Transactions of the Nigerian Association of Mathematical Physics
***Volume 11, (January - June, 2020 Issue*), pp69 – 74**
© Trans. of NAMP

ON THE STUDY OF OUTLIERS IN DESIGN EXPERIMENT

Odior K.A. and Ekerikevwe K.I.

Department of Statistics, Delta-State Polytechnic, Otefe- Oghara, Delta- State.

Abstract

The analysis of variance (ANOVA) procedure is one of the most referred, flexible and practical statistical tool for comparing several populations means in any designed experiment. The inefficiency of this procedure in design experiment when outliers occurred is a major problem to practitioners. This paper considers outliers in a designed experiment using different experimental design: Completely Randomized Design (CRD), Randomized Complete Block Design (RCBD) and the Latin Square Design (LSD). The outlying observations were identified using the cook's distance procedure. Our result clearly revealed a substantial distortion in the estimated ANOVA quantities such as Mean Square Error (MSE), F Statistic and the P-Value. Consequently, the statistical inference regarding the significance of the treatment effects was altered due to the presence of in the experiment resulting in misleading conclusion.

***Keywords*:** Outlier, Design Experiment, Completely Randomized Design, Randomized Complete Block Design, Treatment, Analysis of Variance, Statistical Inference

1. Introduction

In Design and Analysis of Experiment (DOE) the objective is to discover and evaluate something about a particular process or system. It involves a purposefully changes in the controlled variables (input) so that we can observe, monitor and identify the reasons for the changes in the output (response) variables. In any experiment, the results and conclusions that can be drawn depends to a large extent on the manner in which the data were collected[1]. A frequent obstacle that is common in every field involving data collection is the presence of outliers. When outliers are present in the data, the whole set up of the experiment is disturbed [2]. Consequently, proper analysis of the experimental data become more complicated and perhaps misleading interpretation becomes inevitable. Therefore, the phenomenon called outlier is a common problem in design and analysis of experiment. In order to reach an informed and valid deductions about the experiment, a statistical tool of the analysis of variance (ANOVA) are generally employed. ANOVA is a statistical tool devised for the analysis of experimental data which over the years has proved to be useful and informative. In order for the statistical inferences to be valid, the observed variables must conform to assumptions that underlie the statistical procedures to be used. All statistical methods rely explicitly or implicitly on a number of assumptions [3]. These assumptions have been present in statistical studies for long time and have been the framework for all classical methods in regression analysis, analysis of variance and multivariate analysis [4].

One common feature of many real-life statistical data both observational and in designed experiments is that they contain observations which are inconsistent and significantly far away from the remaining data set [5-7]. These observations are technically called outliers. In designed and analysis of experiment outliers are observations (data points) that control the significance of the treatments effects. Thus, conclusion drawn from data contaminated with outliers lack inferential validity. Besides, the outlying observations in designed, experiment and perhaps is ANOVA method result in false and misleading statistical quantities and inference. The presence of an outlier is often an indication of weakness in the statistical model, the data or both.

However, the statistical methodology of ANOVA framework is powerful and resilient under classical assumptions. In reality these assumptions are hardly met, due to the presence of outlying observations. ANOVA statistical procedure is strongly affected by the presence of outliers since its analysis is based on sample group means and variance which are not statistically robust [5].

The classical analysis of variance (ANOVA) statistical tool is one of the most widely used method of data analysis especially in designed and analysis of experiment where comparison of treatment effects is desired Bird [8]. ANOVA has increasingly become a household name and cornerstone in revealing the source of variability in data set into several components (sources) [3]. Therefore, is unarguably a collection of statistical models and their associated estimation procedure use to analyze the difference among group means in a sample in order to make statistical inference about the population mean.

Correspondence Author: Odior K.A., Email: odifullness@gmail.com, Tel: +2348034663466, +2348064647455 (EKI)

ANOVA is a statistical technique for partitioning the total variation of a data set into several components [9]. In practice, ANOVA method enables researchers to ascertain the proportion of total variation attributable to each source of variation in the data set. The applicability of ANOVA in testing of statistical hypothesis concerning population means is well established in statistical literature. The popularity and usefulness of this statistical tool is hinged on its capacity to separate the total variability found within the data set into several components or sources [10]. It may seem odd that the name of the technique is called analysis of variance rather than analysis of mean. The name ANOVA is appropriate because inferences about population means are made by analyzing variance. The ANOVA F test evaluates and accesses whether the group means on the dependent variable differ significantly from each other in a designed experiment.

The data generated from designed experiments are analyzed under certain classical assumptions. Empirical evidences involving real-life data extracted from review of several scientific journals indicates that these assumptions are not always met [11]. Besides statistical literatures revealed that data generated from previous experiments conducted in different parts of the world suffers from the problem of non-normality and heterogeneous error distribution of variances [10].

The statistical methodology of ANOVA framework is efficient when the necessary classical assumptions are sufficiently fulfilled. Thus, in the presence of outliers in a designed experimental data, ANOVA become inefficient since its analysis is based on sample group means and variance which is not statistically robust [5]. Analysis of means has a poor resistance to outliers.

2. Problem of Outliers in Designed and Analysis of Experiment Using ANOVA

The ANOVA is one of the most referred classical statistical methods for decades for testing the hypothesis that K population means are equal, where $K \geq 2$. It maintains its efficient efficiency and optimality when underlying assumptions are sufficiently satisfied. It is a powerful statistical method that split the total variation of a data set into constituent parts and measures the different sources of variation.

In this communication, we considered the simplest form of ANOVA, the one-way ANOVA. The statistical model suitable for the type of design we adopted and express the responses as a linear function of the grand mean, the treatment effects and the error term as follows:

$$Y_{ij} = \mu + \alpha_i + e_{ij} \quad i = 1,2,...,k, \tag{1}$$

$$j = 1,2,....,n$$

The classical assumptions for this model are

$$E(e_{ij}) = 0, \tag{2}$$

$$Var(\square_{ij}) = \sigma^2 < \infty; \text{for all i, j.} \tag{3}$$

$e_{ij}$are assumed to independent and normally distributed with mean zero and constant variance,

$$\sigma^2 \left[i.e., e_{ij} \approx NID(O, \sigma^2) \right] \tag{4}$$

Given the ANOVA assumptions, if the null hypothesis is true, if it follows that the Variance Ratio Statistics (VRS) is distributed as

$$F = \frac{Between\ group\ variability}{Within\ group\ variability} \tag{5}$$

Thus, the ANOVA statistics is defined as $F = \frac{DB/(m-1)}{DW/(n-m)}$ with degree of freedom (m-1)(n-m)

Where DB is the deviation between groups and DW is the deviation within groups. Due to the presence of the sample in bothDW and DB, the value of the F statistic is strongly influenced by the presence of outliers in the designed experiments.

3. Methodology

ANOVA is one of the most flexible and practical techniques for comparing several population means in any designed experiment. Considering the data outlay for a completely randomized design as given below

Treatment

1	y_{11}	y_{12}	…	y_{1n}	y_1
2	y_{21}	y_{22}	…	y_{2n}	y_2
.	.	.	…	…	…
.	.	.	…	…	…
.	.	.	…	…	…
t	y_{t1}	y_{t2}	…	$y_{t_{nt}}$	y_t

The treatment means differ by an amount α_i the treatment effect. Thus, a test of hypothesis can be established as

H_o: $\mu_1 = \mu_2 = … = \mu_i$

Versus

H_1: Not all the μ_i's are equal.

Our test statistics can be derived using the concept of the partitioning of the total sum of squares (TSS) of the measurements about their mean. The total sum of squares is partitioned into two separate sources of variability: one due to variability among treatments (between treatments) and one due to the variability within each treatment (error which accounts for the variability that is not explained by treatment differences.

The partition of TSS can be defined as follows:

$$\sum_{ij}\left(y_{ij}-\bar{y}_i\right)^2 = \sum_i ni(y_i-\bar{y}_i)^2 + \sum_{ij}\left(y_{ij}-\bar{y}_i\right)^2 - 2 \qquad (6)$$

When the number of replications is the same for all treatments, the partition assumes

$$\sum_{ij}\left(y_{ij}-\bar{y}_i\right)^2 = n\sum_i(y_i-\bar{y}_i)^2 + \sum_{ij}\left(y_{ij}-\bar{y}_i\right)^2 - 2 \qquad (7)$$

The above information can be summarized in an ANOVA table for a completely randomized design.

Table 1: Layout of one-way ANOVA

Source	SS	DF	MS	F
Treatment	SST	M-1	MST = SST/(M-1)	MST/MSE
Error	SSE	N-M	MSE = SSE/N-M	
Total	TSS	N-1		

In a designed experiment were outliers occurred, F statistic is affected thus leading to a wrong statistical inference.

The cook's distance is a distance measure, which indicates the inference of its data point on the estimation of parameter vector. The distance measure can be expressed in general form as

$$CD_i^2 = \frac{(\beta_{(i)}-\hat{\beta})'(X'X)(\hat{\beta}_{(i)}-\beta)'}{P} \qquad (8)$$

4. Data Analysis and Result

In considering outliers in a designed experiment the following designed experimental model were examined.

4.1 One-Way Design

An experiment on the effect of three feeds: Dead organism, homemade fish meal and industrial made on the increase in weight of the fishes. The fish were kept in three tanks. The weight (kg) gained was observed in every five days for 30 days.

Table 2: The original data

Replication	Tank 1	Tank 2	Tank 3
	Dead organisms	Homemade fish meal	Industrial made
1	0.2	0.1	0.3
2	0.5	0.3	0.5
3	0.6	0.7	0.7
4	0.9	1.0	0.8
5	1.1	1.2	1.2
6	1.4	1.3	1.5

Below is the analysis of variance with the original data

Table 3: ANOVA table for the original data

Tests of Between-Subjects Effects

Dependent Variable: OBSERVATION

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	11.375[a]	3	3.792	18.142	.000
TREATMENT	11.375	3	3.792	18.142	.000
Error	3.135	15	.209		
Total	14.510	18			

a. R Squared = .784 (Adjusted R Squared = .741)

Table 4: Contamination with one outlier

Replication	Tank one	Tank 2	Tank 3
	Dead organisms	Homemade fish meal	Industrial made
1	0.2	0.1	0.3
2	0.5	0.3	0.5
3	0.6	0.7	0.7
4	0.9	1.0	2.0
5	1.1	1.2	2.5
6	1.4	1.3	19.2

Table 5: Cooks distance statistic

OBSERVATION	COOKS D
1	0.00
2	0.00
3	0.00
4	0.00
5	0.00
6	0.00
7	0.00
8	0.00
9	0.00
10	0.00
11	0.00
12	0.00
13	0.07
14	0.06
15	0.05
16	0.02
17	0.01
18	**0.98**

Table 6: ANOVA table with one outlier

Tests of Between-Subjects Effects

Dependent Variable: OBSERVATION

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	113.048[a]	3	37.683	2.048	.150
TREATMENT	113.048	3	37.683	2.048	.150
Error	276.022	15	18.401		
Total	389.070	18			

a. R Squared = .291 (Adjusted R Squared = .149)

From table 6, we observed a remarkable effect of a single outlier in the dataset, resulting in an increase in the MSE. This was accompanied with the change in the statistical inference regarding the significance of the treatment effects without the outlier. The treatment effects as well as the replication effects were significant at 5% level of significance while with a single outlier, the same became non-significant.

4.2 Two-Way Design

Table 7: The original data

Treatment	Number of fishes		
	Two catfishes	Four catfishes	Three Six catfishes
Dead organisms – tank one	1.43	1.48	1.61
Homemade fish meal- tank two	1.39	1.43	1.53
Industrial made – tank	1.26	1.33	1.46

Table 8: ANOVA table for the original data

Tests of Between-Subjects Effects

Dependent Variable: OBSERVATION

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	18.632[a]	5	3.726	15244.709	.000
Type of feed used for treatment	.047	2	.024	96.727	.000
Number of fishes in the tank	.038	2	.019	77.227	.001
Error	.001	4	.000		
Total	18.633	9			

a. R Squared = 1.000 (Adjusted R Squared = 1.000)

Table 9: Data contamination with one outlier

Treatment	Number of fishes		
	Two catfishes	Four catfishes	Three Six catfishes
Dead organisms – tank one	1.43	1.48	1.61
Homemade fish meal- tank two	1.39	3.43	1.53
Industrial made – tank	0.10	1.33	20.46

Table 10: ANOVA table with one outlier

Tests of Between-Subjects Effects

Dependent Variable: OBSERVATION

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Model	328.892[a]	5	65.778	.930	.543
Type of feed used for treatment	73.250	2	36.625	.518	.631
Number of fishes in the tank	52.484	2	26.242	.371	.712
Error	282.964	4	70.741		
Total	611.856	9			

a. R Squared = .538 (Adjusted R Squared = -.041)

Table 11: Cooks distance statistic

OBSERVATION	COOKS D
1	0.11
2	0.03
3	0.09
4	0.02
5	0.43
6	0.35
7	0.07
8	0.39
9	**0.80**

It is evident from table 10 that the introduction of the outliers, the MSE significantly increase with a corresponding decrease in the F statistic, an indication of non-significance. However, this is a direct opposite of the result in table 8.

4.3 Latin Square Design

Table 12: Original dataset

Treatment	Number of fishes		
	Two catfishes	Four catfishes	Three Six catfishes
Dead organisms – tank one	1.34α	1.47γ	1.53β
Homemade fish meal- tank two	1.46β	1.29α	1.39γ
Industrial made – tank	1.30γ	1.39β	1.59α

Table 13: ANOVA for original data

Tests of Between-Subjects Effects

Dependent Variable: OBSERVATION

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	18.140[a]	7	2.591	145.767	.007
Type of feed used for treatment	.033	2	.017	.938	.516
Number of fishes in the tank	.007	2	.004	.197	.835
Interval of days for water changing in tank	.009	2	.004	.243	.805
Error	.036	2	.018		
Total	18.175	9			

a. R Squared = .998 (Adjusted R Squared = .991)

Table 14: Two outliers contamination

Treatment	Number of fishes		
	Two catfishes	Four catfishes	Three Six catfishes
Dead organisms – tank one	1.34α	1.47γ	1.53β
Homemade fish meal- tank two	0.0009β	1.29α	1.46γ
Industrial made – tank	1.30γ	1.39β	3.59α

Table 15: Cooks statistic

OBSERVATION	COOKS D
1	0.16
2	**0.88**
3	0.29
4	0.17
5	0.16
6	0.02
7	0.63
8	**0.39**
9	0.31

Table 16: ANOVA with two outliers

Tests of Between-Subjects Effects

Dependent Variable: OBSERVATION

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Type of feed used for treatment	64.186	2	32.093	1.151	.465
Number of fishes in the tank	57.728	2	28.864	1.036	.491
Interval of days for water changing in tank	58.057	2	29.029	1.041	.490
Error	55.748	2	27.874		
Total	331.563	9			

a. R Squared = .832 (Adjusted R Squared = .243)

Comparing table 13 and 14, using the computed values of MSE and F statistics, it is very clear that the introduction of outliers completely affected the result leading a false statistical inference.

5. Conclusion

In this article, outliers in a designed experiment using three different experimental designs were examined for two scenarios: experimental data without outliers and with outliers. The result suggests empirically that outliers have a great effect in design experiment using the ANOVA procedure. This is evident by variance ratio statistic and the P values resulting in the non-significance of the treatment effects.

References

[1] Montgomery, D. O. (2010). Design and Analysis of Experiments. John Wiley and Sons, New York.

[2] Lalmohan, B. and Gupta, V. K. (2001). A Useful Statistic for Studying Outliers in Experimental Design.Sankhya, B63, Pg. 338 – 350

[3] Avi, G. (2006). Robust Analysis of Variance.Process Design and Quality Improvement. International Journal of Productivity and Quality Management, Vol. 1(3)

[4] Kulinskaya, E. and Michael, B. O. (2007). Robust Weighted One-Way ANOVA Improved Approximation and Efficiency. Journal of Statistical Planning and Inference, Vol. 2(4)

[5] Bruno, B. and Roberta, V. (2007). Robust Analysis of Variance: An Approach Based on the Forward Search. Computational Statistics and Data Analysis, Vol. 5(1), Pg. 5172 – 5183

[6] Staudtle, R. G. and Sheather, S. J. (1990). Robust Estimation and Testing.Wiley, New York.

[7] Barnett, V. and Lewis, T. (1993). Outliers in Statistical Data, 3rd ed. Wiley, New York.

[8] Bird, K. D. (2004). Analysis of Variance Via Confidence Interval. Sage Publications Ltd, London

[9] Oyeka, C. A. (1992). Applied Statistical Methods in the Sciences.Nober Avocation Publishing Company, Enugu.

[10] Dinesh, I. R. and Padmini, V. P. (2015). Robust ANOVA: An Illustrative Study in Horticultural Crop Research. International Journal of Mathematics and Computational Sciences, Vol. 9(2).

[11] Blanca, M. J., Arnau, J. and Rebecca, B. (2017). Non-normal Data: Is ANOVA still a Valid Opition? Psicothema, Vol. 29, No. 4