either 2.3 or 2.5, and with $\Delta n = 0.50$ it ranged from 3.3 to 5.7.
4. Shape of the distribution and equal and unequal shape in the groups. Twenty-two distributions were investigated, involving several degrees of deviation from normality and with both equal and unequal shape in the groups. For equal shape and slight and moderate departures from normality,

the distributions had values of skewness (γ_1) and kurtosis (γ_2) ranging between -1 and 1, these values being representative of real data (Blanca et al., 2013). The values of γ_1 and γ_2 are presented in Table 2 (distributions 1-12). For severe departures from normality, distributions had values of γ_1 and γ_2 corresponding to the double exponential, chi-square with 8 degrees of freedom, and exponential distributions (Table 2, distributions 13-15). For unequal shape, the values of γ_1 and γ_2 of each group are presented in Table 3. Distributions 16-21 correspond to slight and moderate departures from normality and distribution 22 to severe departure.

5. Pairing of group size with degree of contamination in the distribution. This condition was included with unequal shape and unequal sample size. The pairing was positive when the largest group size was associated with the greater contamination, and vice versa. The pairing was negative when the largest group size was associated with the smallest contamination, and vice versa. The specific conditions with unequal sample size are shown in Table 1.

Ten thousand replications of the 1308 conditions resulting from the combination of the above variables were performed at a significance level of .05. This number of replications was chosen to ensure reliable results (Bendayan, Arnau, Blanca, & Bono, 2014; Robey & Barcikowski, 1992).

Data analysis

Empirical Type I error rates associated with *F*-test were analyzed for each condition according to Bradley's robustness criterion (1978).

Results

Tables 2 and 3 show descriptive statistics for the Type I error rate across conditions for equal and unequal shapes. Although the tables do not include all available information (due to article length limitations), the maximum and minimum values are sufficient for assessing robustness. Full tables are available upon request from the corresponding author.

All empirical Type I error rates were within the bounds of Bradley's criterion. The results show that *F*-test is robust for 3 groups in 100% of cases, regardless of the degree of deviation from a normal distribution, sample size, balanced or unbalanced cells, and equal or unequal distribution in the groups.

Discussion

We aimed to provide a systematic examination of *F*-test robustness to violations of normality under homogeneity of variance, applying Bradley's (1978) criterion. Specifically, we sought to answer the following question: Is *F*-test robust, in terms of Type I error, to slight, moderate, and severe departures from normality, with various sample sizes (equal or unequal sample size) and with same or different shapes in the groups? The answer to this question is a resounding yes, since *F*-test controlled Type I error to within the bounds of Bradley's criterion. Specifically, the results show that *F*-test remains robust with 3 groups when distributions have values of skewness and kurtosis ranging between -1 and 1, as well as with data showing a greater departure from normality, such as the exponential, double exponential, and chi-squared (8) distributions. This applies even when sample sizes are very small (i.e., n= 5) and quite different in the groups, and also when the group distributions differ significantly. In addition, the test's robustness is independent of the pairing of group size with the degree of contamination in the distribution.

Our results support the idea that the discrepancies between studies on the effect of non-normality may be primarily attributed to differences in the robustness criterion adopted, rather than to the degree of contamination of the distributions. These findings highlight the need to establish a standard criterion of robustness to clarify the potential implications when performing Monte Carlo studies. The present analysis made use of Bradley's criterion, which has been argued to be one of the most suitable criteria for

Table 1

Specific conditions studied under non-normality for unequal shape in the groups as a function of total sample size (N), means group size (N/J), coefficient of sample size variation (Δn), and pairing of group size with the degree of distribution contamination: (+) the largest group size is associated with the greater contamination and vice versa, and (-) the largest group size is associated with the smallest contamination and vice versa

			n Pairing	
N	**N/J**	**Δn**	**+**	**–**
30	10	0.16	8, 10, 12	12, 10, 8
		0.33	6, 10, 14	14, 10, 6
		0.50	5, 8, 17	17, 8, 5
45	15	0.16	12, 15, 18	18, 15, 12
		0.33	9, 15, 21	21, 15, 9
		0.50	6, 15, 24	24, 15, 6
60	20	0.16	16, 20, 24	24, 20, 16
		0.33	12, 20, 28	28, 20, 12
		0.50	8, 20, 32	32, 20, 8
75	25	0.16	20, 25, 30	30, 25, 20
		0.33	15, 25, 35	35, 25, 15
		0.50	10, 25, 40	40, 25, 10
90	30	0.16	24, 30, 36	36, 30, 24
		0.33	18, 30, 42	42, 30, 18
		0.50	12, 30, 48	48, 30, 12
120	40	0.16	32, 40, 48	48, 40, 32
		0.33	24, 40, 56	56, 40, 24
		0.50	16, 40, 64	64, 40, 16
150	50	0.16	40, 50, 60	60, 50, 40
		0.33	30, 50, 70	70, 50, 30
		0.50	20, 50, 80	80, 50, 20
180	60	0.16	48, 60, 72	72, 60, 48
		0.33	36, 60, 84	84, 60, 36
		0.50	24, 60, 96	96, 60, 24
210	70	0.16	56, 70, 84	84, 70, 56
		0.33	42, 70, 98	98, 70, 42
		0.50	28, 70, 112	112, 70, 28
240	80	0.16	64, 80, 96	96, 80, 64
		0.33	48, 80, 112	112, 80, 48
		0.50	32, 80, 128	128, 80, 32
270	90	0.16	72, 90, 108	108, 90, 72
		0.33	54, 90, 126	126, 90, 54
		0.50	36, 90, 144	144, 90, 36
300	100	0.16	80, 100, 120	120, 100, 80
		0.33	60, 100, 140	140, 100, 60
		0.50	40, 100, 160	160, 100, 40

examining the robustness of statistical tests (Keselman, Algina, Kowalchuk, & Wolfinger, 1999). In this respect, our results are consistent with previous studies whose Type I error rates were within the bounds of Bradley's criterion under certain departures from normality (Black et al., 2010; Clinch & Keselman, 1982; Feir-Walsh & Thoothaker, 1974; Gamage & Weerahandi, 1998; Kanji, 1976; Lantz, 2013; Lix et al., 1996; Patrick, 2007; Schmider et al., 2010; Zijlstra, 2004). By contrast, however, our results do not concur, at least for the conditions studied here, with those classical handbooks which conclude that *F*-test is only robust if the departure from normality is moderate (Keppel, 1982; Montgomery, 1991), the populations have the same distributional shape (Kirk, 2013), and the sample sizes are large and equal (Winer et al., 1991).

Our findings are useful for applied research since they show that, in terms of Type I error, *F*-test remains a valid statistical procedure under non-normality in a variety of conditions. Data transformation or nonparametric analysis is often recommended when data are not normally distributed. However, data transformations offer no additional benefits over the good control of Type I error achieved by *F*-test. Furthermore, it is usually difficult to determine which transformation is appropriate for a set of data, and a given transformation may not be applicable when groups differ in shape. In addition, results are often difficult to interpret when data transformations are adopted. There are also disadvantages to using non-parametric procedures such as the Kruskal-Wallis test. This test converts quantitative continuous data into rank-ordered data, with a consequent loss of information. Moreover, the null hypothesis associated with the Kruskal-Wallis test differs from that of *F*-test, unless the distribution of groups has exactly the same shape (see Maxwell & Delaney, 2004). Given these limitations, there is no reason to prefer the Kruskal-Wallis test under the conditions studied in the present paper. Only with equal shape in the groups might the Kruskal-Wallis test be preferable, given its power advantage over *F*-test under specific distributions (Büning, 1997; Lantz, 2013). However, other studies suggest that *F*-test is robust, in terms of power, to violations of normality under certain conditions (Ferreira, Rocha, & Mequelino, 2012; Kanji, 1976; Schmider et al., 2010), even with very small sample size (n = 3; Khan & Rayner, 2003). In light of these inconsistencies, future research should explore the power of *F*-test when the normality assumption is not met. At all events, we encourage researchers to analyze the distribution underlying their data (e.g., coefficients of skewness and kurtosis in each group, goodness of fit tests, and normality graphs) and to estimate a priori the sample size needed to achieve the desired power.

Table 2
Descriptive statistics of Type I error for F-test with equal shape for each combination of skewness (γ_1) and kurtosis (γ_2) across all conditions

Distributions	γ_1	γ_2	n	Min	Max	Mdn	M	SD
1	0	0.4	=	.0434	.0541	.0491	.0493	.0029
			≠	.0445	.0556	.0497	.0496	.0022
2	0	0.8	=	.0444	.0534	.0474	.0479	.0023
			≠	.0458	.0527	.0484	.0487	.0016
3	0	-0.8	=	.0468	.0512	.0490	.0491	.0014
			≠	.0426	.0532	.0486	.0487	.0024
4	0.4	0	=	.0360	.0499	.0469	.0457	.0044
			≠	.0392	.0534	.0477	.0472	.0032
5	0.8	0	=	.0422	.0528	.0477	.0476	.0029
			≠	.0433	.0553	.0491	.0491	.0030
6	-0.8	0	=	.0427	.0551	.0475	.0484	.0038
			≠	.0457	.0549	.0487	.0492	.0024
7	0.4	0.4	=	.0426	.0533	.0487	.0488	.0031
			≠	.0417	.0533	.0486	.0487	.0026
8	0.4	0.8	=	.0449	.0516	.0483	.0485	.0019
			≠	.0456	.0537	.0489	.0489	.0020
9	0.8	0.4	=	.0372	.0494	.0475	.0463	.0033
			≠	.0413	.0518	.0481	.0475	.0026
10	0.8	1	=	.0458	.0517	.0494	.0492	.0017
			≠	.0463	.0540	.0502	.0501	.0023
11	1	0.8	=	.0398	.0506	.0470	.0463	.0028
			≠	.0430	.0542	.0489	.0485	.0029
12	1	1	=	.0377	.0507	.0453	.0451	.0042
			≠	.0366	.0512	.0466	.0462	.0032
13	0	3	=	.0443	.0517	.0477	.0479	.0022
			≠	.0435	.0543	.0490	.0489	.0024
14	1	3	=	.0431	.0530	.0487	.0486	.0032
			≠	.0462	.0548	.0494	.0499	.0017
15	2	6	=	.0474	.0524	.0496	.0497	.0017
			≠	.0442	.0526	.0483	.0488	.0022

Table 3
Descriptive statistics of Type I error for F-test with unequal shape for each combination of skewness (γ_1) and kurtosis (γ_2) across all conditions

Distributions	Group	γ_1	γ_2	n	Min	Max	Mnd	M	SD
16	1	0	0.2	=	.0434	.0541	.0491	.0493	.0029
	2	0	0.4	≠	.0433	.0540	.0490	.0487	.0025
	3	0	0.6						
17	1	0	0.2	=	.0472	.0543	.0513	.0509	.0024
	2	0	0.4	≠	.0409	.0579	.0509	.0510	.0033
	3	0	-0.6						
18	1	0.2	0	=	.0426	.0685	.0577	.0578	.0077
	2	0.4	0	≠	.0409	.0736	.0563	.0569	.0072
	3	0.6	0						
19	1	0.2	0	=	.0481	.0546	.0501	.0504	.0020
	2	0.4	0	≠	.0449	.0574	.0497	.0499	.0024
	3	-0.6	0						
20	1	0.2	0.4	=	.0474	.0524	.0496	.0497	.0017
	2	0.4	0.6	≠	.0433	.0662	.0535	.0545	.0057
	3	0.6	0.8						
21	1	0.2	0.4	=	.0462	.0537	.0503	.0501	.0024
	2	0.6	0.8	≠	.0419	.0598	.0499	.0502	.0025
	3	1	1.2						
22	1	0	3	=	.0460	.0542	.0490	.0494	.0027
	2	1	3	≠	.0424	.0577	.0503	.0499	.0029
	3	2	6						

As the present study sought to provide a systematic examination of the independent effect of non-normality on *F*-test Type I error rate, variance homogeneity was assumed. However, previous studies have found that *F*-test is sensitive to violations of homogeneity assumptions (Alexander & Govern, 1994; Blanca, Alarcón, Arnau, & Bono, in press; Büning, 1997; Gamage & Weerahandi, 1998; Harwell et al., 1992; Lee & Ahn, 2003; Lix et al., 1996; Moder, 2010; Patrick, 2007; Yiğit & Gökpinar, 2010; Zijlstra, 2004), and several procedures have been proposed for dealing with heteroscedasticity (e.g., Alexander & Govern, 1994; Brown-Forsythe, 1974; Chen & Chen, 1998; Krishnamoorthy, Lu, & Mathew, 2007; Lee & Ahn, 2003; Li, Wang, & Liang, 2011; Lix & Keselman, 1998; Weerahandi, 1995; Welch, 1951). This suggests that heterogeneity has a greater effect on *F*-test robustness than does non-normality. Future research should therefore also consider violations of homogeneity.

To sum up, the present results provide empirical evidence for the robustness of *F*-test under a wide variety of conditions (1308) involving non-normal distributions likely to represent real data. Researchers can use these findings to determine whether *F*-test is a valid option when testing hypotheses about means in their data.

Acknowledgements

This research was supported by grants PSI2012-32662 and PSI2016-78737-P (AEI/FEDER, UE; Spanish Ministry of Economy, Industry, and Competitiveness).

References

Alexander, R. A., & Govern, D. M. (1994). A new and simpler approximation for ANOVA under variance heterogeneity. *Journal of Educational and Behavioral Statistics, 19*, 91-101.

Bendayan, R., Arnau, J., Blanca, M. J., & Bono, R. (2014). Comparison of the procedures of Fleishman and Ramberg et al., for generating non-normal data in simulation studies. *Anales de Psicología*, *30*, 364-371.

Black, G., Ard, D., Smith, J., & Schibik, T. (2010). The impact of the Weibull distribution on the performance of the single-factor ANOVA model. *International Journal of Industrial Engineering Computations, 1*, 185-198.

Blanca, M. J., Alarcón, R., Arnau, J., & Bono, R. (in press). Effect of variance ratio on ANOVA robustness: Might 1.5 be the limit? *Behavior Research Methods*.

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology*, *9*, 78-84.

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

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

Büning, H. (1997). Robust analysis of variance. *Journal of Applied Statistics, 24*, 319-332.

Chen, S.Y., & Chen, H.J. (1998). Single-stage analysis of variance under heteroscedasticity. *Communications in Statistics – Simulation and Computation, 27*, 641-666.

Clinch, J. J., & Kesselman, H. J. (1982). Parametric alternatives to the analysis of variance. *Journal of Educational Statistics, 7*, 207-214.

Feir-Walsh, B. J., & Thoothaker, 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*, 789-799.

Ferreira, E. B., Rocha, M. C., & Mequelino, D. B. (2012). Monte Carlo evaluation of the ANOVA's *F* and Kruskal-Wallis tests under binomial distribution. *Sigmae, 1*, 126-139.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*, 521-532.

Gamage, J., & Weerahandi, S. (1998). Size performance of some tests in one-way ANOVA. *Communications in Statistics - Simulation and Computation, 27*, 625-640.

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*, 237-288.

Golinski, C., & Cribbie, R. A. (2009). The expanding role of quantitative methodologists in advancing psychology. *Canadian Psychology, 50*, 83-90.

Harvey, C., & Siddique, A. (2000). Conditional skewness in asset pricing test. *Journal of Finance, 55*, 1263-1295.

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.

Kanji, G. K. (1976). Effect of non-normality on the power in analysis of variance: A simulation study. *International Journal of Mathematical Education in Science and Technology, 7*, 155-160.

Keppel, G. (1982). *Design and analysis. A researcher's handbook* (2nd ed.). New Jersey: Prentice-Hall.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1999). A comparison of recent approaches to the analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology, 52*, 63-78.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B.,..., Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*, 350-386.

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*, 187-206.

Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in *AERJ* and *JCP* articles from 1988 to 1997: A methodological review. *The Journal of Experimental Education, 69*, 280-309.

Kirk, R. E. (2013). *Experimental design. Procedures for the behavioral sciences* (4th ed.). Thousand Oaks: Sage Publications.

Kobayashi, K. (2005). Analysis of quantitative data obtained from toxicity studies showing non-normal distribution. *The Journal of Toxicological Science, 30*, 127-134.

Krishnamoorthy, K., Lu, F., & Mathew, T. (2007). A parametric bootstrap approach for ANOVA with unequal variances: Fixed and random models. *Computational Statistics & Data Analysis 51*, 5731-5742.

Lantz, B. (2013). The impact of sample non-normality on ANOVA and alternative methods. *British Journal of Mathematical and Statistical Psychology, 66*, 224-244.

Lee, S., & Ahn, C. H. (2003). Modified ANOVA for unequal variances. *Communications in Statistics - Simulation and Computation, 32*, 987-1004.

Li, X., Wang, J., & Liang, H. (2011). Comparison of several means: A fiducial based approach. *Computational Statistics and Data Analysis, 55*, 1993-2002.

Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin.

Lix, L.M., & Keselman, H.J. (1998). To trim or not to trim: Tests of mean equality under heteroscedasticity and nonnormality. *Educational and Psychological Measurement, 58*, 409-429.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance *F* test. *Review of Educational Research*, *66*, 579-619.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah: Lawrence Erlbaum Associates.

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

Moder, K. (2010). Alternatives to *F*-test in one way ANOVA in case of heterogeneity of variances (a simulation study). *Psychological Test and Assessment Modeling, 52*, 343-353.

Montgomery, D. C. (1991). *Design and analysis of experiments* (3rd ed.). New York, NY: John Wiley & Sons, Inc.

Patrick, J. D. (2007). *Simulations to analyze Type I error and power in the ANOVA F test and nonparametric alternatives* (Master's thesis, University of West Florida). Retrieved from http://etd.fcla.edu/WF/WFE0000158/Patrick_Joshua_Daniel_200905_MS.pdf

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika, 23*, 114-133.

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, *45*, 283-288.

SAS Institute Inc. (2013). *SAS® 9.4 guide to software Updates*. Cary: SAS Institute Inc.

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, 6*, 147-151.

Tiku, M. L. (1964). Approximating the general non-normal variance-ratio sampling distributions. *Biometrika, 51*, 83-95.

Van Der Linder, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*, 181-204.

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

Weerahandi, S. (1995). ANOVA under unequal error variances. *Biometrics, 51*, 589-599.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

Yiğit, E., & Gökpınar, F. (2010). A Simulation study on tests for one-way ANOVA under the unequal variance assumption. *Communications Faculty of Sciences University of Ankara, Series A1, 59*, 15-34.

Zijlstra, W. (2004). *Comparing the Student´s t and the ANOVA contrast procedure with five alternative procedures* (Master's thesis, Rijksuniversiteit Groningen). Retrieved from http://www.ppsw.rug.nl/~kiers/ReportZijlstra.pdf