Original Article

Is It Really Robust?

Reinvestigating the Robustness of ANOVA Against Violations of the Normal Distribution Assumption

Emanuel Schmider,[1] Matthias Ziegler,[1] Erik Danay,[1] Luzi Beyer,[1] and Markus Bühner[2]

[1]Humboldt University Berlin, Germany, [2]Karl-Franzens University Graz, Austria

Abstract. Empirical evidence to the robustness of the analysis of variance (ANOVA) concerning violation of the normality assumption is presented by means of Monte Carlo methods. High-quality samples underlying normally, rectangularly, and exponentially distributed basic populations are created by drawing samples which consist of random numbers from respective generators, checking their goodness of fit, and allowing only the best 10% to take part in the investigation. A one-way fixed-effect design with three groups of 25 values each is chosen. Effect-sizes are implemented in the samples and varied over a broad range. Comparing the outcomes of the ANOVA calculations for the different types of distributions, gives reason to regard the ANOVA as robust. Both, the empirical type I error α and the empirical type II error β remain constant under violation. Moreover, regression analysis identifies the factor "type of distribution" as not significant in explanation of the ANOVA results.

Keywords: ANOVA, assumption violation, normal distribution, high-quality samples, Monte Carlo

The object of this paper is to empirically investigate the robustness of the univariate one-way fixed-effects analysis of variance (ANOVA) against deviations from the assumption of a normally distributed dependent variable. This is tested by applying samples from basic populations. The distribution of individual values in each population differs considerably from the normal distribution and was derived with the method of Monte Carlo simulations. In contrast to previous research, the effort to use data with a high quality was extensive.

Previous Research

Due to violated assumptions the appropriateness of the ANOVA is often doubted. Micceri (1989) analyzed 400 published data sets reporting, most did not have univariate normal distributions. After analyzing 17 journals, Keselman et al. (1998) found that researchers rarely verify the conformance of validity assumptions. The harm of violation of assumptions can be understood quite intuitively considering the role of group means and their variances. When data are skewed, means no longer reflect the central location. When variances are unequal, not every group has the same level of noise, and thus comparisons are invalid. More importantly, the inferences made from the sample statistic to the population parameter based on a test statistic might be flawed (Yu, 2002).

If a one-way ANOVA is used to analyze data sets of randomly selected numbers, the frequency distribution of empirical F values will approximate the probability density curve of the F statistic for the specified degrees of freedom only if the assumptions of the test are respected. Violations of the assumptions will impair this approximation (Ware, 2000), possibly leading to an inflation of the type I or II errors.

Reviews by Glass et al. (1972) and Harwell et al. (1992) sum up a lot of evidence for the robustness of the ANOVA with regard to the empirical α and β values. Other evidence discourages from usage of nonparametric tests due to the loss of precision that comes along with transformations into rank data (Edgington, 1995), lower power (Tanizaki, 1997), and inaccuracy in case of multiple violations (Zimmerman, 1998). All in all, the findings speak for the robustness of the ANOVA concerning violations of the normality assumption and the lack of valuable alternatives (see also Keselman, Algina, Lix, Wilcox, & Deering, 2008; Lix, Keselman, & Keselman, 1996). As reassuring as this seems, previous research could be inferior concerning the quality of samples. Firstly, early studies (before around, 1985) worked with random number generators of questionable quality (Park & Miller, 1988). Secondly, it was not tested that samples underlying Monte Carlo simulations are representative of their distribution or not. We will resolve these deficits by application of high-quality samples that incorporate a goodness of fit test between the samples and their underlying distributions.

Method

For our simulations, we will take random numbers from three different distributions: Normal-, rectangular- and exponential distribution. The normal distribution function is given by

DOI: 10.1027/1614-2241/a000016

Table 1. Moments with their meanings and values for the normal distribution (Equation 1), the rectangular distribution (Equation 2), and the exponential distribution (Equation 3)

n'th central moment	Meaning	Normal distribution	Rectangular distribution	Exponential distribution
1	Mean	0	0	0
2	Variance	1	1	1
3	Skewness	0	0	2
4	Kurtosis	3	1.8	9

Annotations. The first two moments are identical due to construction.

$$f_{\text{normal}}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad (1)$$

with the mean μ and the standard deviation σ.

The rectangular distribution function is defined by

$$f_{\text{rect.}}(x) = \begin{cases} 0, & \text{if } x < a \text{ or } x > b \\ \frac{1}{b-a}, & \text{if } a \le x \le b \end{cases}. \quad (2)$$

The exponential distribution function is given by

$$f_{\text{rect.}}(x) = \begin{cases} 0, & \text{if } x < \beta \\ \frac{\alpha}{e^{\alpha\beta}} e^{-\alpha x}, & \text{if } x > \beta \end{cases}, \quad (3)$$

with the decay parameter α. It has no line symmetry, but is skewed to the right (skewness = 2) whereas the kurtosis is larger than the normal distribution's one.

In the following, we will fix the free parameters of each distribution, such that the first two moments correspond to a standard normal distribution: first moment (mean) = 0 and second moment (dispersion) = 1. The moments of the three standardized distributions are displayed in Table 1 and the curves plotted in Figure 1.

Design of the Monte Carlo Simulations

A univariate, one-way experimental design with three groups (n = 25) and fixed-effects was chosen. Empirical studies often use a design with three groups with 25 persons in each group because this size is often recommended as the threshold for robustness. Beside this, differences between groups were also varied using effect-sizes f. The corresponding means were calculated with G*power (Erdfelder et al., 1996) and are given in Table 2.

These effect-sizes cover the range from low to high effects as proposed by Cohen (Rothstein et al., 1990). Thus, if ANOVAs with an effect-size of, for example, f = 0.2 were analyzed, independent samples were simulated by programming the random number generator (Gough, 2003) such that the groups yield the following mean values: 0 for group 1; 0.2450 for group 2; and 0.4900 for group 3. By doing this, the variance within a group remains the same, which is important. The sample size of 75 can be considered optimal in terms of power for an effect-size of f = 0.37. This effect-size was added to check power empirically. Thus, a 3 (Distributions) × 8 (Effect-sizes) design with 24 conditions was chosen. For each condition 50,000 data sets were simulated.

Getting Appropriate Samples

By drawing samples from the respective distributions normally, rectangularly, and exponentially distributed data were obtained. However, a sample is not always a good representative for the basic population it is drawn from. For this reason, the samples taken from the respective distributions were analyzed prior to conducting the actual ANOVAs. Aim of this was to extract those samples that are prototypical for their basic population. To determine prototypicality the goodness of fit was computed with the Kolmogorov-Smirnov test (K-S test). Only the 10% best samples were chosen by applying a linear algorithm, that picks out, by exhaustive comparison, the 5,000 samples with the best fit.

Statistical Analyses

For each of the 24 conditions 5,000 ANOVAs were computed. Outcomes were coded as 0 (H_0 cannot be rejected) in case of no significant differences between the groups or 1 (H_0 has to be rejected) in case of significant differences.

The results from the ANOVAs were then regressed on the effect-size and type of distribution with a logistic regression analysis. The type of distribution is a categorial variable and was therefore recoded into binary parameters.

Results

Goodness of Fit

Figure 2 contains plots of samples drawn from a normally distributed basic population.

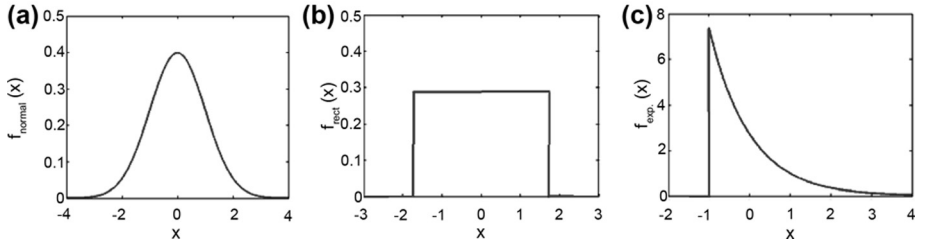

Figure 1. (a) Normal distribution (Equation 1). (b) Rectangular distribution (Equation 2). (c) Exponential distribution (Equation 3). Parameters are fixed concerning Table 1.

Table 2. Effect-sizes *f* and their corresponding mean differences for the three groups

Mean difference	$f = 0$	$f = 0.1$	$f = 0.2$	$f = 0.3$	$f = 0.37$	$f = 0.4$	$f = 0.5$	$f = 0.6$
d_1	0	0	0	0	0	0	0	0
d_2	0	0.1225	0.2450	0.3675	0.4533	0.4900	0.6125	0.7350
d_3	0	0.2450	0.4900	0.7350	0.9065	0.9800	1.2250	1.4700

Annotations. d_1 = mean for group 1, d_2 = mean for group 2, and d_3 = mean for group 3.

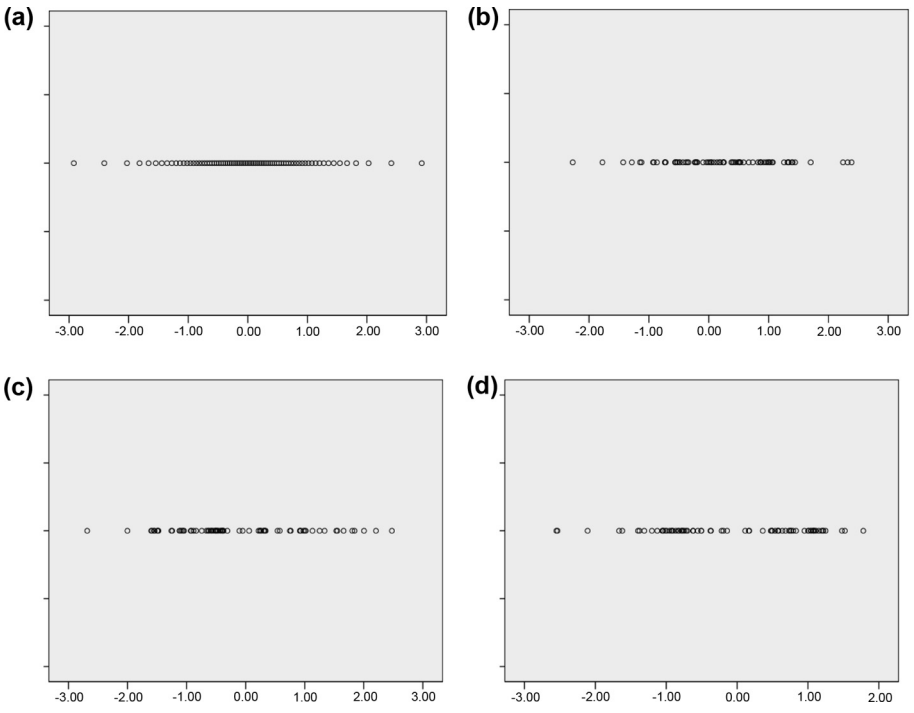

Figure 2. (a) Perfect sample from a normal distribution. (b) Good sample, $\alpha = 1.00$. (c) Bad sample, $\alpha = 0.02$. (d) Bad sample, $\alpha = 0.03$.

A perfect sample of the normal distribution is shown in Figure 2a. Good samples should closely resemble the perfect distribution. This holds for the sample shown in Figure 2b. In Figure 2c however, there is a quite low density at 0 as only three data points range between −0.3 and 0.2. Moreover, the symmetry is broken. Figure 2d, on the other hand, is rather symmetric but concentrating many values at −1 and 1, whereas the density at 0 is much lower. All in all, 5,000 prototypical samples were drawn in each condition.

ANOVAs

Results for the ANOVAs are shown in Table 3.

As can be seen, the number of significant tests increases with increasing effect-size as was expected. However, even in the case of no actual group differences (effect-size = 0), there were some significant outcomes (around 5% independent for all types of distributions). This can be explained by the fact that random numbers sometimes lead to unbalanced mean values between the groups and therefore yield significant differences. The percentage of such false results closely approximates the conventional level for type I errors. On the other hand, in case of an optimal sample size, the type II error β should comply with the number of (wrongly) not significant outcomes. According to Table 3, the percentage of not significant outcomes is around 80% for $f = 0.37$. Thus, the type II error β, chosen as a basis when constructing the design, could exactly be replicated.

Invariance of α and β for different distributions results from comparison by estimating their confidence intervals. According to Glass et al. (1972), the value of the standard error for processing 5,000 ANOVA simulations, assuming $\alpha = 5\%$, is given by about 0.005% or 0.5%. The confidence intervals of the empirical α and β belonging to the three distributions overlap more than 50% with each other. This indicates that the differences are not significant (Cumming & Finch, 2005). Therefore we can state that α and β stay constant under application of non-normal distributions.

Regression Analysis

The results for the regression analysis are given in Table 4. There is a significant influence of effect-size, but no significant impact of type of distribution.

Table 3. Results of the 5,000 ANOVA tests for each condition

Distribution	Effect-size *f*	Significant ANOVAs (Σ)	Significant ANOVAs (%)
Normal	0	257	5.14
Normal	0.1	509	10.18
Normal	0.2	1,550	31.00
Normal	0.3	3,069	61.38
Normal	0.37	4,013	80.26
Normal	0.4	4,342	86.84
Normal	0.5	4,891	97.82
Normal	0.6	4,983	99.66
Rectangular	0	263	5.26
Rectangular	0.1	533	10.66
Rectangular	0.2	1,548	30.96
Rectangular	0.3	3,118	62.36
Rectangular	0.37	4,038	80.76
Rectangular	0.4	4,374	87.48
Rectangular	0.5	4,901	98.02
Rectangular	0.6	4,998	99.96
Exponential	0	236	4.72
Exponential	0.1	562	11.24
Exponential	0.2	1,645	32.90
Exponential	0.3	3,172	63.44
Exponential	0.37	4,027	80.54
Exponential	0.4	4,343	86.86
Exponential	0.5	4,817	96.34
Exponential	0.6	4,971	99.42

Annotations. For each combination of distribution with effect-size 5,000 tests were computed.

Table 4. Results for the logistic regression analysis

	B	*SE*	Wald	*df*	*p*	Exp (B)
Distribution			3.791	2	0.150	
Distribution (1)	−0.036	0.021	2.842	1	0.092	0.965
Distribution (2)	0.000	0.021	0.000	1	1.000	1.000
Effect-size	13.231	0.072	33770.980	1	0.000	557451.892
Constant	−3.383	0.025	18227.812	1	0.000	0.034

Annotations. *B* = regression coefficient; *SE* = standard error.

Nagelkerkes R^2 was .64. Considering that the kind of distribution was not a significant factor, this was primarily achieved by the factor effect-size.

The results have been reproduced by recurrence of all previous steps with new random numbers, leading to the same outcomes.

Discussion

The present study aimed at investigating the robustness of the ANOVA against violations of the underlying assumption of normally distributed data. Unlike previous studies a high-quality random number generator was used to simulate normally, rectangularly, and exponentially distributed data. Beside distribution shape effect-size was also varied. All other influences, such as group variance or group size, were held constant. Thus, results can causally be explained by the manipulations of distribution shape and effect-size. The results give strong support for the robustness of the ANOVA under application of non-normally distributed data.

A lot of effort was dedicated to process the Monte Carlo simulations with data of very high quality. A filtering process that compares the samples with the desired type of distribution and allows only the best 10% of the samples to pass the actual calculations followed the generation of high-quality random numbers. Taking into account that 10 times more random numbers than finally used were simulated, altogether 90 million random numbers were simulated. We did not find investigations, to our best knowledge, that cared that much about high-quality samples as a basis of their Monte Carlo simulations.

Limitations and Outlook

The present study only focused on one assumption of the ANOVA. Violations of other assumptions have been shown to negatively influence ANOVAs. It would be interesting to check further variations of the design with high-quality samples under violations of the normality assumption.

For now the commonly given advice to use samples of 25 participants per condition in ANOVA designs to circumvent possible negative influences of violations of normality assumptions seems well heeded.

References

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*, 170–180.

Edgington, E. S. (1995). *Randomization tests*. New York: M. Dekker.

Erdfelder, E., Faul, F., & Buchner, A. (1996). G*power: A general power analysis program. *Behavior Research Methods, Instruments, & Computers, 28*, 1–11.

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

Gough., B. (2003). GNU Scientific Library Reference Manual, (2nd ed.) Network Theory Ltd..

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.

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods, 13*, 110–129.

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

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.
Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156–166.
Park, S. K., & Miller, K. W. (1988). Random number generators: good ones are hard to find. *Communications of the ACM, 31*, 10.
Rothstein, H. R, Borenstein, M., Cohen, J., & Pollack, S. (1990). Statistical power analysis for multiple regression/correlation: A computer program. *Educational and Psychological Measurement, 50*, 819–830.
Tanizaki, H. (1997). Power comparison of non-parametric tests: Small-sample properties from Monte Carlo experiments. *Journal of Applied Statistics, 24*, 603–632.
Ware, M. E. (2000). *Demonstrations and activities in the teaching of psychology,* Vol. I, Erlbaum.
Yu, C. H. (2002). An overview of remedial tools for the violation of parametric test assumptions in the SAS system. *Proceedings of 2002 Western Users of SAS Software Conference.*
Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education, 67*, 55–68.

Matthias Ziegler

Humboldt University Berlin
Institute for Psychological Diagnostics
Unter den Linden 6
10099 Berlin
Germany
E-mail matthias.ziegler@cms.hu-berlin.de

