STATISTICS IN TRANSITION new series, December 2022
Vol. 23, No. 4, pp. 77–90, DOI 10.2478/stattrans-2022-0043
Received – 09.09.2021; accepted – 01.08.2022

Robustness of randomisation tests as alternative analysis methods for repeated measures design

Abimibola Victoria Oladugba[1], **Ajali John Obasi**[2], **Oluchukwu Chukwuemeka Asogwa**[3]

ABSTRACT

Randomisation tests (*R*-tests) are regularly proposed as an alternative method of hypothesis testing when assumptions of classical statistical methods are violated in data analysis. In this paper, the robustness in terms of the type-I-error and the power of the *R*-test were evaluated and compared with that of the *F*-test in the analysis of a single factor repeated measures design. The study took into account normal and non-normal data (skewed: exponential, lognormal, Chi-squared, and Weibull distributions), the presence and lack of outliers, and a situation in which the sphericity assumption was met or not under varied sample sizes and number of treatments. The Monte Carlo approach was used in the simulation study. The results showed that when the data were normal, the *R*-test was approximately as sensitive and robust as the *F*-test, while being more sensitive than the *F*-test when data had skewed distributions. The *R*-test was more sensitive and robust than the *F*-test in the presence of an outlier. When the sphericity assumption was met, both the *R*-test and the *F*-test were approximately equally sensitive, whereas the *R*-test was more sensitive and robust than the *F*-test when the sphericity assumption was not met.

Key words: randomisation test, repeated measures design, sensitivity, robustness, Monte Carlo.

1. Introduction

Research in many areas of application as affirmed by Ma et al. (2012) normally involves study plans in which measurements or responses are repeatedly obtained from an experimental unit (EU). According to Davis (2002), repeated measurements refer broadly to data in which the response of each experimental unit or subject is observed on multiple treatment conditions or time points. Repeated measures design (RMD)

[1] Corresponding Author. Department of Statistics, University of Nigeria, Nsukka, Nigeria. E-mail: abimibola.oladugba@unn.edu.ng. ORCID: https://orcid.org/0000-0002-6402-8833.

[2] Department of Statistics, University of Nigeria, Nigeria. ORCID: https://orcid.org/0000-0002-4761-9682.

[3] Department of Mathematics/Computer Science/Statistics and Informatics, Alex Ekwueme Federal University Ndufu Alike Ikwo, Nigeria. ORCID: https://orcid.org/0000-0001-7297-9201.

is an experimental design that involves multiple measures of the same variable(s) taken on the same EU either under different treatment conditions or over two or more time periods (Kreuger and Tian, 2004).The major advantage of RMD is that it uses exactly the same individuals or subjects in all treatment conditions thereby eliminating the influence of individual differences from the analysis and also being economical in the use of resources and enabling the subjects to be their own control as measurements are taken under both control and other experimental conditions (Reed III, 2003; Howitt and Cramer, 2011).

An approach to RMD data analysis is the repeated measures analysis of variance (RM ANOVA) that is based on the *F*-test statistic which has assumptions that must be met to ensure valid results are obtained from the analysis and therefore is limited in its application (Dragset, 2009). The assumptions include random sampling of EU from the population, normality of responses, and equality of all pairwise differences in variance between experimental conditions called sphericity (Girden, 1992, Lindman, 1992). The *F*-test is a statistical test in which the sampling distribution of the test statistic has an *F*-distribution when the null hypothesis is true (Oladugba et al., 2014). In statistical analysis, if the assumptions for any parametric test cannot be satisfied, there is risk of passing invalid inference if such test is deployed. So, researchers either transform the response data so that the resulting variable meets the conditions of the intended test to be used or resort to a different test such as the non-parametric test, which is not affected by the assumptions of the parametric test (Zimmerman and Zumbo, 1990) but transformation of data according to Sawilowsky et al. (1989) can have poor power properties. Also, the use of ranks in nonparametric tests leads to loss of information, thus the researchers cannot rely with high confidence level on ranking or transformation of data as an alternative to the *F*-test when its assumptions are not met (Gleason, 2013).

Randomization test (R-test) or permutation test can provide excellent solutions in the presence of unsuitable conditions for the use of the *F*-test or when the researchers want to maintain the use of the original data. The *R*-test is a way of hypothesis testing that can be deployed for analysis of experimental data when assumptions of parametric tests are not tenable (Edgington, 1995; Kherad-Pajouh and Renaudi, 2014). It provides an efficient approach to hypothesis testing. In other words, the *R*-test is perceived as an alternative method to data analysis in conditions when assumptions of parametric procedures are not met (Craig and Fisher, 2019; Berry et al., 2018). *R*-test performs well in conditions not favourably for the *F*-test and is as sensitive and robust as the *F*-test when parametric test assumptions are met (Mundry, 1999; Mewhort, 2005; Mewhort et al., 2010).

Since the validity of any statistical inference depends largely on satisfaction of the assumptions of the underlying model, researchers should not anticipate any statistical

test to be the most appropriate in any situation but rather subject proposed statistical test to scrutiny to ensure it is better than other alternatives in terms of sensitivity and robustness (Peres-Neto and Olden, 2001). The sensitivity of a test is the ability of a test to make right decision vis-à-vis rejection or acceptance of a hypothesis also known as power of a test; it is greatly influenced by sample size and presence of outliers (Cohen, 1988) and assumption of sphericity (constant variance) for RMD (Dragset, 2009) while robustness, on the other hand, refers to the ability of a test to yield correct conclusion or perform optimally in terms of controlling the type-I-error (p) that is not to falsely detect an effect when some of the distributional assumptions are not met or under unfavourable conditions (Vorapongsathorn et al., 2004).

Hence, this paper used the *R*-test to analyse the RMD and compared the results to that of the *F*-test in order to find out which was more sensitive and robust under the conditions that data are normal and non-normal (exponential, lognormal, Chi-square, and Weibull distributions), in the absence and presence of outliers, when sphericity assumption was met or not in variant number of treatments and sample sizes.

2. Materials and methods

2.1. Material

The data presented in Table 1 were obtained from Gravetter and Wallnau (2007). The responses generated from the study were based on the time (in seconds) lapsed until participants reported they felt nothing called latency when a stimulus (of 500-milligram weight) was gently placed on a region of the body. The study compared the adaptation for four regions of the body for a sample of 7 participants.

Table 1. Data on sensory adaption experiment

	Area of stimulation (Treatment)			
Subjects	**Back of hand**	**Lower back**	**Middle of Palm**	**Chin below lower Lip**
1	6.5	4.6	10.2	12.1
2	5.8	3.5	9.7	11.8
3	6.0	4.2	9.9	11.5
4	6.7	4.7	8.1	10.7
5	5.2	3.6	7.9	9.9
6	4.3	3.5	9.0	11.3
7	7.4	4.8	10.8	12.6

2.2. The *F*-test method for analysis of single factor RMD

The *F*-test procedure for hypothesis testing in analysis of RMD involves computing the *F*-statistic associated with the problem. In this section, the model, ANOVA table

presented in Table 2 and the *F*-test procedures for analysing single factor RMD are defined as follows.

The model for this design is defined as:

$$y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij} \quad i = 1, 2, \ldots, n; j = 1, 2, \ldots, t$$

$$\sum_{j=1}^{t} \tau_j = 0; \beta_i \sim N(0, \sigma_\beta^2); \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

where, y_{ij} is the response from the i^{th} subject at treatment j; μ is the grand mean; τ_j is the fixed effect of the j^{th} treatment (the treatments are assumed to have fixed effects thus the zero-sum constraint); β_i is the random effect for i^{th} subject and ε_{ij} is a random error component specific to i^{th} subject at j^{th} treatment.

Table 2. ANOVA table for single factor RMD

Source of Variation	SS	df	Mean Square	F_0
Subject	SS_B	n -1	MS_S	
Treatments	SS_T	t – 1	MS_T	$\frac{MS_T}{MS_E}$
Error	SS_E	$(t - 1)(n - 1)$	MS_E	
Total	SS_T	tn – 1		

where $MS_S = \frac{SS_B}{n-1}$; $MS_T = \frac{SS_T}{t-1}$; $MS_E = \frac{SS_E}{(t-1)(n-1)}$. The sums of squares are then defined as follows:

$SS_T = \sum_{i=1}^{n}\sum_{j=1}^{t}(\bar{y}_{.j} - \bar{y}_{..})^2$; $SS_S = \sum_{i=1}^{n}\sum_{j=1}^{t}(\bar{y}_{i.} - \bar{y}_{..})^2$; $SS_E = \sum_{i=1}^{n}\sum_{j=1}^{t}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} - \bar{y}_{..})^2$

2.3. Randomization test procedure

The hypothesis to be tested is:

H_o: the different treatments had the same effect *vs* H_1: there is a differential effect of at least one treatment

$\alpha = 0.05$

Test statistic

Here, *F*-statistic was used as the test statistic. It summarizes the differences between means and eliminates the effects of between-subject variability.

Procedure

With repeated measures, we permute the data within subject. If there is no effect of treatments, then the set of scores from any subject can be exchanged across treatments.

The steps are as follows:

- Compute the *F*-statistic for the original data, and denote that as F_{cal}.
- Permute the data within each subject, and do it for every subject.

- Calculate an F-statistic for each of the permuted data.
- If this F-statistic is greater than F_{cal}, increment the counter.
- Repeat the preceding three steps B times, where $B \geq 10{,}000$.
- Divide the value in the counter by B to obtain the probability of obtaining an F-statistic as large as F_{cal} if the null hypothesis were true. Denote this value as empirical type-I-error (p-value).
- Reject the null hypothesis of no difference due to treatment if p-value is less than our chosen level of significance.

2.4. Randomization test procedure for RMD

The R-test for analysing single factor RMD involves the following procedures. Compute a test statistic that sufficiently explains the experimental data (the F-statistic in this case) for the data in Table 1. Afterwards, the data are rearranged within the subject repeatedly and the test statistic is recomputed for all resultant data permutations. Randomization test uses the obtained results from all data permutations and the original result of the experiment to form a reference set which is used to decide the significance of the test. The fraction of the data permutation in the reference set having test statistic values greater than or equal to the value obtained from the original results before data were permuted is the type-I-error (significance or probability value).

In permuting data in RMD, Edgington (1995) proposed two schemes, namely systematic and random permutation schemes. In this paper, the random permutation scheme was adopted and carried out in the following way. Firstly, the data are arranged in a table with k columns and n rows, where k is the number of treatments and n is the number of subjects. An index number 1 to n was assigned to the subjects and 1 to k to the treatments, so that each measurement has associated with it a compound index number, the first part which indicates the subject and the second indicates the treatments. Accordingly, index (2, 3) for instance referred to the measurement for the second person under the third treatment. Then a random number generation algorithm was used to randomly determine for each subject independently of the other subjects which of the k measurements is to be assigned to the first treatment, which of the remaining k-1 measurements to the second treatment, and so on. The random determination of order of measurements within each subject performed over all subjects constitutes a single permutation or arrangement of the data. The arrangement is repeated for a large number of times like 10,000 permutations, and for each permutation, the test statistic is computed. The p-value is computed as the number of the test statistic value, including the obtained test statistics values that are as large as the obtained test statistics value.

2.5. Outlier detection and sphericity assumption

Outliers were randomly injected into the dataset in Table 1, and Tukey's method of outlier detection as explained by Songwon (2006) was used in detecting them. One of the ways to test for sphericity in RMD is the use of Mauchly's test. Mauchly's test tests the hypothesis that the variances of the differences between any two conditions are equal. Thus, if the significance level of Mauchly's test is less than or equal to the alpha level, sphericity is violated. Mauchly's test of sphericity in SPSS version 22 was used to verify this condition.

2.6. Monte Carlo Simulation

In order to analyse RMD with the *R*-test and the *F*-test so as to check their robustness, a Monte Carlo simulation was conducted using RMD in Table 1 with $n = 7$ subjects and $t = 4$ treatments. Three variables were manipulated: (i) sample sizes (n); (ii) number of treatments (t); and (iii) distribution structure of the data (normal, exponential, lognormal, Chi-square and Weibull distributions). The performance of the two tests was investigated with three sample conditions $n = 5$, 7, and 9, and three treatment conditions $t = 3$, 4, and 5, under 5 distributional structures of the data in the presence and absence of outliers and when sphericity assumption is met or not, respectively.

The *R statistical package* was used to implement the Monte Carlo technique sampling of 10,000 permutations from the possible $(t!)^n$ permutations for the *R*-test. In the simulation, the experiment was repeated 1000 times for each distribution. In each repetition, the resulting tables of data set were analysed appropriately using the *F*-test and the *R*-test methods to obtain the rate of type-I-error and power. The percentage of significant tests out of 10,000 iterations was considered as the rejection rate.

The comparison procedures were considered in two scenarios. Firstly, in the scenario that the null hypothesis (H_0: $\mu_i = 0$) is true, the rejection rate of the null hypothesis was regarded as the type-I-error rate for each test. The test that had the closest type-I-error to the nominal $\alpha = 0.05$ was considered as the more robust of the two. Secondly, in the scenario that the alternative hypothesis (Ha: $\mu_i \neq 0$) was true, the rejection rate of the null hypothesis was considered as the power for each test. The test that had larger power was taken to be more sensitive than the other.

2.7. Distribution structure of data

Data were simulated from five theoretic distributions. The normal distribution was used to test condition under which normality assumption holds. The skewed distributions used include Chi-square, exponential, lognormal, and Weibull distributions; this represents condition under which the distribution assumption (normality) does not hold. The probability density function of the five distributions is defined as follows.

(a) Normal distribution

The normal distribution has probability density function (pdf) as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < \mu < \infty, \ \sigma > 0, \ x > 0$$

The parameters (μ and σ^2) of the normal distribution were estimated using the maximum likelihood estimators (MLE). For the normal distribution, data were simulated using mean, $\bar{x}$ = 7.7250 and variance, σ^2= 9.1180, of the experimental data.

(b) Exponential distribution

The exponential distribution has pdf with parameter θ is given by

$$f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}} \quad , \ \theta > 0, x \geq 0$$

Data were simulated to follow the exponential distribution using the MLE of the exponential distribution parameters obtained as $\hat{\theta}$= 7.7250 as fitted using *fitdistrplus* package in *R* statistical computing.

(c) Chi-square distribution

The pdf of Chi-square distribution with parameter *n*, is given as

$$f(x) = \frac{x^{\frac{n}{2}-1} e^{-x/2}}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})}, \ x > 0$$

Using *fitdistrplus* package in *R* statistical computing, the parameter of the Chi-square distribution,
n = 4.559 ~ 5, was used for simulation of data where n is the mean of Chi-square distribution.

(d) Lognormal distribution

The pdf for the two-parameter (μ and σ^2) lognormal distribution is

$$f(\mathrm{X}|\mu,\sigma^2) = \frac{1}{X\sqrt{(2\pi\sigma^2)}} e^{-\left[\frac{(\ln(X)-\mu)^2}{2\sigma^2}\right]}, \qquad X > 0, \ -\infty < \mu < \infty, \ \sigma > 0$$

The MLE of μ and σ^2 were obtained as:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \boldsymbol{ln}\,(\boldsymbol{X_i})}{n} = 7.7880 \text{ and } \hat{\sigma}^2 = \frac{\left(\sum_{i=1}^{n} (\boldsymbol{ln}\,(\boldsymbol{X_i}) - \frac{\sum_{i=1}^{n} \boldsymbol{ln}\,(\boldsymbol{X_i})}{n}\right)^2}{n} = 12.0120$$

(e) Weibull distribution

The two-parameter Weibull distribution has pdf given as

$$f(x/k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{(x/\lambda)^k}, \quad x \geq 0, \ k > 0, \lambda > 0$$

Data were simulated using the MLE of the Weibull distribution parameters obtained as $k = 7.020$ and $\lambda = 9.10$ as fitted using *fitdistrplus* package in *R* statistical computing.

3. Results

In Tables 3, 4, 5 and 6, the simulation results (type-I-error and power) for the *F*-test and the *R*-test based on the three manipulated variables (sample size, number of treatments and distribution structure of the data) are presented. Following from the methods mentioned in Section 2 as implemented in *R Statistical package*, for each sample size, the optimal values of the type-I-error and power were recorded. The sample size was denoted as n, the values in bracket indicate the number of treatment (t) that produced optimal type-I-error and highest power as the number of treatments were varied. The values in bold are either the optimal type-I-error or the highest power for each of the test.

Table 3 shows the type-I-error of the *F*-test and the *R*-test for the data in the absence of outliers. The results indicated that as n increased, type-I-error decreased for data with normal distribution, for Chi-square, lognormal, exponential and Weibull, it initially increased but afterwards decreased for the *F*-test while the *R*-test produced type-I-error that increased as n increased under the normal distribution but reduced as n increased for Chi-square, exponential and Weibull, while for data with lognormal distribution, the type-I-error decreased as n decreased. On the other hand, the power values for the normal data decreased initially but later increased as the sample size increased, it increased initially and subsequently decreased for Chi-square and Weibull distributions for the *F*-test and increased for exponential and lognormal data distributions. The *R*-test on the other hand had increasing power as n increased for exponential, Weibull, and Chi-square but had an increasing trend for lognormal although with a slight initial decrease at $n = 7$. When outliers were introduced, the type-I-error and power values are presented in Table 4. The results indicated that type-I-error for the *F*-test under all the data distributions had a decreasing trend as n increased but an increasing trend for Weibull distribution. Furthermore, the power for the *F*-test exhibited a slight decreasing trend for Chi-square and lognormal, while it increased for normal, exponential and Weibull as n increased. On the other hand, the power values of the *R*-test for all data distribution were increasing as sample size increased.

The results for when sphericity assumption was met are displayed in Table 5. The type-I-error for the *F*-test in this table revealed that as n increased, normal and exponential data distributions initially had a slight increasing trend but substantially increased afterwards for Weibull data distribution while a decreasing trend was observed for Chi-square and lognormal data distribution. The results of sphericity

assumption not met as displayed in Table 6 showed that type-I-error and power for both tests decreased for all distributions as the sample size increased.

Table 3. Simulation results (type-I-error and power) for *F*-test and *R*-test in the absence of outliers

		type-I-error		Power	
Distribution	***n (t)***	***F*-test**	***R*-Test**	***F*-test**	***R*-Test**
Normal	5(5)	0.**0500**	0.**0429**	0.9437	0.9226
	7(4)	0.0428	0.0546	0.9122	0.9179
	9(5)	0.0408	0.0595	0.**9914**	0.**9805**
Exponential	5(5)	0.0725	0.**0501**	0.7093	0.9032
	7(4)	0.1442	0.0613	0.7528	0.9211
	9(5)	0.**0611**	0.0581	0.**7828**	0.**9469**
Lognormal	5(5)	0.**0413**	0.0593	0.7229	0.8534
	7(5)	0.0662	0.0439	0.7237	0.8629
	9(5)	0.0599	0.**0490**	0.**8009**	0.**8979**
Chi-square	5(4)	0.0705	0.**0524**	0.6184	0.7528
	7(4)	0.1009	0.0687	0.**6729**	0.7367
	9(5)	0.**0704**	0.0591	0.5646	0.**8086**
Weibull	5(5)	0.0849	0.0640	0.5256	0.7439
	7(5)	0.1225	0.0580	0.**6804**	0.7811
	9(5)	0.0783	0.**0441**	0.5959	0.**8724**

Table 4. Simulation results (type-I-error and power) for *F*-test and *R*-test in the presence of outliers

		type-I-error		Power	
Distribution	***n (t)***	***F*-test**	***R*-Test**	***F*-test**	***R*-Test**
Normal	5(4)	0.1029	0.0699	0.5790	0.7498
	7(4)	0.**0824**	0.0601	0.5617	0.**8209**
	9(5)	0.0873	0.**0588**	0.**5869**	0.7998
Exponential	5(5)	0.2018	0.**0566**	0.5958	0.7909
	7(5)	0.**0755**	0.1003	0.**6963**	0.7304
	9(5)	0.1046	0.0708	0.5540	0.**8202**
Lognormal	5(5)	0.0815	0.0597	0.**6876**	0.7588
	7(5)	0.1174	0.0632	0.6011	0.6901
	9(5)	0.**0792**	0.**0512**	0.5906	0.**8094**
Chi-square	5(4)	0.2171	0.0696	0.5377	0.8132
	7(4)	0.1024	0.0741	0.5213	0.7995
	9(5)	0.**0843**	0.**0516**	0.**6628**	0.**8180**
Weibull	5(5)	0.**0818**	0.**0536**	0.5448	0.7800
	7(5)	0.1032	0.0684	0.**5994**	0.7468
	9(5)	0.0929	0.0684	0.5834	0.**7933**

Table 5. Simulation results (type-I-error and power) for *F*-test and *R*-test with sphericity assumption met

		type-I-error		Power	
Distribution	***n (t)***	***F*-test**	***R*-Test**	***F*-test**	***R*-Test**
Normal	5(4)	0.**0506**	0.0420	0.9637	0.9024
	7(5)	0.0431	0.0446	0.9202	0.9231
	9(5)	0.0521	0.**0511**	0.**9884**	0.**9531**
Exponential	5(4)	0.1011	0.0429	0.8663	0.9001
	7(5)	0.1502	0.0559	0.8818	0.8965
	9(5)	0.**0841**	0.**0523**	0.**9212**	0.**9045**
Lognormal	5(4)	0.0706	0.0462	0.8291	0.8088
	7(5)	0.0762	0.**0518**	0.8119	0.8321
	9(5)	0.**0699**	0.0442	0.**8921**	0.**8899**
Chi-square	5(5)	0.**0589**	0.**0493**	0.6610	0.8011
	7(5)	0.1209	0.0621	0.6690	0.7822
	9(5)	0.1022	0.0489	0.**6710**	0.**8399**
Weibull	5(4)	0.**0820**	0.0531	0.5006	0.7877
	7(4)	0.1015	0.**0429**	0.**5094**	0.8807
	9(5)	0.1183	0.0401	0.5009	0.**8991**

Table 6. Simulation results (type-I-error and power) for *F*-test and *R*-test with sphericity assumption not met

		type-I-error		Power	
Distribution	***n (t)***	***F*-test**	***R*-Test**	***F*-test**	***R*-Test**
Normal	5(5)	0.1112	0.0612	0.**6821**	0.8080
	7(5)	0.1230	0.0610	0.5417	0.8526
	9(5)	0.**0811**	0.**0588**	0.6809	0.**8595**
Exponential	5(4)	0.2074	0.0640	0.5958	0.8522
	7(5)	0.**0603**	0.0595	0.4993	0.8032
	9(5)	0.1032	0.**0467**	0.**6240**	0.**8704**
Lognormal	5(4)	0.1401	0.0531	0.4876	0.7863
	7(5)	0.0631	0.0699	0.4211	0.7902
	9(5)	0.**0503**	0.**0518**	0.**5906**	0.**8186**
Chi-square	5(5)	0.0813	0.4040	0.5307	0.7863
	7(5)	0.1109	0.0601	0.5213	0.**8039**
	9(5)	0.**0705**	0.**0517**	0.**6028**	0.7995
Weibull	5(4)	0.1207	0.0485	0.5448	0.7904
	7(4)	0.**0779**	0.0590	0.**5994**	0.**8002**
	9(5)	0.1052	0.**0508**	0.5891	0.7808

4. Discussion of results

The *R*-test and the *F*-test were used to analyse RMD with and without outlier and sphericity respectively. From the results, under the normal assumption, the type-I-error of both tests was within limits regarded as being robust with the *F*-test producing a better value at $n = 5$ ($p = 0.05$) while the power of both the *F*-test and *R*-test was very high (0.9914 and 0.9805 respectively) and it increased as the sample size and the number of treatments increased. This implies that both tests were approximately equally sensitive and robust under normal assumption. For the exponentially distributed data, as the sample size increased, the optimal type-I-error for the *F*-test was at $n = 9$, $t = 5$ ($p = 0.0611$) and $n = 5$, $t = 5$ for the *R*-test ($p = 0.051$), whereas the highest power for the *F*-test and the *R*-test was 0.7828 and 0.9469 respectively at $n = 9$, $t = 5$, which shows that the *R*-test was more powerful that the *F*-test and more robust too for exponential data. For lognormal distribution, the optimal type-I-error for both tests as the sample size increased was 0.0413 and 0.0490 for the *F*-test and the *R*-test respectively, while both tests exhibited power of 0.8009 and 0.8979 at $n = 9$ respectively for the *F*-test and the *R*-test. For the Chi-square distribution, the *F*-test had optimal type-I-error of 0.0704 at $n = 9$, $t = 5$ and 0.0524 for *R*-test at $n = 9$, $t = 5$. Also, the highest power of the *F*-test and the *R*-test was 0.6729 ($n = 7$) and 0.8076 ($n = 9$) respectively. For the Weibull distribution, the *R*-test was more robust with $p = 0.0441$ and more powerful with power = 0.8724.

When outliers are present, the *R*-test was more powerful and robust in all distributions: normal assumption ($p = 0.0588$, power = 0.8209), exponential distribution ($p = 0.0566$, power = 0.8202), lognormal ($p = 0.0512$, power=0.8094), Chi-square ($p = 0.0516$, power =0.8180), Weibull ($p = 0.0536$, power = 0.7933).

When sphericity condition was met, the *F*-test was more powerful and robust ($p = 0.0506$, power = 0.9531) for data with normal distribution while the *R*-test was more powerful and robust for lognormal data ($p = 0.0518$, power = 0.8899), Chi-square ($p = 0.0493$, power = 0.8399), Weibull distribution ($p = 0.0429$, power = 0.8991). For exponential data, the *F*-test was more robust for data ($p = 0.0523$) while the *R*-test was more powerful (0.9212). Furthermore, when sphericity assumption was not met, the *F*-test was only more robust for lognormal ($p = 0.0503$) while the *R*-test was more powerful (power = 0.8186). Meanwhile, the *R*-test was more robust and powerful for other distributions – normal ($p = 0.0588$, power = 0.8595), exponential ($p = 0.0467$, power = 0.8704), Chi-square ($p = 0.0517$, power = 0.8039), and Weibull ($p = 0.0508$, power = 0.8002).

5. Conclusion

In this paper, the *R*-test was used in analysing RMD with or without outlier and sphericity respectively. The test offers the freedom of choice of test statistic that sufficiently suits a particular statistical problem for researchers and is free from any distributional or test assumptions, but rather depends only on the randomization technique – thus the name the randomization test. The study also employed the classical test (*F*-test) for analysing RMD, which is hinged on a number of conditions for reliable valid inference. This paper compared both tests to ascertain which controlled the type-I-error better and had higher power than the other. These criteria of comparison were referred to as robustness and sensitivity respectively.

The results in Tables 3, 4, 5 and 6 showed that under the normal distribution when sphericity held, both tests were equally robust and approximately powerful with optimal values at $n = 5$, $t = 5$ ($p = 0.05$ power = 0.9914) for the *F*-test and at $n = 9$, $t = 5$ ($p = 0.0421$, power = 0.9805) for the *R*-test. When data had skewed distributions (exponential, Chi-square, lognormal and Weibull), the *R*-test was more robust and powerful. In the presence of an outlier and when sphericity condition was not met, the *F*-test was less robust and sensitive than the *R*-test. In the analysis of RMD when normality and sphericity conditions were met, the *R*-test was comparably as robust and sensitive as the *F*-test. When data had skewed distributions (exponential, lognormal, Chi-square and Weibull), the *F*-test was less robust and sensitive as the sample size and the number of treatments increased. Also, in the presence of an outlier and when sphericity condition was met or not, the *R*-test was more robust and sensitive than the *F*-test. In a nutshell, the *R*-test was approximately as sensitive as the *F*-test in RMD when data follow normal and sphericity conditions met but more sensitive when data were skewed (exponential, Chi-square, lognormal and Weibull).

In general, since the *R*-test is always as robust and sensitive and even more robust and sensitive than the *F*-test, to alleviate the burden of assessing parametric assumptions which is done before the use of the *F*-test, researchers are advised to go ahead with *R*-test which is not based on any assumption and is easily carried out with modern-day high-capacity computers.

References

Berry, K., Johnston, J., Mielke, P., (2018). *Permutation Statistical Methods: A Permutation Statistical Approach*, doi: 10.1007/978-3-319-98926-6_2.

Cohen, J., (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). New Jersey: *Lawrence Earlbaum Associates*.

Craig, A. R., Fisher, W. W., (2019). Randomization tests as alternative analysis methods for behavior-analytic data. *Journal of the Experimental Analysis of Behavior*, 111(2), pp. 309–328.

Davis, C. S., (2002). Statistical Methods for the Analysis of Repeated Measurements. New York, NY: *Springer Publishers*.

Dragset, I. G., (2009). Analysis of longitudinal data with missing values: Methods and Applications in Medical Statistics (Master's Thesis). Available from *Norwegian university of science and technology digital theses database*.

Edgington, E. S., (1995). Randomization Tests (3rd Ed). New York, NY: *Marcel Dekker*.

Girden, E. R., (1992). ANOVA: Repeated measures. *Sage Publications*, Newbury Park, CA.

Gleason, J., (2013). Comparative power of the ANOVA, randomization ANOVA, And Kruskal-Wallis test (Doctoral Dissertation). Available from *Wayne State University Digital Dissertations database*.

Gravetter, F. J., Wallnau, L. B., (2007). Statistics for the behavioral science. Canada: *Vicki Knight*.

Howitt, D., Cramer, D., (2011). Introduction to Research Methods in Psychology (3rd ed.). *Essex: Pearson Education Limited*.

Kherad-Pajouh, S., Renaud, O., (2014). A general permutation approach for analyzing repeated measures ANOVA and mixed-model designs. *Computational Statistics and Data Analysis*, 21 (5), pp. 42–59.

Krueger, C., Tian, L., (2004). A Comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points. *Biological Research for Nursing*, 6, pp. 151–157.

Lindman, H. R., (1992). Analysis of Variance in Experimental Design. New York: *Springer-Verlag*.

Ma, Y., Mazumdar M., Memtsoudis, S. G., (2012). Beyond repeated-measures analysis of variance: advanced statistical methods for the analysis of longitudinal data in anesthesia research. *Regional Anesthesia and Pain Medicine*, 37, pp. 99–105.

Mewhort, D. J. K., (2005). A comparison of the randomization test with the F test when error is skewed. *Behavior Research Methods*, 37 (3), pp. 426–435.

Mewhort, D. J. K., Johns, B. T., Kelly, M., (2010). Applying the permutation test to factorial designs. *Behavior Research Methods*, 42 (2), pp. 366–372.

Mundry, R., (1999). Testing related samples with missing values: a permutation approach. *Animal Behaviour*, 58, pp. 1143–1153.

Oladugba, A. V., Udom, A. U., Ugah, T. E., Ukaegbu, E. C, Madukaife, M. S., Sanni, S. S., (2014). Principles of Applied Statistics. Nsukka: *University of Nigeria Press Ltd.*

Peres-Neto, P. R., Olden, J., (2001). Assessing the robustness of randomization tests: examples from behavioral Studies. *Animal Behaviour*, 61, pp. 79–86.

Reed III, J., (2003). Analysis of variance (ANOVA) models in emergency medicine. The *Journal of Emergency and Intensive Care Medicine*, 7(2), pp. 21–34.

Sawilowsky, S. S., Blair, R. C., Higgins, J. J., (1989). An investigation of the type-I-error and power properties of the rank transformation procedure in factorial ANOVA. *Journal of Educational Statistics*, 14 (3), pp. 255–267.

Songwon, S., (2006). A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets (Master's thesis) available from University of Pittsburgh, *Graduate School of Public Health database*.

Vorapongsathorn, T., Taejaroenkul, S., Viwatwongkasem, C., (2004). A comparison of type-I-error and power of Bartlett's test, Levene's test and Cochran's test under violation of assumptions. S*ongklanakarin Journal of Science and Technology*, 26(4), pp. 537–547.

Zimmerman, D. W., Zumbo, B. D., (1990). Effect of outliers on the relative power of parametric and nonparametric statistical tests. *Perceptual and Motor Skills*, 71, pp. 339–349.