GRUPO EDITORIAL DE PSICOFUNDACIÓN

Psicothema (2023) 35(1) 21-29

Psicothema

www.psicothema.com • ISSN 0214-9915

Colegio Oficial de Psicólogos del Principado de Asturias

Psicothema

Article

Non-normal Data in Repeated Measures ANOVA: Impact on Type I Error and Power

María J. Blanca[1], Jaume Arnau[2], F. Javier García-Castro[1], Rafael Alarcón[1] and Roser Bono[2]

1 University of Malaga.
2 University of Barcelona.

ARTICLE INFO

Received: July 06, 2022
Accepted: September 25, 2022

Keywords:
Violation of normality
Within-subject design
Robustness
Power
ANOVA

ABSTRACT

Background: Repeated measures designs are commonly used in health and social sciences research. Although there are other, more advanced, statistical analyses, the F-statistic of repeated measures analysis of variance (RM-ANOVA) remains the most widely used procedure for analyzing differences in means. The impact of the violation of normality has been extensively studied for between-subjects ANOVA, but this is not the case for RM-ANOVA. Therefore, studies that extensively and systematically analyze the robustness of RM-ANOVA under the violation of normality are needed. This paper reports the results of two simulation studies aimed at analyzing the Type I error and power of RM-ANOVA when the normality assumption is violated but sphericity is fulfilled. **Method:** Study 1 considered 20 distributions, both known and unknown, and we manipulated the number of repeated measures (3, 4, 6, and 8) and sample size (from 10 to 300). Study 2 involved unequal distributions in each repeated measure. The distributions analyzed represent slight, moderate, and severe deviation from normality. **Results:** Overall, the results show that the Type I error and power of the F-statistic are not altered by the violation of normality. **Conclusions:** RM-ANOVA is generally robust to non-normality when the sphericity assumption is met.

Datos no Normales en el ANOVA de Medidas Repetidas: Impacto en el Error Tipo I y Potencia

Palabras clave:
Violación de la normalidad
Diseño intrasujeto
Robustez
Potencia
ANOVA

RESUMEN

Antecedentes: El diseño de medidas repetidas es uno de los más usados en ciencias sociales y de la salud. Aunque hay otras alternativas más avanzadas, el análisis de varianza de medidas repetidas (ANOVA-MR) sigue siendo el procedimiento más empleado para analizar las diferencias de medias. El impacto de la violación de la normalidad ha sido muy estudiado en el ANOVA intersujeto, pero los estudios son muy escasos en el ANOVA-MR. Por ello, el objetivo de este trabajo es realizar dos estudios de simulación Monte Carlo para analizar el error de Tipo I y la potencia cuando se incumple este supuesto bajo el cumplimiento de la esfericidad. **Método:** El estudio 1 incluye 20 distribuciones, tanto conocidas como desconocidas, manipulando el número de medidas repetidas (3, 4, 6 y 8) y el tamaño muestral (de 10 a 300). El estudio 2 incluye diferentes distribuciones en cada medida repetida. Las distribuciones analizadas representan desviación leve, moderada y severa de la normalidad. **Resultados:** En general, los resultados muestran que tanto el error Tipo I como la potencia del estadístico F no se alteran con la violación de la normalidad. **Conclusiones:** El ANOVA-MR es generalmente robusto a la no normalidad cuando la esfericidad se satisface.

Cite as: Blanca, M. J., Arnau, J., García-Castro, F. J., Alarcón, R., & Bono, R. (2023). Non-normal Data in Repeated Measures ANOVA: Impact on Type I Error and Power. *Psicothema, 35*(1), 21-29. https://doi.org/10.7334/psicothema2022.292
Corresponding author: María J. Blanca, blamen@uma.es

Repeated measures designs are widely used in health and social sciences research (Fernández et al., 2010), not only in psychology but also in fields such as general medicine (Singh et al., 2013), psychiatry (Gueorguieva & Krystal, 2004), epidemiology (Gunasekara et al., 2014), pharmacology (Maurissen & Vidmar, 2017), neurotoxicology (Tamura & Buelke-Sam, 1992), anesthesiology (Schober & Vetter, 2018), ophthalmology (Armstrong, 2017), pulmonology (De Livera et al., 2014), and veterinary science (Zhao et al., 2019). In the methodological literature, the analysis of repeated measures data continues to generate debate, as illustrated by the considerable number of books (e.g., Davis, 2002; Islam & Chowdhury, 2017; Moskowitz & Hershberger, 2013; Raghavarao & Padgett, 2014; Verma, 2016), tutorials, and review articles that have been published since 2000 (e.g., Armstrong, 2017; Bathke et al., 2009; Blanca, 2004; De Livera et al., 2014; Fernández et al., 2007; Gueorguieva & Krystal, 2004; Keselman et al., 2001, 2002; Maurissen & Vidmar, 2017; Schober & Vetter, 2018; Singh et al., 2013; Tippey et al., 2015; Vallejo & Lozano, 2006). The conventional univariate test of significance within the general linear model for the analysis of repeated measures is repeated measures analysis of variance (RM-ANOVA), which uses the F-statistic to determine statistical significance. The model is defined by:

$$Y_{ij} = \mu + \alpha_j + \pi_i + \varepsilon_{ij}$$

where Y_{ij} represents the observation for subject i at time j; μ is the grand mean of the population means; α_j is the fixed effect of time j; π_i represents the random effect for subject i; and ε_{ij} is the error effect associated with subject i at time j. This error effect is a random variable, defined as NID(0, σ_ε^2), and it is independent of π_i. The RM-ANOVA procedure requires fulfillment of the assumptions of normality and sphericity, among others. Although other approaches (e.g., mixed model, multivariate analysis, adjusted F test, etc.) have been proposed for the analysis of repeated measures data when these assumptions are not met, RM-ANOVA remains one of the most widely used statistical procedures in various areas of knowledge (Armstrong, 2017; Blanca et al., 2018; Goedert et al., 2013).

Monte Carlo simulation studies aim to analyze how the violation of assumptions affects the robustness of statistical procedures. Type I error is defined as the probability of rejecting the null hypothesis when it is true. This probability is called the significance level or α, with a conventionally preset value of .05. In the context of ANOVA, obtaining inflated Type I error rates leads to the conclusion that there is a treatment effect, or differences in means, when this is not the case. The probability of erroneously accepting the null hypothesis is referred to as Type II error, labeled β. Power is defined as the probability of correctly rejecting the null hypothesis (1 - β), i.e., the probability of detecting an effect when it actually exists. Conventionally, a power value of .80 is considered adequate (Cooper & Garson, 2016; Kirk, 2013). Power depends on factors such as significance level, sample size, and effect size (Cohen, 1988).

A robust statistical procedure is one that is resistant to deviations from its underlying assumptions (Box, 1953). In terms of Type I error, a procedure is robust when the actual probability of Type I error is close to the nominal significance level of .05. The violation of an assumption does not automatically imply that a test is invalidated, but it is essential to be aware of the consequences of a violation so as to understand the potential mistakes that could occur in the statistical decision-making process. Although the impact of the violation of normality has been extensively studied for between-subjects ANOVA (e.g., Blanca et al., 2017; Schmider et al., 2010), this is not the case for RM-ANOVA; most studies of the latter are focused on analyzing the impact of the violation of sphericity or of both sphericity and normality simultaneously (e.g., Berkovits et al., 2000; Haverkamp & Beauducel, 2017, 2019).

Some methodological books suggest that non-normality may increase the Type I error and decrease the power of RM-ANOVA (Verma, 2016), with some authors proposing the transformation of the dependent variable or the use of a non-parametric procedure as analytic alternatives (Tabachnick & Fidell, 2007). In this context, Sheskin (2003) states that if one or more of the assumptions of a parametric test are violated, data may be transformed into a format that makes it compatible for analysis with the appropriate non-parametric test. Similarly, Wilcox (2022) has argued that the F-statistic has undesirable properties under non-normality, especially in situations with outliers and heavy-tailed distributions, and he proposes robust statistical procedures to address this problem. By contrast, a meta-analysis by Keselman et al. (1996) suggests that RM-ANOVA is generally insensitive to non-normality, although Type I error may increase slightly when the shape of the distribution is asymmetric. More recent studies also show that RM-ANOVA tends to be robust to the violation of normality (Berkovits et al., 2000; Kherad-Pajouh & Renaud, 2015), although these studies were aimed at comparing the performance of other statistical procedures with that of the F-statistic, especially in small samples.

Regarding power, most studies likewise focus mainly on comparing different statistical procedures and do not analyze whether there is a loss of power when RM-ANOVA is used with non-normal as opposed to normal distributions. For example, Bosley (2019) compared the performance of RM-ANOVA with that of three non-parametric and two robust procedures, testing three and five repeated measures and different distributions (normal, uniform, chi-square with 2 degrees of freedom, and Student's t with 3 degrees of freedom). Overall, the results showed higher power for RM-ANOVA. Conversely, Meltzer (2001) compared six statistical procedures and concluded that in terms of Type I error and power there were more effective analyses than RM-ANOVA, one of which was the linear mixed model.

Although the Type I error and power of RM-ANOVA have been previously addressed, there are, to the best of our knowledge, no studies that extensively and independently analyze the effect of non-normality. Consequently, there are no clear guidelines that can inform applied researchers in the statistical analysis of repeated measures data when normality is violated. Our aim in this paper was therefore to analyze the Type I error and statistical power of RM-ANOVA in a wide variety of conditions that may be found in real research situations. To this end, two studies were carried out. In the first, we focus on designs involving 3, 4, 6, and 8 repeated measures and consider different sample sizes representing small, medium, and large samples, with several distributions of the data, including both known and unknown distributions implying slight, moderate, and severe deviation from normality. The second study considers the case of designs involving 3 and 4 repeated

measures with unequal distributions in each repeated measure, a condition that has not been studied previously for RM-ANOVA, although it has been addressed in relation to between-subjects ANOVA (e.g., Blanca et al., 2017). In both studies, data were generated with an unstructured (UN) covariance matrix with sphericity approximately equal to 1 ($\varepsilon \approx .95$) in order to analyze independently the effect of non-normality. The UN matrix is the most general structure (Kowalchuk et al., 2004) and the one most typically found in longitudinal behavioral data (Arnau et al., 2014; Bono et al., 2010).

Empirical Type I error rate and statistical power are analyzed in both studies. The former was interpreted according to Bradley's (1978) criterion, a widely accepted standard that facilitates the comparison of results across similar studies (Arnau et al., 2012; Livacic et al., 2010; Vallejo et al., 2010). According to this criterion, a procedure is considered robust if the Type I error rate is between .025 and .075 for a significance level of .05. This simplifies the interpretation of results and allows us to identify those procedures that are liberal, conservative, and robust to violations. For power, the values of means were set so as to yield a power of approximately .80 for the normal distribution for each sample size, with this value being used as a reference to compare the power obtained with each non-normal distribution.

Study 1. Equal Distributions in the Repeated Measures

The aim here was to analyze empirical Type I error rates and power of the F-statistic in one-way RM-ANOVA with non-normal distributions and equal distributions in the repeated measures.

Method

Instruments

A Monte Carlo simulation study was performed using the SAS/IML (interactive matrix language) software and the PROC GLM module (SAS 9.4; SAS Institute Inc., 2013). A series of macros was created that allowed generation of the data and estimation of the general linear model. These macros are available upon request from the corresponding author. First, we generated an UN covariance matrix with sphericity approximately equal to 1 ($\varepsilon \approx .95$). We used this matrix because, as already noted, it is the most general structure (Kowalchuk et al., 2004) and the one most typically found in longitudinal behavioral data (Arnau et al., 2014; Bono et al., 2010). Next, non-normal data were generated using the procedure proposed by Fleishman (1978), which uses a polynomial transformation to simulate data with specific values of skewness and kurtosis. Normal data were generated using the Cholesky transformation of the covariance matrix. Finally, simulated data were analyzed with the PROC GLM of SAS to obtain probability values associated with the F-statistic of RM-ANOVA.

Procedure

In order to examine the Type I error rate, differences between repeated measures were set to zero. A one-way repeated measures design was considered (no between-subject factor was included), manipulating the following variables:

1. Within-subject levels (K). The repeated measures were K = 3, 4, 6, and 8.
2. Total sample size. Keselman et al. (1998) found that more than half (55.3%) of the studies with repeated measures reported a sample size of 60 or fewer, although the range varied from 6 to 1000. Accordingly, we considered a wide range of sample sizes so as to study small (lower than 30), medium (from 30 to 75), and large samples (above 75) (Bono et al., 2016): 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100, 120, 150, 180, 210, 240, 270, and 300.
3. Shape of the distribution with equal distributions in the repeated measures. The values of skewness (γ_1) and kurtosis (γ_2) for each distribution are shown in Table 1. A total of 20 distributions were investigated, including the normal distribution (distribution 0). Blanca et al. (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. In light of these findings, we considered the 12 distributions (distributions 1-12) used by Blanca et al. (2017), with values of skewness and kurtosis within this interval, representing slight and moderate departure from the normal distribution. Seven well-known distributions (distributions 13-19) were also added so as to consider extreme departures from normality, and they are also representative of real data (Bono et al., 2017; Micceri, 1989). The latter distributions were as follows: a distribution with values of γ_1 and γ_2 corresponding to the double exponential; chi-square with 8 degrees of freedom; exponential; lognormal ($\zeta = 1$ and $\sigma = 0.5$); and three gamma distributions with different values of the shape parameter α (0.75, 2, and 4).

In order to analyze empirical power, the values of means were selected to give *a priori* a target power value of approximately .80. This power was then used as a reference to compare the empirical power of RM-ANOVA for each non-normal distribution. Empirical power was calculated with the syntax *power = 1 - probf(fcrit, numdf, dendf, ncp)*, where *probf* is the probability function of SAS for the *F* distribution, *fcrit* represents the theoretical F-statistic, *numdf* and *dendf* are the degrees of freedom of the numerator and denominator, and *ncp* defines the non-centrality parameter. The following variables were manipulated:

1. Within-subject levels. The repeated measures were K = 3, 4, 6, and 8.
2. Sample size. The sample sizes were set to 10, 20, 50, 100, 200, and 300.
3. Shape of the distribution with equal distributions in the measures repeated. The same 20 distributions considered for empirical Type I error rates were investigated.
4. Mean pattern. Three mean patterns were included for each K. With K = 3, one of the means was different from the means of the other repeated measures (e.g., 1, 1, 2; 1, 2, 1). With K = 4, 6, and 8, the means were manipulated so that a) one was different from the rest (e.g., 1, 1, 1, 2), and b) half were different and equal to each other (e.g., 1, 1, 2, 2). For all K, the means were also manipulated so that the increase between them was linear and proportional (e.g., 1, 1.5, 2, 2.5).

Ten thousand replications of the 1520 and 1440 conditions for Type I error and power, respectively, 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 et al., 2014; Robey & Barcikowski, 1992).

Table 1.
Skewness (γ_1) and kurtosis (γ_2) coefficients for each simulated distribution.

Distributions	γ1	γ2
0 (Normal)	0	0
1	0	0.4
2	0	0.8
3	0	-0.8
4	0.4	0
5	0.8	0
6	-0.8	0
7	0.4	0.4
8	0.4	0.8
9	0.8	0.4
10	0.8	1
11	1	0.8
12	1	1
13	0	3
14	1	3
15	2	6
16	1.75	5.9
17	2.31	8
18	1.41	3
19	1	1.5

Data Analysis

The proportion of rejection of the null hypothesis represented the empirical Type I error rates associated with the F-statistic of RM-ANOVA. As noted earlier, Bradley's (1978) criterion of robustness was used to interpret the results, according to which a procedure is considered robust if the Type I error rate is between .025 and .075 for a nominal alpha level of .05. When the empirical Type I error rate is above the upper limit, the test is considered liberal, and when it is below the lower limit it is considered conservative.

For the power analysis, empirical power for each experimental condition was recorded. Discrepancy was calculated, defined as the difference between the power obtained with the non-normal distribution and that obtained with the normal distribution in each experimental condition.

Results

Table 2 shows descriptive statistics for empirical Type I error rates for each distribution across all the conditions manipulated. The results indicate that Type I error rates were almost always within the interval [.025, .075], with means around .05 in all conditions (shape of distributions, sample size, and number of repeated measures). Only in one case, corresponding to distribution 17, K = 4 with N = 10, was the Type I error rate greater than .075, specifically .078. More detailed results are available upon request from the corresponding author.

Table 3 shows descriptive statistics for the empirical power and discrepancy. Overall, all minimum values of empirical power were around .80, and means of discrepancy were near 0 in all conditions studied.

Table 2.
Minimum and maximum values, median, mean, and standard deviation of the empirical Type I error rate for each distribution across all conditions (K = 3, 4, 6, and 8; N ranged from 10 to 300).

Distributions	Min	Max	Md	M	SD
0 (Normal)	.045	.059	.053	.053	.003
1	.046	.060	.053	.053	.003
2	.045	.064	.053	.053	.004
3	.047	.060	.054	.053	.003
4	.045	.061	.053	.053	.003
5	.045	.062	.053	.053	.003
6	.046	.061	.053	.053	.003
7	.044	.061	.053	.053	.004
8	.047	.059	.053	.053	.003
9	.046	.059	.052	.053	.003
10	.046	.060	.053	.053	.003
11	.047	.066	.053	.053	.003
12	.047	.064	.054	.054	.003
13	.045	.060	.051	.052	.003
14	.045	.060	.052	.052	.003
15	.047	.069	.054	.055	.005
16	.046	.064	.054	.054	.004
17	.045	.078	.054	.056	.007
18	.048	.067	.054	.054	.004
19	.044	.059	.053	.057	.003

Table 3.
Minimum and maximum values, mean, and standard deviation of empirical power and discrepancy for each distribution across all conditions (K = 3, 4, 6, and 8; N = 10, 20, 50, 100, 200, and 300; and different mean patterns). (Discrepancy = power of the respective non-normal distribution – power of the normal distribution).

	Empirical power				Discrepancy			
Distributions	Min	Max	M	SD	Min	Max	M	SD
0 (Normal)	.801	.842	.811	.009	-	-	-	-
1	.799	.847	.811	.010	-.009	.011	.000	.004
2	.796	.844	.811	.010	-.012	.013	.000	.004
3	.796	.846	.810	.010	-.010	.008	-.001	.004
4	.798	.845	.812	.010	-.008	.010	.001	.004
5	.796	.845	.814	.011	-.008	.020	.003	.006
6	.790	.848	.809	.011	-.021	.006	-.002	.005
7	.797	.848	.812	.010	-.011	.013	.001	.004
8	.795	.845	.813	.010	-.007	.013	.002	.004
9	.796	.850	.814	.011	-.010	.020	.002	.006
10	.799	.844	.814	.011	-.009	.024	.003	.006
11	.797	.844	.815	.011	-.006	.030	.004	.007
12	.798	.849	.816	.012	-.009	.026	.004	.007
13	.799	.845	.813	.010	-.007	.013	.002	.004
14	.799	.850	.817	.011	-.006	.027	.006	.007
15	.799	.867	.822	.015	-.006	.055	.011	.014
16	.798	.861	.821	.014	-.008	.047	.010	.012
17	.795	.872	.824	.018	-.006	.064	.013	.016
18	.798	.854	.818	.013	-.007	.039	.007	.010
19	.801	.849	.816	.012	-.009	.027	.005	.007

Study 2. Unequal Distributions in each Repeated Measure

The aim here was to analyze empirical Type I error rates and power of the *F*-statistic in RM-ANOVA with non-normal distributions and unequal distributions in each repeated measure.

Method

Instruments

A Monte Carlo simulation study was performed using the same program and data generation procedure as in Study 1.

Procedure

A one-way repeated measures design was considered (no between-subject factor was included). The following variables were manipulated for empirical Type I error rates:

1. Within-subject levels. The repeated measures were K = 3 and 4.
2. Sample size. The sample sizes were the same as in Study 1: 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100, 120, 150, 180, 210, 240, 270, and 300.
3. Shape of the distribution with unequal distributions in the repeated measures. Seven distributions were considered for each K. The values of γ_1 and γ_2 for each repeated measure are shown in Table 4. Distributions 20-25 and 27-32 correspond to slight and moderate departures from normality, whereas distributions 26 and 33 reflect severe departure. For K = 3 and severe departure, we used the well-known distributions corresponding to the double exponential, chi-square with 8 degrees of freedom, and exponential. For K = 4 and severe departure, we added the gamma distribution ($\alpha = 0.75$) at the last repeated measure.

With respect to empirical power, the manipulated variables were the same as in Study 1 in terms of sample size (6 conditions) and patterns of means (3 conditions) for K = 3 and 4. The shapes of the distribution were the same as for Type I error with unequal distributions in the repeated measures (7 conditions for each K).

Ten thousand replications of the 266 and 252 conditions for Type I error and power, respectively, resulting from the combination of the above variables were performed at a significance level of .05.

Data Analysis

Empirical Type I error rates and power were recorded and analyzed as in Study 1.

Results

Table 5 shows descriptive statistics for empirical Type I error rates for each distribution across all sample sizes. Overall, the results indicated that Type I error rates were within the interval [.025, .075], with means around .05 in all conditions. More detailed results are available upon request from the corresponding author.

Table 4.
Values of skewness (γ_1) and kurtosis (γ_2) for distributions of each repeated measure.

Distributions	Repeated measures	γ_1	γ_2
20	1	0	0.2
	2	0	0.4
	3	0	0.6
21	1	0	0.2
	2	0	0.4
	3	0	-0.6
22	1	0.2	0
	2	0.4	0
	3	0.6	0
23	1	0.2	0
	2	0.4	0
	3	-0.6	0
24	1	0.2	0.4
	2	0.4	0.6
	3	0.6	0.8
25	1	0.2	0.4
	2	0.6	0.8
	3	1	1.2
26	1	0	3
	2	1	3
	3	2	6
27	1	0	0.2
	2	0	0.4
	3	0	0.6
	4	0	0.8
28	1	0	0.2
	2	0	0.4
	3	0	-0.6
	4	0	-0.8
29	1	0.2	0
	2	0.4	0
	3	0.6	0
	4	0.8	0
30	1	0.2	0
	2	0.4	0
	3	-0.6	0
	4	-0.8	0
31	1	0.2	0.4
	2	0.4	0.6
	3	0.6	0.8
	4	0.8	1
32	1	0.2	0.4
	2	0.6	0.8
	3	1	1.2
	4	1.2	1.4
33	1	0	3
	2	1	3
	3	2	6
	4	2.31	8

Table 6 shows the empirical power and discrepancy with respect to the power of the normal distribution for 3 and 4 repeated measures across all sample sizes and mean patterns. Overall, as in Study 1, all minimum values of empirical power were around .80 and means of discrepancy were near 0.

Table 5.
Minimum and maximum values, median, mean, and standard deviation of Type I error rates for 3 and 4 repeated measures as a function of distribution across all conditions of N (which ranged from 10 to 300).

K	Distributions	Min	Max	Md	M	SD
3	20	.047	.053	.050	.050	.002
	21	.047	.055	.051	.051	.002
	22	.045	.053	.052	.051	.002
	23	.048	.053	.050	.050	.002
	24	.046	.053	.050	.049	.002
	25	.047	.055	.050	.050	.002
	26	.046	.053	.049	.049	.002
4	27	.045	.055	.050	.050	.003
	28	.045	.055	.050	.050	.002
	29	.046	.053	.051	.051	.002
	30	.047	.053	.050	.050	.002
	31	.046	.055	.051	.051	.002
	32	.045	.056	.049	.050	.003
	33	.045	.057	.050	.050	.003

Table 6.
Minimum and maximum values, mean, and standard deviation of empirical power and discrepancy for each distribution across all conditions (K = 3 and 4; N = 10, 20, 50, 100, 200, and 300; and different mean patterns).

		Empirical power				Discrepancy			
K	Distributions	Min	Max	M	SD	Min	Max	M	SD
3	20	.799	.830	.814	.007	-.005	.008	.001	.004
	21	.797	.826	.813	.008	-.007	.008	.000	.004
	22	.805	.826	.815	.006	-.005	.006	.002	.004
	23	.796	.830	.813	.008	-.008	.014	.000	.006
	24	.803	.828	.814	.007	-.004	.007	.001	.003
	25	.805	.828	.816	.006	-.004	.009	.003	.004
	26	.808	.831	.821	.007	.001	.020	.008	.007
4	27	.801	.831	.814	.009	-.005	.006	.002	.003
	28	.801	.829	.813	.009	-.010	.006	.001	.004
	29	.799	.830	.815	.009	-.006	.009	.003	.004
	30	.795	.827	.812	.009	-.012	.011	.000	.006
	31	.799	.832	.815	.009	-.005	.011	.003	.004
	32	.804	.834	.816	.009	-.006	.011	.003	.005
	33	.804	.835	.820	.008	-.002	.022	.008	.007

Note: Discrepancy = power obtained in the respective non-normal distribution – power obtained with the normal distribution.

Discussion

The aim of this paper was to analyze the Type I error and statistical power of RM-ANOVA in a wide variety of conditions that may be encountered in real research situations. To this end, two studies were carried out. In the first, we focused on designs with 3, 4, 6, and 8 repeated measures and considered different sample sizes representing small, medium, and large samples with different distribution shapes, including both known and unknown distributions reflecting slight, moderate, and severe deviation from the normal distribution. The second study considered the case of designs involving 3 and 4 repeated measures with unequal distributions in each repeated measure. In both studies we analyzed empirical Type I error and power. The former was interpreted using Bradley's (1978) criterion, while for the latter we compared the power obtained with each non-normal distribution with that obtained with the normal distribution. The value of means was set so as to yield a power of approximately .80 for the normal distribution for each sample size.

Regarding Type I error, the results of Study 1 with equal distribution in the repeated measures indicated, overall, that Type I error rates are within the bounds for considering a statistical procedure as robust according to Bradley's (1978) criterion. Only one Type I error rate was greater than .075, specifically .078, and this corresponded to a design with four repeated measures, a gamma distribution with $\alpha = 0.75$, and $\gamma 1 = 2.31$, $\gamma 2 = 8$ with $N = 10$, that is to say, with severe departure from normality and a very small sample size. The results of Study 2, with unequal distribution in the repeated measures, supported the robustness of RM-ANOVA under non-normality; all Type I error rates were within the interval [.025, .075] and means were around .05 in all conditions.

When interpreting these results it is important to consider the large number of conditions that have been simulated. The two studies included 33 types of distribution (with equal and unequal distributions in the repeated measures), sample sizes between 10 and 300, and designs involving 3, 4, 6, and 8 repeated measures. Across the two studies and a total of 1786 simulated conditions, the Type I error rate was only greater than .075 in one case. In other words, RM-ANOVA is liberal at a rate of 0.05%, whereas it is robust in 99.95% of the conditions studied here. More specifically, the procedure may be considered robust under non-normality with distributions with skewness and kurtosis as large as 2.31 and 8, respectively. These results extend knowledge about the robustness of this parametric procedure to a larger number of conditions than have been considered in previous studies (Berkovits et al., 2000; Kherad-Pajouh & Renaud, 2015).

Regarding the power of RM-ANOVA, the results show that this does not decrease with the violations of normality considered in the present study. Empirical power was around .80, and the discrepancy between the power obtained with each non-normal distribution and that obtained with normal distribution was near 0. This finding held for all conditions, with equal and unequal distributions in the repeated measures, different sample sizes, and different mean patterns, including a linear pattern.

Considering Type I error and power together, we can conclude that departure from normality, at least in the conditions studied here, does not affect the F-statistic when sphericity is fulfilled. This conclusion is in line with Keselman et al. (1996), who suggested, based on the results of a meta-analysis, that the procedure is generally insensitive to non-normality. In contrast to their study, however, we did not detect an increase in Type I error with asymmetric distributions.

The present findings are useful for applied research insofar as they show that RM-ANOVA is a valid statistical procedure under non-normality in a variety of conditions, provided that the sphericity

assumption is met. Therefore, and in contrast to what is recommended in some texts (Tabachnick & Fidell, 2007), transformation of the dependent variable or the use of non-parametric procedures may not be necessary even in the absence of normality. As Blanca et al. (2017) pointed out, these procedures entail a loss of information and pose problems in the interpretation of the results obtained. Our results notwithstanding, researchers are still encouraged to analyze the distribution underlying their repeated measures data and to assess the assumption of sphericity, which is more relevant in the case of RM-ANOVA (Davis, 2002; Kirk, 2013).

This study has a number of limitations that need to be acknowledged. First, Bradley's criterion was used for the interpretation of results. Although this is the established criterion for the interpretation of robustness in the majority of simulation studies, it is not widely known among applied researchers. In this respect, it is important to clarify the implications of this criterion for research: given a nominal significance level of .05, the actual value of Type I error may be different from this value but with a maximum deviation that is considered acceptable (i.e., not exceeding .075 and not dropping below .025). Second, we used a covariance matrix with an approximate sphericity of 1 that may not represent some real research situations. However, we did aim to analyze the effect of non-normality extensively and independently of the effect of violation of sphericity. Future studies are warranted to address the impact of deviations from sphericity and normality by also considering different covariance matrix structures. Third, we have not considered the presence of missing values that may be frequent in data with repeated measures (Davis, 2002; Graham, 2009; Keselman et al., 2001; Vallejo et al., 2011). The general linear model eliminates non-complete cases from the analysis, so it would be interesting in future studies to analyze the behavior of different imputation procedures for these missing values. Finally, the results are limited to distributions with skewness and kurtosis as large as 2.31 and 8, respectively, and more extreme departures have not been analyzed. Researchers may also consult Wilcox (2022) for alternative procedures to RM-ANOVA based on robust methods for dealing with non-normal distributions, such as comparison of means based on trimmed means and bootstrap methods.

Funding

This research was supported by grant PID2020-113191GB-I00 from the MCIN/AEI/ 10.13039/501100011033.

References

Armstrong, R. (2017). Recommendations for analysis of repeated-measures designs: Testing and correcting for sphericity and use of MANOVA and mixed model analysis. *Ophthalmic & Physiological Optics, 37*(5), 585–593. https://doi.org/1.1111/opo.12399.

Arnau, J., Bendayan, R., Blanca, M. J., & Bono, R. (2014). Should we rely on the Kenward–Roger approximation when using linear mixed models if the groups have different distributions? *British Journal of Mathematical and Statistical Psychology, 67*(3), 408–429. https://doi.org/10.1111/bmsp.12026

Arnau, J., Bono, R., Blanca, M. J., & Bendayan, R. (2012). Using the linear mixed model to analyze nonnormal data distributions in longitudinal designs. *Behavior Research Methods, 44*, 1224–1238. https://doi.org/10.3758/s13428-012-0196-y

Bathke, A., Schabenberger, O., Tobias, R., & Madden, L. (2009). Greenhouse-Geisser adjustment and the ANOVA-type statistic: Cousins or twins? *The American Statistician, 63*(3), 239–246. https://doi.org/1.1198/tast.2009.08187

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*(1), 364–371. https://doi.org/10.6018/analesps.30.1.135911

Berkovits, I., Hancock, G., & Nevitt, J. (2000). Bootstrap resampling approaches for repeated measure designs: Relative robustness to sphericity and normality violations. *Educational and Psychological Measurement, 60*(6), 877–892. https://doi.org/1.1177/00131640021970961

Blanca, M. J. (2004). Alternativas de análisis estadístico en los diseños de medidas repetidas [Approaches to the statistical analysis of repeated measures designs]. *Psicothema, 16*(3), 509–518.

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema, 29*(4), 552–557. https://doi.org/1.7334/psicothema2016.383

Blanca, M. J., Alarcón, R., & Bono, R. (2018). Current practices in data analysis procedures in psychology: What has changed? *Frontiers in Psychology, 9*, Article 2558. https://doi.org/10.3389/fpsyg.2018.02558

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 9*(2), 78–84. https://doi.org/10.1027/1614-2241/a000057

Bono, R., Arnau, J., Blanca, M. J., & Alarcón, R. (2016). Sphericity estimation bias for repeated measures designs in simulation studies. *Behavior Research Methods, 48*(4), 1621–1630. https://doi.org/10.3758/s13428-015-0673-1

Bono, R., Arnau, J., & Vallejo, G. (2010). Modelización de diseños split-plot y estructuras de covarianza no estacionarias: un estudio de simulación [Modeling split-plot data and nonstationary covariance structures: A simulation study]. *Escritos de Psicología - Psychological Writings, 3*(3), 1–7. https://doi.org/10.5231/Psy.Writ.2010.2903

Bono, R., Blanca, M. J., Arnau, J., & Gómez-Benito, J. (2017). Non-normal distributions commonly used in health, education, and social sciences: A systematic review. *Frontiers in Psychology, 8*, Article 1602. https://doi.org/10.3389/fpsyg.2017.01602

Bosley, T. (2019). *Comparative power of the Friedman, Neave and Worthington match, Skillings-Mack, trimmed means repeated measures ANOVA, and bootstrap trimmed means repeated measures ANOVA tests* [Doctoral dissertation, Wayne State University]. https://digitalcommons.wayne.edu/oa_dissertations/2318/

Box, G. E. P. (1953). Non-normality on test on variance. *Biometrika, 40*(3–4), 318–335. https://doi.org/10.1093/biomet/40.3-4.318

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

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

Cooper, J. A., & Garson, G. D. (2016). *Power analysis.* Statistical Associates Blue Book Series.

Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*. Springer.

De Livera, A., Zaloumis, S., & Simpson, J. (2014). Models for the analysis of repeated continuous outcome measures in clinical trials: Analysis of repeated continuous measures. *Respirology, 19*(2), 155–161. https://doi.org/1.1111/resp.12217

Fernández, P., Livacic-Rojas, P., & Vallejo, G. (2007). Cómo elegir la mejor prueba estadística para analizar un diseño de medidas repetidas [How to choose the best statistical analysis for analyzing a repeated measures design]. *International Journal of Clinical Psychology, 7*(1), 153–175.

Fernández, P., Vallejo, G., Livacic-Rojas, P. E., & Tuero, E. (2010). *Características y análisis de los diseños de medidas repetidas en la investigación en España en los últimos 10 años* [Characteristics and analysis of repeated measures designs used in research in Spain over the last 10 years]. In M. J. Blanca et al. (coords.), *Actas del XI Congreso de Metodologías de las Ciencias Sociales y de la Salud* (pp. 193–198). Universidad de Málaga.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*(4), 521–532. https://doi.org/1.1007/BF02293811

Goedert, K., Boston, R., & Barrett, A. (2013). Advancing the science of spatial neglect rehabilitation: An improved statistical approach with mixed linear modeling. *Frontiers in Human Neuroscience, 7*, Article 211. https://doi.org/1.3389/fnhum.2013.00211

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Gueorguieva, R., & Krystal, J. (2004). Move over ANOVA: Progress in analyzing repeated-measures data and its reflection in papers published in the Archives of General Psychiatry. *Archives of General Psychiatry, 61*(3), 310–317. https://doi.org/10.1001/archpsyc.61.3.310

Gunasekara, F., Richardson, K., Carter, K., & Blakely, T. (2014). Fixed effects analysis of repeated measures data. *International Journal of Epidemiology, 43*(1), 264–269. https://doi.org/1.1093/ije/dyt221

Haverkamp, N., & Beauducel, A. (2017). Violation of the sphericity assumption and its effect on Type-I error rates in repeated measures ANOVA and multi-level linear models (MLM). *Frontiers in Psychology, 8*, Article 1841. https://doi.org/10.3389/fpsyg.2017.01841

Haverkamp, N., & Beauducel, A. (2019). Differences of Type I error rates for ANOVA and Multilevel-Linear-Models using SAS and SPSS for repeated measures designs. *Meta-Psychology, 3*, Article MP.2018.898. https://doi.org/10.15626/MP.2018.898

Islam, M., & Chowdhury, R. (2017). *Analysis of repeated measures data*. Springer. https://doi.org/1.1007/978-981-10-3794-8

Keselman, H. J., Algina, J., & Kowalchuk, R. (2001). The analysis of repeated measures designs: A review. *British Journal of Mathematical & Statistical Psychology, 54*(1), 1–2. https://doi.org/1.1348/000711001159357

Keselman, H. J., Algina, J., & Kowalchuk, R. K. (2002). A comparison of data analysis strategies for testing omnibus effects in higher-order repeated measures designs. *Multivariate Behavioral Research, 37*(3), 331–357. https://doi.org/10.1207/S15327906MBR3703_2

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*(3), 350–386. https://doi.org/10.3102/00346543068003350

Keselman, J. C., Lix, L. M., & Keselman, H. J. (1996). The analysis of repeated measurements: A quantitative research synthesis. *British Journal of Mathematical and Statistical Psychology, 49*(2), 275–298. https://doi.org/10.1111/j.2044-8317.1996.tb01089.x

Kherad-Pajouh, S., & Renaud, O. (2015). A general permutation approach for analyzing repeated measures ANOVA and mixed-model designs. *Statistical Papers, 56*(4), 947–967. https://doi.org/1.1007/s00362-014-0617-3

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

Kowalchuk, R. K., Keselman, H. J., Algina, J., & Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted F tests. *Educational and Psychological Measurement, 64*(2), 224–242. https://doi.org/10.1177/0013164403260196

Livacic-Rojas, P., Vallejo, G., & Fernández, P. (2010). Analysis of Type I error rates of univariate and multivariate procedures in repeated measures designs. *Communications in Statistics – Simulation and Computation, 39*(3), 624–664. https://doi.org/1.1080/03610910903548952

Maurissen, J., & Vidmar, T. (2017). Repeated-measure analyses: Which one? A survey of statistical models and recommendations for reporting. *Neurotoxicology and Teratology, 59*, 78–84. https://doi.org/1.1016/j.ntt.2016.1.003

Meltzer, J. A. (2001). *The effects on Type I error rate and power of the single-factor repeated measures ANOVA F-test and selected alternatives under non-normality and non-uniformity* [Doctoral dissertation, The State University of New Jersey]. ProQuest Dissertations Publishing.

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

Moskowitz, D. S., & Hershberger, S. L. (2013). *Modeling intraindividual variability with repeated measures data: Methods and applications*. Taylor & Francis. https://doi.org/1.4324/9781410604477

Raghavarao, D., & Padgett, L. (2014). *Repeated measurements and cross-over designs*. John Wiley & Sons.

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*(2), 283–288. https://doi.org/10.1111/j.2044-8317.1992.tb00993.x

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

Schober, P., & Vetter, T. (2018). Repeated measures designs and analysis of longitudinal data: If at first you do not succeed – try, try again. *Anesthesia and Analgesia, 127*(2), 569–575. https://doi.org/1.1213/ANE.0000000000003511

Sheskin, D. J. (2003). *Handbook of parametric and nonparametric statistical procedures*. Chapman and Hall/CRC.

Singh, V., Rana, R., & Singhal, R. (2013). Analysis of repeated measurement data in the clinical trials. *Journal of Ayurveda and Integrative Medicine, 4*(2), 77–81. https://doi.org/1.4103/0975-9476.113872

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*(4), 147–151. https://doi.org/10.1027/1614-2241/a000016

Tabachnick, B. G., & Fidell, L. (2007). *Experimental designs using ANOVA*. Thomson.

Tamura, R., & Buelke-Sam, J. (1992). The use of repeated measures analyses in developmental toxicology studies. *Neurotoxicology and Teratology, 14*(3), 205–21. https://doi.org/1.1016/0892-0362(92)90018-6

Tippey, K., Ritchey, P., & Ferris, T. (2015). Crossover-repeated measures designs: Clarifying common misconceptions for a valuable human factors statistical technique. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 59*(1), 342–346. https://doi.org/10.1177/1541931215591071

Vallejo, G., Fernández, P., & Livacic-Rojas, P. (2010). Pruebas robustas para modelos ANOVA de dos factores con varianzas heterogéneas [Robust tests for two-way ANOVA models under heteroscedasticity]. *Psicológica, 31*(1), 129–148.

Vallejo, G., Fernández, M. P., Livacic-Rojas, P. E., & Tuero-Herrero, E. (2011). Comparison of modern methods for analyzing repeated measures data with missing data. *Multivariate Behavioral Research, 46*(6), 900–937. https://doi.org/10.1080/00273171.2011.625320

Vallejo, G., & Lozano, L. (2006). Modelos de análisis para diseños multivariados de medidas repetidas [Multivariate repeated measures designs]. *Psicothema, 18*(2), 293–299.

Verma, J. P. (2016). *Repeated measures design for empirical researchers*. Wiley.

Wilcox, R. R. (2022). *Introduction to robust estimation and hypothesis testing* (5th ed.). Academic Press.

Zhao, J., Wang, C., Totton, S., Cullen, J., & O'Connor, A. (2019). Reporting and analysis of repeated measurements in preclinical animal experiments. *PloS One, 14*(8), Article e0220879. https://doi.org/1.1371/journal.pone.022087