JOURNAL OF COMPUTATIONAL INNOVATION AND ANALYTICS
https://e-journal.uum.edu.my/index.php/jcia

THE IMPLEMENTATION OF THE ALEXANDER-GOVERN TEST IN FACTORIAL DESIGN ANALYSIS

[1]Nurul Syafiqah Ishak Latfi & [2]Suhaida Abdullah
[1]Appfuxion Consulting Sdn Bhd, Malaysia
[2]School of Quantitative Sciences,
Universiti Utara Malaysia, Malaysia

[1]Corresponding author: inurulsyafiqahishaklatfi@gmail.com

ABSTRACT

This study proposed to evaluate the performance of the Alexander-Govern test (*AG test*), Analysis of Variance (*ANOVA*), and *t*-test by analyzing the Type I error rate. The *AG* test is regarded as a reliable control Type I error rate. This technique is insensitive in the presence of heteroscedasticity under a normal distribution. Simulation research was carried out using Statistical Analysis Software (SAS) to assess the effectiveness of the tests that are based on the rate of Type I error. By creating the conditions that could highlight the strengths and weaknesses of each test, three variables are being manipulated: sample size, variance heterogeneity, and type of pairings. The performance of the *AG* tests is convincing when it is able to control the Type I error rates better compared to *ANOVA* under all conditions of heterogeneous variances. Meanwhile, the *ANOVA* performs best only when the variances are homogenous. A real data experiment was

applied to validate the result. In the battery life design experiment, the *p*-value using the *AG* test and *ANOVA* are computed and compared. The *AG* test provides valid results when it can test the main effect and the interaction effect, as well as the *ANOVA*. With good performance in the simulation study, the AG test can be considered a good alternative to the *ANOVA* when the assumptions of the homogeneity of variances are violated in the case of factorial design.

Keywords: Alexander-Govern Test, *ANOVA*, Normality Distribution, *t*-test, Type I error rate.

INTRODUCTION

Traditional methods, such as the *t*-test and the Analysis of Variance (*ANOVA*), are commonly used to compare the independence of two groups and more than two groups, respectively, and are always favoured by researchers. Based on a sample of data, the *t*-test determines whether there is a statistically significant difference between the means of the two groups. When conducting a *t*-test, it is typical to make the following assumptions: the measuring scale, random sampling, normality of the data distribution, sufficiency of the sample size, and equality of variance. When data invalidates the assumptions, the *t*-test may not be reliable.

Meanwhile, *ANOVA* is a classic test for comparing three or more means. The assumptions needed are almost the same as those for the t-test, where data need to be normally distributed, samples are independent, and variances are homogeneous. Since these tests have similar assumptions, their ability to identify any differences will be disrupted by the same problems, such as unequal variances. Both tests have poor control of Type I error rate when data is not homogeneous. In the case of unequal variance, also known as heteroscedasticity, the *Welch t*-test is the common option to the classical *t*-test. For the case of more than two groups, various tests were introduced, such as the *Welch* test, *James's* test, and the Alexander-Govern test (*AG* test) (Ochuko et al., 2015c).

Alexander and Govern (1994) invented the *AG* test to test the equality of *k*-independent means in the case of variance heterogeneity (Abdullah et al., 2014). The test compares samples from two or more

groups, possibly of varying sizes, and deals with unequal variance. For variance heterogeneity in normal data, the *AG* test provides excellent control of Type I error rates (Schneider & Penfield, 1997). This test, however, is not robust to non-normal data because it cannot handle noisy data or outliers when sample sizes are unequal (Ochuko et al., 2015a).

In terms of data distribution, researchers commonly refer to the central limit theorem to assume their sample is normally distributed. This theorem needs a sufficient sample size that is at least 30. In many conditions in the real-life situation, a large sample is hard to attain. In some cases, only small data is available. The *t*-test, which usually handles a small sample, will have no problem as long the data is normally distributed. However, when it is not normal, it will affect the ability of the *t*-test to control the Type I error. Consequently, the result is questionable.

The worst-case scenario is when both assumptions, the homogeneity of variances and normal distribution, are violated. Some practitioners failed to check the necessary assumptions and abused the tests under the violation of the assumptions. Consequently, any decision based on the abused tests will not provide the supposedly needed solution. According to Ochuko et al. (2015b), non-normality and variance heterogeneity are obstacles to *ANOVA*. This increases the likelihood of Type I errors while decreasing power (Ochuko et al., 2015a).

As mentioned previously, the classical *t*-test and *ANOVA* are the most used tests when comparing means between independent groups and are widely used in various fields. In experiments, for instance, these tests provide useful information to identify the significant differences between the tested treatments. Montgomery (2020) provided various experimental designs with their respective analyses. Correct experimental design can reduce variations in the data, leading to better analysis of the treatment effect. It also reduces the tendency to have heterogeneous variances in the data. However, heteroscedasticity can still occur, especially when the sources of the variation are unknown.

The factorial design of experiments is one of the important designs used to investigate the effects of different factors on the response variable (Montgomery, 2020). Hence, the extended *ANOVA* is used

to investigate the main effects of different factors and the effect of the interaction between the factors. Even though variances are assumed to be equal, due to less information on the sources of variation, the condition of heteroscedasticity still can occur. Under this circumstance, the accuracy of the results based on the common extended *ANOVA* might be jeopardized. To date, the use of other tests, such as the *AG* test, is not common in the field of experimental design. No references were found to investigate the implementation of the *AG* test, specifically in the case of factorial design. Hence, this study initiated some simulation trials to identify the *AG* test's performance based on the Type I error rates in analyzing the main effects of the factors and the interaction in the factorial design. Note that the analysis of the real-data experiment was applied to validate the *AG* test.

RELATED WORK

Factorial designs are the most efficient design in experimental studies on the effect of at least two factors (Antony, 2014; Nguyen et al., 2019). It allows investigation in each trial of replication of the experiment of all possible combinations of the levels of the factors (Montgomery, 2020). Crawley (2012) stated that factors are frequently used to represent statistical blocks where these are internally homogeneous units in which each experimental treatment is repeated.

The *ANOVA* analyses the effect of the investigated factors on the response variable by comparing the means. It assumes variances of all groups are homogeneous, where the accuracy of the results also depends on the assumption of normal data distribution. It is employed to determine whether the measure of the means is equal and is resistant to minor deviations from normality, especially when the sample size is sufficient to ensure normality (Allen, 2017; Crawley, 2012; Hodžić & Islamović, 2020). Allen (2017) stated that *ANOVA* is particularly sensitive to the homogeneity of variance assumption. Therefore, when there is a violation, the analysis's findings could be incorrect since the *p*-value might become excessively large or too conservative.

On the other hand, the *AG* test is used as the alternative to the *ANOVA* to deal with unequal variance (Abdullah et al., 2011; Jamaluddin, 2015; Ochuko et al., 2015c, 2015a). The *AG* test provides excellent control of Type I error rates for variance heterogeneity in normal

data. Integration of the *AG* test with some robust estimators, such as trimmed mean, enhances its ability to better control the Type I error rate under non-normal data. It has been tested in several studies using simulation and some real-world data validation (Ochuko et al., 2015c).

Testing the effect of at least two factors needs more advanced *ANOVA*. Montgomery (2020) explained that the usage of two-way *ANOVA*, or more, depends on how many factors are investigated. The variances are assumed to be homogeneous as long as they follow the design procedures. However, in some cases, the variances can still be heterogeneous. Under this condition, the result from the *ANOVA* can be questionable. Hence, the use of the *AG* test is believed to be able to handle the problem of variances heterogeneity. However, to date, no articles have been found related to the study of the implementation of the *AG* test in analyzing factorial design. Mendeş and Yiğit (2013) used transformation techniques and modification on the mean to investigate the effect of the interaction in a factorial design, which involved the case of variances heterogeneous. Their techniques improved the control of Type I error rate, but as mentioned in various research papers such as Abdullah et al. (2014), transformation sometimes cannot handle the heterogeneity of variances, and, in addition, the technique is quite complicated.

The Type I error is a false positive when using any statistical tests, where the error happens when the null hypothesis is rejected while it is false. Evaluating the performance of statistical tests using the ability of the test to control the Type I error rates is common (Abdullah et al., 2011, 2014). A good test is considered if it is able to control the Types I error rates if the error rates fall from 0.5 α to 1.5α as α is the level of significance (Bradley, 1978). For example, a test is deemed robust if the empirical Type I error rates fall within the range of 0.025 to 0.075 because the significance level in this study is set to 0.05. According to Jamaluddin (2015) and Wilcox (2003), when the Type I error rate is less than 0.025, the test is conservative; when it is greater than 0.075, the test is liberal.

METHODOLOGY

This paper will investigate the performance of the *AG* test in testing the effect of factors via factorial design. As discussed previously, the

ANOVA, which is commonly used in the factorial design, is restricted by the assumption of homogeneity variances. The *AG* test is utilized to overcome this problem. This chapter explains the procedure to employ the *AG* test in analyzing a factorial design experiment, specifically identifying factors affecting the response variable. To observe the performance of the *AG* test in the use of factorial design, a simulation study was done to monitor the performance of the *AG* test based on the Type I error rate. This simulation study was designed properly by manipulating some variables, as described in the following sections. The simulation was designed to closely mimic the real experimental conditions. The *AG* test was then validated using real-life data obtained from a previous study of the battery life experiment.

The *AG* Test in Factorial Design

Alexander and Govern (1994) introduced the *AG* test, which has been used to compare two or more groups by utilizing the mean as a gauge of central tendency. Note that this test is able to control the Type I error rate and has high power in the presence of variance heterogeneity for normal data. Nevertheless, it is not resilient to non-normal data (Ochuko et al., 2015c).

The formula of the *AG* test is as follows:

The standard error of each *j*th sample is calculated as

$$S_j = S_{\bar{x}_j} = \sqrt{\frac{\sum_{i=1}^{n_j}(X_i - \bar{X}_j)^2}{n_j(n_j - 1)}}. \tag{1}$$

For each of the *j*th samples, the weight, w_j is calculated so that the $\sum w_j = 1$:

$$w_j = \frac{1/S_j^2}{\sum_{j-1}^{J} 1/S_j^2}. \tag{2}$$

Then, using the weight, w_j, the weighted- mean, X^+ is estimated:

$$X^+ = \sum_{j=1}^{J} w_j \bar{X}_j. \tag{3}$$

The statistic t_j for a sample is determined as:

$$t_j = \frac{\bar{X}_j - X^+}{S_j}. \tag{4}$$

The t_j is distributed to t distribution with $t_j = n_j - 1$ degree of freedom.

Each of the t_j will be normalization transformed to get the z_j.

$$z_j = c + \frac{(c^3 + 3c)}{b} - \frac{(4c^7 + 33c^5 + 240c^{30} + 855c)}{(10b^2 + 8bc^4 + 1000b)}, \tag{5}$$

where $c = [a \ln(1 + t_j^2/v_j]^{1/2}$; $b = 48a^2$ and $a = v_j - 0.5$.

The test statistics, *AG,* is obtained by the summation of the z_j^2 given as follows:

$$AG = \sum_{j=1}^{J} z_j^2. \tag{6}$$

The test statistic, *AG,* is assumed to distribute as a chi-square distribution with *J*-1 degrees of freedom since the *AG* is bigger than χ_{J-2}^2, the null hypothesis will be rejected.

The Design of Data Simulation

The design of data simulation is based on several variables, such as the number of groups, sample sizes, variance values, and types of distributions. This project focused on two factorial designs with two levels for the first and three for the second. It produced a combination of six treatments. Therefore, the number of groups used is six in the simulation data. In an experiment, sample sizes were restricted to the number of experimental units. To set on the sizes of sample, this simulation used the range of small sample ($n = 5$), moderate sample ($n = 20$), and large sample ($n = 30$).

Data is generated using the SAS generator RANNOR, which allows random sampling from a standard normal distribution. To obtain the random sample, the mean of the distributions was set to 0 to make sure the null hypothesis was true. As for the variances, considering the six groups of treatments, they were either homogeneous or heterogeneous. The combination of all these variables produced conditions where, in total, there will be 24 conditions altogether.

Type I Error Rate.

The performance of the *AG* test was evaluated based on its ability to control the Type I error rate using simulation. The null hypothesis was set as all the compared means equal to 0.

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k = 0$$
$$H_1: At\ least\ 1\ mean\ is\ not\ equal\ to\ 0$$

The *AG* test was calculated for each simulation. This process was repeated 5000 times using SAS to generate the p-value using the standard chi-square distribution table with *J* -1 degree of freedom. The value of the significance level was set to 0.05, meaning it is acceptable to have a chance of 5% of making a Type I error. The test is deemed significant if the p-value for the *AG* test is less than 0.05. Otherwise, the test is not significant. From all 5000 times of simulation, the number of rejected tests was counted, and the ratio of this number is considered the Type I error rate. Based on Bradley's criteria, as discussed previously, the acceptable control range of the Type I error rates is between 0.025 and 0.075. Note that values outside the range indicated the test failed to control the Type I error (Ochuko et al., 2015c).

The Real-data Experiment

To validate the *AG* test in the case of factorial design, data from an experiment with a two-factor factorial design was chosen. The experiment investigated the effect of two factors, *i.e.*, materials and temperature, on the life of batteries (Montgomery, 2020). Three materials were used and denoted as type 1, 2, and 3, while two temperature levels were 15°F and 125°F. Four batteries (uniform in size) were tested at each combination of the material types and temperature levels, which made 24 tests run in random order. Table 1 displays the observations from this experiment.

Table 1

The Battery Life (in hours) for the Battery Design Experiment

Material Type	Temperature (°F)	
	15	125
	130	20
	74	82
1	155	70
	180	58

Material Type	Temperature (°F)	
	15	125
2	150	25
	159	58
	188	70
	126	45
3	138	96
	168	104
	110	82
	160	60

Source: Montgomery (2020).

The hypothesis involved in this experiment can be divided into the main effect factors and the interaction between the factors. It can be written as

i. Main effect

Types of materials

$H_0: \mu_1 = \mu_2 = \mu_3$
$H_1: At\ least\ 1\ mean\ battery\ life\ using\ different\ material\ is\ not\ equal$

Temperature

$H_0: \mu_{15} = \mu_{25}$
$H_1: The\ mean\ of\ batery\ life\ under\ temperature\ 15\ is\ not\ equal\ with\ 125$

ii. Interaction effect

$H_0: \mu_{1,15} = \mu_{2,15} = \mu_{3,15} = \mu_{1,125} = \mu_{2,125} = \mu_{3,125}$
$H_1: At\ least\ 1\ mean\ battery\ life\ using\ different\ materials\ and\ temperature\ is\ not\ equal$

In a two-way *ANOVA*, the best estimate of σ^2 is a Mean Square Error (*MSE*). The *MSE* is the test statistic's denominator, and the numerator is the mean sum of squares for each main factor and interaction term. The *F*-statistic is used to answer the alternative hypothesis. The *F*-values for the *ANOVA* and *AG* tests were calculated manually in Excel as well as SPSS software to compare the accuracy of the results.

In conducting a two-way *ANOVA*, the hypothesis regarding the interaction effect needs to be tested first. If the null hypothesis of no interaction is rejected, then there is no interpretation of the results of

the hypotheses involving the main effects. If the interaction term is insignificant, then the two main effects must be examined separately. Before performing the *ANOVA* test, there were a few assumptions: (a) The samples are independent, (b) Each sample is from a normally distributed population, and (c) This test assumes homogeneity of equal variances. Using SPSS software to check the homogeneity of variance, Levene's and Breusch-Pagan's tests were applied. Levene's test is a useful tool for assessing the homogeneity of variance assumption in *ANOVA* and other statistical analyses. Nevertheless, it should be used cautiously and in conjunction with other diagnostic procedures, such as residual and normal probability plots.

The Breusch-Pagan test is useful for assessing the assumption of homoscedasticity in a linear regression model. However, like any statistical test, it should be used in conjunction with other diagnostic tools, such as residual plots, to ensure the validity of the results. Consequently, before performing the *ANOVA* test, the normality of distribution was validated by the Shapiro-Wilk test. If the *p*-value associated with the test is less than the alpha = 0.05, we reject the null hypothesis and conclude that the data set is not normally distributed. Before getting the value for the Shapiro-Wilk test, a standardized residual was calculated.

ANALYSIS AND RESULTS

This section provides findings, the analysis from the simulation study, and the real data. The simulation study used to obtain the Type I error rates which were used to evaluate the performance of the *AG* test compared to the *ANOVA* under various conditions, were set to tally with the conditions of the real data. Table 2 displays the result of the Type I error rates according to the effect of the factor, either the main effect or interaction, as stated in the first column. It is followed by the second column, which refers to the factor itself, where a single factor is used to investigate the main effect, and the combination of at least two factors investigates the interaction effect factors. The sample sizes, as stated in the third column, provide the sizes of the generated sample for each level of the factors. For instance, the first sample sizes of the first row stated (3, 3), which refers to factor 1 with two levels and three sample sizes for each of the levels. The fourth column is the value of the variance set, and the last two columns are the Type I error

rates with respect to the *AG* test, the t-test (for two groups), or *ANOVA* (more than two groups).

Results of the Simulation Study

Based on the results in Table 2, the Type I error rate for the study on the main effects can be seen in the case of two groups (for two levels of factor one) and three groups case (for three levels of factor two). All the bolded values are the Type I error rates that lie within 0.025 to 0.075, indicating the test can control the error. For the case of two groups, regardless of whether the sample sizes are balanced or not, as expected, the *t*-test provides the best control of the Type I error rate under the homogeneous variance. At the same time, the Type I error rates of the *AG* test are also considerable. However, when the variances are not equal, under all conditions, the *t*-test fails to control the Type I error rate with a value greater than 0.075. Meanwhile, the Type I error rates of the *AG* test are still within the values 0.025 and 0.075 under all conditions. In testing the main effect of the three groups' conditions, the *ANOVA* can only control the Type I error rates if the variances are homogeneous. Hence, the *AG* test has better control of Type I error rates in all conditions.

The effect of interaction between the two factors was simulated by combining the levels of each factor, which made six combinations altogether. Under this condition, the performance of the *ANOVA* was slightly different compared with the previous conditions. Under the conditions of balanced sample size, the *ANOVA* is only able to control the Type I error rates when variances are homogeneous. However, when the sample sizes are unbalanced, the Type I error can still be controlled regardless of the homogeneity of variances if the nature of the pairings is positive. Meanwhile, the *AG* test has better control of the Type I error rates in all conditions.

Table 2

Data Simulation Results

The effect factor	Factors	Sample sizes	Variance	Type I Error Rate	
				AG test	*t*-test/*ANOVA*
Main effect	Factor 1 (level 1, level 2)	(3, 3)	1,1	**0.0312**	**0.0506**
			1,36	**0.0550**	0.0924
		(5,5)	1,1	**0.0406**	**0.0512**
			1,36	**0.0528**	0.0800
		(5, 25)	1,1	**0.0554**	**0.0528**
			1,36	**0.0514**	0.0006
			36,1	**0.0498**	0.404
	Factor 2 (level 1, level 2, level 3)	(3,3,3)	1,1,1	**0.0340**	**0.0530**
			1,1,36	**0.0404**	0.1316
		(5, 5, 5)	1,1,1	**0.0502**	**0.0536**
			1,1,36	**0.0498**	0.1134
		(5, 15, 30)	1,1,1	**0.0544**	**0.0494**
			1,1,36	**0.0536**	0.0048
			36,1,1	**0.0466**	0.4118
			1,9,36	**0.0448**	0.0088
			36,9,1	**0.0500**	0.3438
Interaction effect	Factor 1 and Factor 2 (6 treatments)	(5,5,5,5,5,5)	1,1,1,1,1,1	**0.0546**	**0.0472**
			1,1,1,1,1,36	**0.0542**	0.1546
			1,4,4,16,16,36	**0.0586**	0.0866
		(5, 15, 15, 25, 25, 30)	1,1,1,1,1,1	**0.0548**	**0.0526**
			1,1,1,1,1,36	**0.0562**	**0.0478**
			1,4,4,16,16,36	**0.0556**	**0.0272**
			36,1,1,1,1,1	**0.0528**	0.4356
			36,16,16,4,4,1	**0.0554**	0.2778

## Real-Data: Validation of the *AG* Test

To validate the *AG* test, real data, as displayed in Table 1, was used to calculate the *p*-value and make a comparison with the *ANOVA*.

Descriptive Analysis

From the experiment to investigate the effect of materials and temperature on the battery's life, *Figure 1* summarizes the observed data of the battery's life in a boxplot.

Figure 1

Boxplot of the Battery's Life Based on Different Materials
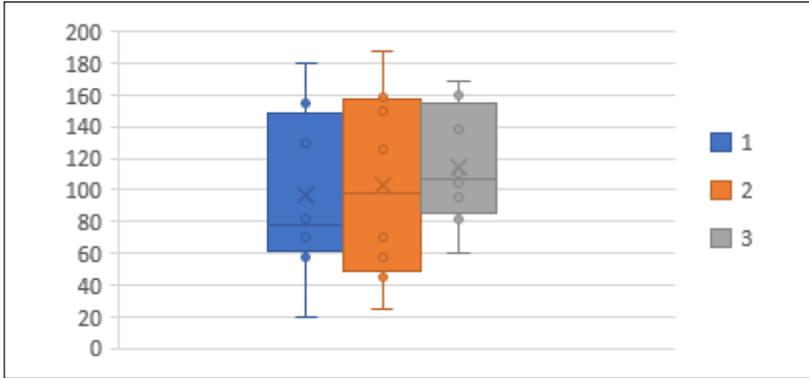
From *Figure 1*, the effect of different materials (material types 1, 2, and 3) did not produce much difference in the lifetime of the batteries where the boxplots are all at the same height.

Figure 2

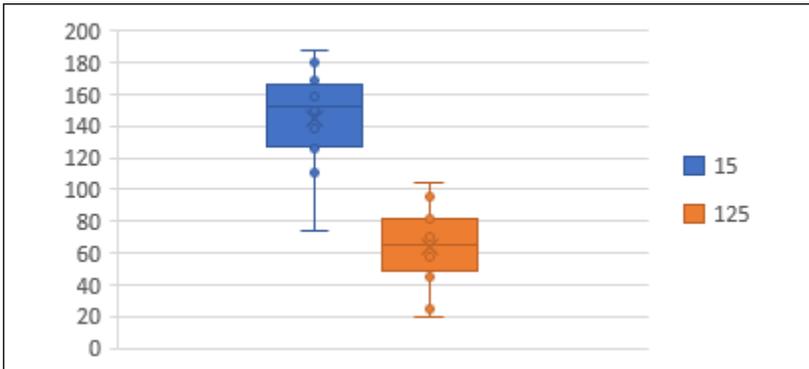*Boxplot of the Battery's Life Based on Different Temperatures*


Figure 2 displays the battery life based on different temperatures (15°F and 125°F), showing that the battery life at 15°F is higher than at 125°F. The boxplots of the battery life for both temperatures also display balance whiskers representing normal distribution.

To see if any interaction between the two factors occurs, *Figure 3* illustrates the life of the batteries respective to the types of materials

and temperatures. To understand the interaction, the failure of one factor to produce the same effect on the response at different levels of another factor can be seen. Based on *Figure 3*, the effect of different types of materials on the life of batteries is the same as the effect of temperature at different levels. Hence, it is believed that no interaction occurs between the types of materials and temperatures.

Figure 3

Plot of the Battery's Life Based on Different Materials and Temperatures
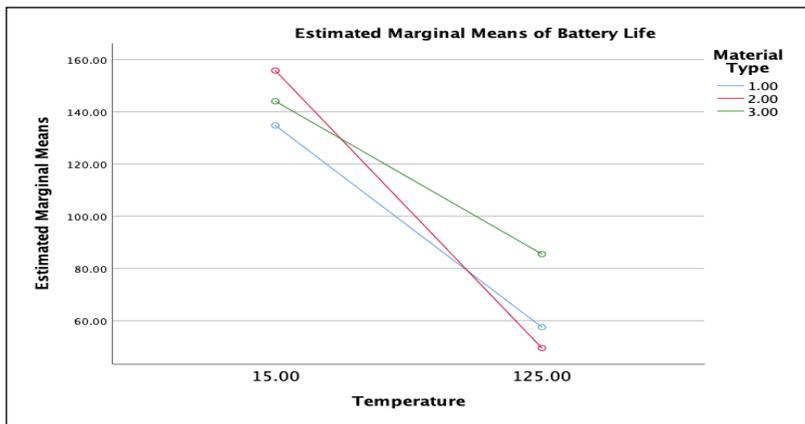
Testing the Assumptions: Variances Homogeneity and Normality

Before performing the *ANOVA* test, one needs to check the assumptions. The homogeneity of variances was tested using Levene's test and Breusch-Pagan. While the normality was checked using the Shapiro-Wilk test. Table 3 illustrates the result of the test. Both tests, the Levene test and the Breusch-Pagan test, show insignificant p-values, indicating all variances are homogeneous. For testing normality, Table 4 shows the result of the Shapiro-Wilks test with the p-value of 0.557, which is sufficiently large to indicate that the hypothesis of normal distribution cannot be rejected.

Table 3

Test of the Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
Battery Life	Based on Mean	.765	5	18	.587
	Based on Median	.670	5	18	.651
	Based on Median and with adjusted df	.670	5	11.413	.654
	Based on trimmed mean	.759	5	18	.591

Homogeneity Tests	P-value	Significant
Modified Breusch-Pagan Test	0.281	No
Breusch-Pagan Test	0.301	No
F Test	0.301	No

Table 4

Test of Normality

Normality Tests	P-value	Results
Shapiro-Wilk	0.557	Do not reject the null hypothesis

The Effect of Different Types of Materials and Temperature on the Life of the Batteries

Table 5 shows the results of the *AG* test and the two-way *ANOVA* for the effect of the types of materials and temperatures on the life of batteries. The results indicate almost the same results produced by the *AG* and *ANOVA*. For the main effect of materials, both tests indicated no significant effect of the different types of materials on the life of the batteries. However, when testing the temperature, the results between both tests were slightly different. *Figure 2* clearly shows that lower temperatures produced more life for the batteries. Based on Table 5, the *ANOVA* test produced a significant result with a very small *p*-value, with the significance level being up to 0.01 compared to the *AG* test, which was only significant at 0.1. It needs to be reminded that based on Table 2, the simulation results of the *AG* test show conservative Type I error rates (lower than 0.05) but still acceptable (within 0.025 to 0.075) under conditions of two groups, balanced sample sizes with

homogeneous variance. From these results, it can be said that the *AG* test is valid for use.

Table 5

Two-way ANOVA

The effect	*AG* test p-value	*ANOVA* p-value
The main effect of the material type	0.951	0.4303
The main effect of the temperature	0.0639	0.000002
Interaction	0.0907	0.2649

CONCLUSION

This study aimed to identify the *AG* test's performance when implemented into the analysis of factorial design, where this study focused on the two factors of factorial design. The simulation study showed that the *AG* test outperformed the *ANOVA* for better controlling the Type I error rates when variances are heterogeneous regardless of the number of groups and sample sizes. The performance of the *AG* test under homogeneous variances can be considered good. Nevertheless, *ANOVA* still provides the best. In real data analysis, the *AG* test is able to investigate the effects of the types of material and temperatures on the life of the batteries. It is the alternative to the *ANOVA*, especially when variances are heterogeneous.

ACKNOWLEDGMENT

REFERENCES

Abdullah, S., Yahaya, S. S. S., & Othman, A. R. (2011). Modified alexander-govern test as alternative to t-test and ANOVA F test. *Sains Malaysiana*, *40*(10), 1187–1192. https://www.ukm.my/jsm/english_journals/vol40num10_2011/vol40num10_2011pg1187-1191.html

Abdullah, S., Yahaya, S. S. S., & Yusof, Z. M. (2014). Testing the equality of students' performance using Alexander-Govern test with adaptive trimmed means. *AIP Conference Proceedings*, *1602*, 1157–1160. https://doi.org/10.1063/1.4882630

Alexander, R. A., & Govern, D. M. (1994). A new and simpler approximation for ANOVA under variance heterogeneity. *Journal of Educational Statistics*, *19*(2), 91–101. https://doi.org/10.3102/10769986019002091

Allen, M. (2017). Factorial analysis of variance. In *The SAGE Encyclopedia of Communication Research Methods*. SAGE Publications, Inc. https://doi.org/10.4135/9781483381411.n192

Antony, J. (2014). Full factorial designs. In *Design of Experiments for Engineers and Scientists* (pp. 63–85). Elsevier. https://doi.org/10.1016/B978-0-08-099417-8.00006-7

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

Crawley, M. (2012). Analysis of variance. In *The R Book* (pp. 498–536). Wiley. https://doi.org/10.1002/9781118448908.ch11

Hodžić, D., & Islamović, F. (2020). Factorial ANOVA experimental design in R. *Proceedings of the 9th International Scientific Conference of Defensive Technologies*, *October 2020*. https://www.researchgate.net/publication/344597074

Jamaluddin, F. B. (2015). *Alexander-govern test using winsorized means*.https://www.scoutsecuador.org/site/sites/default/files/%5Bbiblioteca%5D/5.1 Conservacion de alimentos y Recetas sencillas.pdf%0Ahttp://publications.lib.chalmers.se/records/fulltext/245180/245180.pdf%0Ahttps://hdl.handle.net/20.500.12380/245180%0Ahttp://dx

Mendeş, M., & Yiğit, S. (2013). Type I error and test power of different tests for testing interaction effects in factorial experiments. *Statistica Neerlandica*, *67*(1), 1–26. https://doi.org/10.1111/j.1467-9574.2012.00528.x

Montgomery, D. C. (2020). *Design and analysis of experiments*. John wiley and sons. https://shorturl.at/glQ79

Nguyen, D., Kim, E., Wang, Y., Pham, T. V., Chen, Y.-H., & Kromrey, J. D. (2019). Empirical comparison of tests for one-factor ANOVA under heterogeneity and non-normality: A Monte Carlo study. *Journal of Modern Applied Statistical Methods*, *18*(2), 2–30. https://doi.org/10.22237/jmasm/1604190000

Ochuko, T. K., Abdullah, S., Zain, Z., & Syed Yahaya, S. S. (2015a). Modifying and evaluating the Alexander-Govern test using real data. *Modern Applied Science*, *9*(12), 1–11. https://repo.uum.edu.my/id/eprint/17619

Ochuko, T. K., Abdullah, S., Zain, Z., & Syed Yahaya, S. S. (2015b). The modification and evaluation of the Alexander-Govern test in terms of power. *Modern Applied Science*, *9*(13), 1–21. https://repo.uum.edu.my/id/eprint/17618

Ochuko, T. K., Abdullah, S., Zain, Z., & Syed Yahaya, S. S. (2015c). Winsorized modified one step m-estimator in Alexander-Govern test. *Modern Applied Science*, *9*(10), 51–67. https://repo.uum.edu.my/id/eprint/17623

Schneider, P. J., & Penfield, D. A. (1997). Alexander and govern's approximation: Providing an alternative to anova under variance heterogeneity. *Journal of Experimental Education*, *65*(3), 271–286. https://doi.org/10.1080/00220973.1997.9943459

Wilcox, R. R. (2003). One-way anova. In *Applying Contemporary Statistical Techniques* (pp. 285–328). Elsevier. https://doi.org/10.1016/b978-012751541-0/50030-4