Modification of ANOVA with Trimmed Mean

Nur Farah Najeeha Najdi1, Nor Aishah Ahad1
1School of Quantitative Sciences, Universiti Utara Malaysia (UUM)

Correspondence: Nur Farah Najeeha Najdi (nfarahnajeeha@gmail.com)

Abstract

Analysis of Variance (ANOVA) is a well-known method to test the equality of mean for two or more groups. ANOVA is a robust test under the normality assumption. Arithmetic mean is used in the computation of the ANOVA test. Mean is known to be sensitive towards outlier and this problem will affect the robustness and power of ANOVA. In this study, modification of ANOVA was created using one type of mean to replace arithmetic mean namely trimmed mean. New approaches were be obtained for the computation of ANOVA. This study was conducted based on a simulation study and application on real data. The performance of the modified ANOVA is then compared with the classical ANOVA test in terms of Type I error rate. This innovation enhances the ability of modified ANOVA to provide good control of Type I error rates. The findings were in favor of the modified ANOVA or better known as ANOVA™.

Keywords: anova, modified anova, arithmetic mean, trimmed mean, type I error

Introduction

Analysis of variance (ANOVA) is one of the most frequently used methods in statistics (Moder, 2016). It allows us to compare more than two group means in a continuous response variable. An ANOVA model assumes that:

i. The probability distribution of responses in each group is normal.
ii. Each probability distribution has the same variance.
iii. Samples are independent.

Note that with the normality of the probability distributions and the constant variability, the probability distributions differ only with respect to their means. ANOVA is a very powerful test as long as all prerequisites are met. It is not necessary, nor is it usually possible, that an ANOVA model fit the data perfectly. ANOVA models are reasonably robust against certain types of departures from the model, such as the data not being exactly normally distributed (Kutner, Nachtsheim, Neter & Li, 2005).

Basically, classical parametric tests such as analysis of variance (ANOVA) and independent sample t-test are often used in testing the central tendency measure by researchers rather than other methods since the aforementioned methods provide a good control of Type I error and generally more powerful than other methods when all the assumptions are fulfilled (Wilcox & Keselman, 2010). Analysis of variance (ANOVA) has been stated to be robust to departures from population normality (Glass, Peckham & Sanders, 1972).
The calculation of ANOVA involves the mean. It is also known as the arithmetic mean. The simplest and most classical mean values are the arithmetic, the geometric and the harmonic mean. All of the means are known for their inequalities (Xia, Xu & Qi, 1999). The arithmetic mean is the most commonly used average. It is generally referred to as the average or simply mean. The breakdown point is zero. The arithmetic mean or simply mean is defined as the value obtained by dividing the sum of values by their number or quantity. The formula given below is the basic formula that forms the definition of arithmetic mean and is used in case of ungrouped data where weights are not involved. The formula is as follows:

\[ A = \frac{a_1+a_2+a_3+\cdots+a_n}{n} \] (1)

The classical point and interval estimators use the sample mean and standard deviation. If a graph of the data indicates that the classical assumptions are violated, then an alternative estimator should be considered. Robust estimators can be obtained by giving zero weight to some cases and applying classical methods to the remaining data. Bickel (1965) and (Stigler, 1973) consider trimmed means or sometimes called a truncated mean. Shorack (1974) and also Shorack and Wellner (1986) derive the asymptotic theory for a large class of robust procedures. Special cases include trimmed, Winsorized, metrically trimmed, and Huber type skipped means. Many location estimators can be presented in a unified way by ordering the values of the sample as \( x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_n \).

A trimmed mean is calculated by discarding a certain percentage of the lowest and the highest scores and then computing the mean of the remaining scores. The prevalent method of trimming is to remove outliers from each tail of the distribution of scores. Tukey (1948) introduced trimmed mean as a compromise to the two classical estimators mean and median, to achieve balance between outlier(s) tolerance and efficiency. In addition, the recommendation is to trim 20% from each tail (Rosenberger & Gasko, 1983; Wilcox, 1995). The \( x \)% trimmed mean has a breakdown point of \( x \)% for the chosen level of \( x \).

For example, a mean trimmed 50% is computed by discarding the lower and higher 25% of the scores and taking the mean of the remaining scores. The median is the mean trimmed 100% and the arithmetic mean is the mean trimmed 0%. Therefore, in this study, we will consider the robust method, trimmed mean along with the other means mentioned above. Wilcox (2012) states that trimming means cannot fix every problem but do work remarkably well to adjust for problems of heteroscedasticity (non-equal variances) and non-normality. The \( \alpha \)-trimmed mean is as follows:

\[ \tilde{x}_{ij} = \frac{1}{n^j-\alpha_{ij}} \left[ \sum_{l=\alpha_{ij}+1}^{n^j-\alpha_{ij}} x_{ij} \right] \] (2)

Thus, the trimmed mean correspond to the mean value of data samples where \( p \) highest and \( p \) lowest samples are removed. Application of trimming lowers the influence of extreme data values on the result of averaging.

Outliers can completely break down the results of the ANOVA test when not properly taken into account (Wilcox, 1990). Given this limitation of the ANOVA test, there is a need for ANOVA type tests that are robust. Such an approach using robust estimators provides better control of the probability of the Type I error for one-way ANOVA situations (Lix & Keselman, 1998).

**Problem Statement**

Calculation of the ANOVA has been using the classical mean or known as the arithmetic mean. It is very sensitive to changes in data series, especially to outliers. It has a breakdown point of 0%, so it shows that it is highly affected by extreme values or known as outliers. Therefore, it motivates us to perform this study by modifying the arithmetic mean with other type of mean in ANOVA.
A reliable estimation of the location of the bulk of the observations is needed. Therefore, the trimmed mean can help in providing a better estimation than the classical mean. Outliers and asymmetry less affect the standard error of the trimmed mean than the classical mean. ANOVA test using trimmed means can have more power than the test using the classical mean (Rousselet, Pernet & Wilcox, 2017). Therefore, in this study, we will compare the performance of ANOVA with the modified ANOVA in term of the Type I error rate.

**Objectives of the Study**

The followings are the objectives for this study:

1. To propose modification on classical ANOVA with trimmed mean.
2. To measure performance of the modified ANOVA in term of Type I error.
3. To compare performance between classical ANOVA with modified ANOVA.

**Methodology**

**Design Specification**

The Monte Carlo simulation study was performed using the SAS programming language. Pseudo-random number generators was invoked to obtain random variates from the normal distribution by using the SAS generator (SAS, 2006). Normal variates with mean, \( \mu = 0 \) and \( \sigma = 1 \) were generated. Nominal alpha was set at \( \alpha = 0.05 \). Figure 3.1 shows the flow of the simulation study processes. The modified ANOVA is then applied on real data.

![Figure 1: Flow chart of Simulation Study](image-url)
Each statistical method is assessed to compare three group means in a simulated dataset. We focus on three groups because it is quite common and can be found in a lot of trials. The sample size is to be manipulated in this simulation study such that small, medium and large sample size to test the robustness of the modified ANOVA. Since this study uses ANOVA to compare three group means, the size of each group will be given as (a, b, c). We use group sizes of (15, 15, 15), (30, 30, 30) and (50, 50, 50) to represent the small, medium and large sample size, respectively.

Type I error is the rejection of a true null hypothesis resulting in the faulty conclusion of statistically significant treatment effects. From the simulation, the number of count will increase by ‘1’ if the p-value is less than $\alpha=0.05$ in which we reject the $H_0$. Then, the total number of counts were divided by the total simulations which is 1,000. Type I error rate corresponding to each method was determined and compared.

**Robustness Criterion**

In each simulation scenario, a test is considered to be significant when a p-value is less than the nominal $\alpha=0.05$. The number of significant tests will be counted in simulated datasets in a scenario and the Type I error rate was calculated. The robustness of a method is determined by its ability in controlling the Type I error. Researchers have developed a few robustness criterions. Sullivan and D’Agostino in 2003 illustrated that a test which does not exceed 10% of the nominal significance level as robust. Guo and Luh (2000) interpreted that if a test’s empirical Type I error rate is not higher than 0.075, with a 5% significance level, then it is robust.

Meanwhile, this study adopted the robustness liberal criterion by Bradley. This criterion was chosen since it was widely used by many robust statistic researchers (e.g. (Keselman, Kowalchuk, Algina, Lix, & Wilcox, 2000); (Othman, et al., 2004)) to judge robustness. A test can be considered robust if its empirical rate of Type I error, $\alpha$, is within the interval $0.5\alpha$ and $1.5\alpha$ (Bradley, 1978). If the nominal level is $\alpha = 0.05$, the empirical Type I error rate should be between 0.025 and 0.075.

**ANOVA Test**

The total variability can be split into several parts. The total amount of variability among observations can be measured by summing the squares of the differences between each $x_{ij}$ and $\bar{x}$:

- **SST**: Total sum of squares
  \[
  SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 \tag{3}
  \]

The variability has two sources:

- Variability between group means (specifically, variation around the overall mean $\bar{x}$)
  \[
  SSA = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2, \tag{4}
  \]

- Variability within groups means (specifically, variation of observations about their group mean $\bar{x}_i$)
  \[
  SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^{k} (n_i - 1) s_i^2 \tag{5}
  \]

It is the case that,

\[
SST = SSA + SSE \tag{6}
\]

F-test is a measure of the variability between treatments divided by a measure of the variability within treatments. Table 3.1 shows the calculations for F-test for ANOVA.
Table 1: Table of ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model/ Group</td>
<td>SSA</td>
<td>k</td>
<td>( \frac{SSA}{k-1} )</td>
<td>( MSA )</td>
</tr>
<tr>
<td>Residual/ Error</td>
<td>SSE</td>
<td>n-k</td>
<td>( \frac{SSE}{n-k} )</td>
<td>MSE</td>
</tr>
<tr>
<td>Total</td>
<td>SST</td>
<td>n-1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Trimmed Mean**

This study will trim the mean of the data by 20% as have been suggested by Rosenberger and Gasko (1983) and Wilcox (1995; 2005) that 20% trimming should be used. The sample trimmed mean is computed as shown below. Let \( x_{1j}, x_{2j}, \ldots, x_{n_{ij}} \) be an ordered sample of group \( j \) with size \( n_j \).

Calculate the \( \alpha \)-trimmed mean of group \( j \) by using:

\[
\tilde{x}_{ij} = \frac{1}{h} \left\{ \sum_{i=g_1+1}^{n_j-g_2} x_{ij} \right\}
\]  \hspace{1cm} (7)

where

\[
h = n_j - g_1 - g_2
\]  \hspace{1cm} (8)

\[
g_1 = \lceil n_j \alpha_u \rceil
\]  \hspace{1cm} (9)

\[
g_2 = \lceil n_j \alpha_1 \rceil
\]  \hspace{1cm} (10)

**Modified ANOVA**

This paper focuses on modifying the ANOVA with a type of mean that is trimmed mean. Their performances will be observed using the Type I error rate. As have been mentioned before, there are three parts in calculating the F value for ANOVA, which are SST, SSA and SSE.

Recalling the formula are as follows,

\[
\text{SST} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2
\]

\[
\text{SSA} = \sum_{i=1}^{k} n_i (\tilde{x} - \bar{x})^2, \text{ and}
\]

\[
\text{SSE} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \tilde{x})^2 = \sum_{i=1}^{k} (n_i - 1)s_i^2
\]

\[
\text{SST} = \text{SSA} + \text{SSE}
\]

The \( \bar{x} \) will be replaced by the trimmed mean as shown in equation (7) to produce modified ANOVA presented by ANOVA\(_{YM}\).

**Data Analysis And Results**

**Simulation Results**

The robustness of a method is determined by its ability to control the Type I error. Table 2 displays the empirical Type I error rates for all the procedures. Based on Bradley’s liberal criterion of robustness (Bradley, 1978), a test can be considered robust if the rate of Type I error, \( \alpha \) is within the interval 0.5\( \alpha \) and 1.5\( \alpha \). For the nominal level of \( \alpha = 0.05 \), the Type I error rates should be between 0.025 and
0.075. The best procedure is the one that can produce Type I error rate closest to the nominal (significance) level.

Table 2: Type I error rates

<table>
<thead>
<tr>
<th>Methods</th>
<th>N=45(15,15,15)</th>
<th>N=90(30,30,30)</th>
<th>N=150(50,50,50)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>0.0213</td>
<td>0.0351</td>
<td>0.0323</td>
</tr>
<tr>
<td>ANOVA_{TM}</td>
<td>0.0366</td>
<td>0.0372</td>
<td>0.0403</td>
</tr>
</tbody>
</table>

Both of the methods showed robust Type I error rates. The best procedure is ANOVA_{TM} or modified ANOVA with trimmed mean which produces the nearest Type I error rate to the nominal level which is α = 0.05.

**Analysis on Real Data**

The performance of the ANOVA and also the modified ANOVA were demonstrated on real data. The data used in this study was the Index of African Governance. The Index of African Governance was a project of Harvard University's Kennedy School of Government's Program on Intrastate Conflict and Conflict Resolution and of the World Peace Foundation. This data set includes data on human development; participation and resource human; and safety and security of every Africa country.

There are three groups of them. The sample sizes for Group 1, 2 and 3 were 85, 53 and 32 respectively. The Shapiro-Wilk test has been employed in order to determine the normality of data analysis.

Table 3: Tests of Normality for BMI of Group 1

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Shapiro-Wilk</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human development (Group 1)</td>
<td>.872</td>
<td>85</td>
</tr>
<tr>
<td>a. Lilliefors Significance Correction</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Tests of Normality for BMI of Group 2

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Shapiro-Wilk</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human development (Group 2)</td>
<td>.921</td>
<td>53</td>
</tr>
<tr>
<td>a. Lilliefors Significance Correction</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Tests of Normality for BMI of Group 3

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Shapiro-Wilk</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human development (Group 3)</td>
<td>.870</td>
<td>32</td>
</tr>
<tr>
<td>a. Lilliefors Significance Correction</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Hypothesis:

H_0: The data is normally distributed.
H_1: The data is not normally distributed.
α = 0.05
Based on the Shapiro-Wilk test as shown in Table 3, 4 and 5, all of the groups are normally distributed since the p-value is greater than $\alpha = 0.05$, which failed to reject the null hypothesis.

**Table 6: Descriptive statistics for each group**

<table>
<thead>
<tr>
<th>Group</th>
<th>Sample size (N)</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>85</td>
<td>64.5</td>
<td>2.74</td>
<td>0.81</td>
<td>31</td>
<td>91</td>
</tr>
<tr>
<td>2</td>
<td>53</td>
<td>62.9</td>
<td>2.93</td>
<td>0.48</td>
<td>30</td>
<td>82</td>
</tr>
<tr>
<td>3</td>
<td>32</td>
<td>74.0</td>
<td>3.75</td>
<td>0.47</td>
<td>31</td>
<td>88</td>
</tr>
</tbody>
</table>

There are a total of 85 students from three different groups. In this data, the mean of human development score of each group is 64.5, 62.9 and 74.0 respectively. They are quite the same in terms of average, thus showing that the respondents have about the same opinions thus same score regardless of how they are being grouped. Furthermore, Group 1 has the lowest variation with value of 2.74 as compared to the other two groups.

**Table 7: Results of the test using different methods**

<table>
<thead>
<tr>
<th>Methods</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>0.0317</td>
</tr>
<tr>
<td>ANOVA&lt;sub&gt;TM&lt;/sub&gt;</td>
<td>0.0221</td>
</tr>
</tbody>
</table>

The results of the test in the form of p-values are given in Table 7. For comparison, the data were tested using all the five procedures mentioned in this study namely ANOVA and the modified ANOVA, which is ANOVA<sub>TM</sub>. As can be observed in Table 7, both of the methods show significant results in which they reject the null hypothesis. The ANOVA<sub>TM</sub> shows a better detection with the strongest significance (p = 0.0221) as compared to the other method. As shown in the simulation results in Table 2, the ANOVA<sub>TM</sub> does produce robust Type I error rates. Even though ANOVA<sub>TM</sub> shows stronger significance (p = 0.0221) as compared to the classical ANOVA, but ANOVA<sub>TM</sub> in general only gave a brief information on the data since the data has been trimmed by 20%. Thus, misrepresentation of the result could occur.

**Conclusions**

The goals of this paper are to propose modification on classical ANOVA and also to measure and compare the performances in term of Type I error. The Type I error will increase when the method is less robust compared to other methods. This will cause wary rejection of the null hypothesis and the power of the test can be reduced. Undetected differences are one of the consequences. This study has integrated the ANOVA with other type of mean, which is trimmed mean.

This paper has shown some improvement in the statistical solution for detecting differences between location parameters. The findings showed that the modified robust procedure, ANOVA<sub>TM</sub> is comparable with the classical ANOVA in controlling Type I error rates. In the analysis on real data, ANOVA<sub>TM</sub> (p = 0.0221) showed a slightly stronger significance than the classical ANOVA (p = 0.0317).

**Limitation of the Study**

Even though the study has achieved its goals, some constraints were inevitable. Study limitations would be time restrictions since only 12 weeks are part of one semester. This implies that there is very little time to analyze more information.
Recommendation of the Study

To improve the performance of the modified ANOVA methods, other types of mean or robust scale estimators should be considered to be replaced in the formula. There are plenty of them that can be chosen from. For example, generalized means or power mean. It is an abstraction of the quadratic, arithmetic, geometric and harmonic means. Other example of type of mean that can be used is weighted arithmetic mean. Furthermore, it is recommended that in collecting data for any experiment, efforts should be made to always have large set of data to enable the central limit theorem to come into play should the normality assumption be violated. If the researcher want to use trimmed mean, then it is better to use a larger size of data sets. Other than that, type of data should be taken into account. This study has used growth data, therefore further research should use other data such as financial and economical data.

References


