The Robust Test Statistic in Comparing Two Independent Groups using Trimming and Winsorization

Suhaida Abdullah, Sharipah Soaad Syed Yahaya and Zahayu Md. Yusof
School of Quantitative Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia

***Abstract*-- The classical independent t-test is often in jeopardy when the assumptions of normality or homogeneity of variances being violated. The test performs worsen when these violations occur simultaneously. Alexander-Govern test offers the alternative solution to the classical t-test when dealing with heterogeneous variances conditions. However, it produces good control of Type I error rates only if the data are normally distributed, which is a known fact that normality is hardly achieved in real life situation. As a remedy, in this study, we modify the Alexander-Govern test using trimmed mean and Winsorized mean as the location measures. Generally, the modified test using trimmed mean performs better compared to the original test in terms of Type I error rates. However, the test using Winsorized mean failed to control the Type I error rate well under most condition considered.**

***Index Term*-- T-test; Alexander-Govern test; trimmed mean; Winsorized mean**

1. Introduction and Background of Study

Statistical methods are powerful in extracting information from data. However, choosing incorrect statistical tests would mislead the information and affecting the conclusion. This will lead to a more serious problem where the information might jeopardize any decision-making process.

The *t*-test and ANOVA are well-known as the most commonly used statistical methods when comparing two or more independent groups. However, some researchers sometimes are unaware of the assumptions needed to these methods. The methods are adversely affected by non-normality, particularly when variances are heterogeneous and group sizes are unequal [1].

Those who are mindful of the problems might choose to use nonparametric methods. It is a good alternative where these methods need to fulfill fewer assumptions. As a free distribution approach, the practitioners can use it without need any assumption on the data distribution. Instead of using the original observation, most of the nonparametric methods use rank values which make this approach free from the effect of an outlier. However, using the rank values produce a less powerful test. Therefore some researchers might turn to another approach known as robust methods.

The robust method is not a new approach where it was introduced decades ago. But it becomes more popular nowadays due to its good performance in dealing with violation of assumptions. To handle various problems in classical methods, researchers developed many new robust methods as alternatives. The Alexander-Govern (*AG*) test is one of the robust methods for comparing independent groups when data are heterogeneous [2].

[3] and [4] investigated on how the *AG* test performs. They found that this test is a good alternative to *ANOVA* because of its simple computation and the overall superiority when considering both Type I error rates and power under experimental conditions. However, this test seems to suffer when the normal distribution assumption is not fulfilled [4].

Currently, there is no study yet done, modifying the AG test with robust estimator which consider friendly to data distribution. Motivated to produce better statistical test, which able to handle the non-normality and the heterocedasticity, the main objective of this paper is to modify the *AG* test using robust estimators known as adaptive trimmed mean and adaptive *Winsorized* mean.

2. The Proposed Statistical Test

The original AG test is testing the equality of means with null hypothesis of

$$H_0 : \mu_1 = \mu_2 = ... = \mu_J$$

where $\mu_1=\mu_2=\ldots=\mu_J$ are the mean of J independent groups.

Every J groups with size, n_j has a sample mean ($\bar{X}_j$) and each of the mean has a standard error (S_j) which derived as:

$$S_j = S_{\overline{X}_j} = \sqrt{\frac{\sum_{i=1}^{n_j}(X_i - \overline{X}_j)^2}{n_j(n_j-1)}} \quad (1)$$

Then the weight (w_j) is calculated as:

$$w_j = \frac{1/S_j^2}{\sum_{j=1}^{J} 1/S_j^2} \quad (2)$$

such that $\sum w_j = 1$. The weighted mean (X^+) is computed as:

$$X^+ = \sum_{j=1}^{J} w_j \overline{X}_j \quad (3)$$

One-sample *t* statistic t_j is then calculated using the weighted mean as follow:

$$t_j = \frac{\overline{X}_j - X^+}{S_j} \quad (4)$$

where each of the t_j is distributed as *t* distribution with $v_j = n_j - 1$ degrees of freedom. The *z* statistic is a normalized transformation of each of the *t* statistic value:

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

where c = $[a\ln(1+t_j^2/v_j)]^{1/2}$; b = $48a^2$ and a = $v_j - 0.5$. The *A*G test statistic is obtained by total up the z_j^2 values:

$$AG = \sum_{j=1}^{J} z_j^2 \quad (6)$$

AG statistic is approximately distributed to χ^2 distribution with (*J*-1) degrees of freedom.

In this study, the adaptive trimmed mean or the adaptive winsorized mean substitute the common mean as it central tendency measure. In order to identify the shape of data distribution, these adaptive trimmed and winsorized mean use hinge estimator HQ_1 to determine how many data should be trimmed or winsorized.

2.1 Adaptive trimmed mean

The adaptive trimmed mean is calculated as

$$m(\gamma_l, \gamma_u) = \frac{1}{h}\sum_{i=g_1+1}^{n_j-g_2} Y_i \quad (7)$$

where $g_1 = [n_j\gamma_l]$, $g_2 = [n_j\gamma_u]$, $h = n_j - g_1 - g_2$, γ_l = lower trimming percentage, γ_u = upper trimming percentage and n_j is the sample size. The percentage of lower and upper trimming identified using hinge estimator HQ_1 [5]. However, the total percentage of trimming is predetermined just like the usual trimmed mean. The standard error of the adaptive trimmed mean is computed as

$$s_{m(\alpha_1,\alpha_2)} = \sqrt{\frac{SS(\alpha_1,\alpha_2)}{h(h-1)}} \tag{8}$$

Where

$$\begin{aligned} SS(\alpha_1,\alpha_2) = {} & (g_1+1)[Y_{(g_1+1)} - \hat{x}_t(\alpha_1,\alpha_2)]^2 + [Y_{(g_1+2)} - \hat{x}_t(\alpha_1,\alpha_2)]^2 + \ldots \\ & + [Y_{(n_j-g_2-1)} - \hat{x}_t(\alpha_1,\alpha_2)]^2 + (g_2+1)[Y_{(n_j-g_2)} - \hat{x}_t(\alpha_1,\alpha_2)]^2 \\ & - \{g_1[Y_{(g_1+1)} - \hat{x}_t(\alpha_1,\alpha_2)] + g_2[Y_{(n_j-g_2)} - \hat{x}_t(\alpha_1,\alpha_2)]\}^2 / n_j \end{aligned} \tag{9}$$

2.2 Adaptive *Winsorized* mean

The adaptive *Winsorized* mean is derived by

$$\bar{x}_{aw}(\alpha_1,\alpha_2)_j = \frac{(m_{1j}+1)x_{m_{1j}+1} + x_{m_{1j}+2} + \ldots + x_{n_j-m_{2j}-1} + (m_{2j}+1)x_{n_j-m_{2j}}}{n_j} \tag{10}$$

Where $m_{1_j} = [\alpha_1 n_j]$, $m_{2_j} = [\alpha_2 n_j]$, α_1 = the proportion of the observations to be *Winsorized* for the lower tail distribution, α_2 = the proportion of the observations to be *Winsorized* for the upper tail distribution and n_j = number of sample size for the j^{th} group.

Each of the adaptive *Winsorized* mean, $\bar{x}_w(\alpha_1,\alpha_2)_j$ will have a standard error, $S_{aw}(\alpha_1,\alpha_2)_j$ and calculated as:

$$S_{aw(\alpha_1,\alpha_2)_j} = \sqrt{\frac{s^2_{aw}(\alpha_1,\alpha_2)}{n(n-1)}} \tag{11}$$

where

$$\begin{aligned} s^2_{aw}(\alpha_1,\alpha_2)_j = {} & (m_{1_j}+1)[x_{m_{1_j}+1} - \bar{x}_{aw}(\alpha_1,\alpha_2)_j]^2 + [x_{m_{1_j}+2} - \bar{x}_{aw}(\alpha_1,\alpha_2)_j]^2 + \ldots + \\ & [x_{n_j-m_{2_j}-1} - \bar{x}_{aw}(\alpha_1,\alpha_2)_j{}^2] + (m_{2_j}+1)[x_{nj-m_{2_j}} - \bar{x}_{aw}(\alpha_1,\alpha_2)_j{}^2] - \\ & \{m_{1_j}[x_{m_{1_j}+1} - \bar{x}_{aw}(\alpha_1,\alpha_2)_j] + m_{2_j}[x_{n_j-m_{2_j}} - \bar{x}_{aw}(\alpha_1,\alpha_2)_j]\}^2 / n_j \end{aligned} \tag{12}$$

Due to the modification, there will be two new proposed *AG* test statistics denoted as *AG_ATM* and *AG_AWM* to represent the *AG* test with adaptive trimmed mean and adaptive winsorized mean respectively. Both *AG_ATM* and *AG_AWM* are also approximately distributed as a χ^2 with (*J*-1) degrees of freedom.

,

3. Methods and Empirical Investigation

To evaluate the proposed test procedures performances, this study uses to manipulate a few variables to create different conditions. For the number of groups, this study focuses on two groups case with balanced and unbalanced sample sizes. For balanced sample sizes, each group will have the equal sample size of 20 while for unbalanced groups, the sample size is 15 and 25. To investigate the robustness of AG to variance heterogeneity, the variance ratios chosen are 1:1 for homogeneous variances condition and 1:36 for the heterogeneous condition. With regards to nonnormality, the investigation considered four types of distributions representing four different shapes of data (i.e. Normal, symmetric with heavy tails, skewed with moderate tails, and skewed with heavy tails). For easy manipulation of the shapes, this study uses the g-h distribution. The g-h distribution transformed from the normal distribution with constant g controlling the value of skewness and h controlling the value of

kurtosis. The level of skewness and kurtosis will increase as the value of g and h increase, respectively. The data are symmetric when g = 0 and h = 0. The values of (g, h) used in this study were (0, 0), (0.5,0), (0.5,0) and (0.5,0.5). Table 1 summarizes the skewness and kurtosis values for four selected situations [6].

[7] investigated the power of the Tukey multiple comparison statistics and found that the power of the test was slightly affected by variance heterogeneity, unequal sample sizes, nonnormality and positive and negative pairings of unequal group sizes with unequal variances. Hence, in addition to the above variables, nature of pairing was also taken into consideration. There were two types of pairing, positive and negative. Positive pairing is when the group with the smallest sample size being paired with the smallest variance while the group with the largest sample size being paired with the largest variance. For the negative pairing, the group with largest sample size being paired with the smallest variance and the group with smallest sample size being paired with the largest variance.

This study uses SAS generator RANNOR to generate pseudo-random variates. Observations from the g-h distributions were generated by transforming the standard normal variables to the g-h random variables using the following equation:

$$Y_{ij} = \begin{cases} \dfrac{\exp(gZ_{ij})-1}{g}\exp(hZ_{ij}^{2}/2) & \text{for } g \neq 0 \\ Z_{ij}\exp(hZ_{ij}^{2}/2) & \text{for } g = 0 \end{cases} \qquad (13)$$

In examining the Type I error rates, the group location measures were set to zero. As there were various opinions on how much the data should be trimmed, this study also examined on a different percentage of trimming. [8] suggested using 20% for symmetric trimming while [8] concluded that, 15% is the most optimal proportion to be trimmed. In this paper, only 15% amount of trimming and winsorization will be considered in *AG_ATM* and *AG_AWM* test.

Table I
Some properties of the *g-h* distribution.

g	h	Skewness	Kurtosis	Shape
0.0	0.0	0.0	3.0	Normal
0.0	0.5	0.0	11986.2	Symmetric with heavy tails
0.5	0	1.81	9.7	Skewed with moderate tails
0.5	0.5	120.1	18393.6	Skewed with heavy tails

4. Results and Discussion

The performance of the investigated procedures in terms of robustness (insensitivity to the violation of assumptions) under a particular condition was based on the stringent criterion of robustness. Under this criterion, a procedure tested at the significance level of $\alpha = 0.05$ is considered robust when the Type I error rate is within the 0.045 and 0.055 intervals.

A more liberal criterion of robustness proposed by [9] was considered as well for identifying moderate performances. For this criterion, any procedure with Type I error rates between the range of 0.025 and 0.075 were considered as robust. Table 2 represents the Type I error rates for *AG_ATM* and *AG_AWM* for all 20 conditions being investigated. The robustness of these tests was compared to the original *AG* test and classical *t*-test as well. The stringent values of Type I error rates were marked as ** and the liberal values with *.

From the results in Table 2, under a normal distribution, the performance of the original *AG* test is indisputable. It is found to be robust in all conditions regardless of the variance ratio. The *AG_ATM* also performs well (robust) under all conditions. In contrast, the winsorization process does not seem to be a good alternative where it fails to improve the performance of the *AG* test even though when the data is normally distributed. It is robust in two conditions only. While the classical *t*-test performs well as expected, with only one condition which is not robust.

When the tail of the distribution becomes heavier while skewness remains normal, the condition clearly put some impact on the robustness of the *AG* test where most of the Type I error rates are reduced. There is only one condition which is not robust when the sample is unbalanced with the positive pairing. For the classical *t*-test, it is robust when the sample size is a balance. While the *AG_AWM* test shows the worst with only robust in one condition. Out of all, the *AG_ATM* demonstrates the most convincing results where the Type I error rates still under control in all conditions.

Under skewed distribution with normal tail, the original *AG* test still can sustain its robustness under most of the conditions. It not robust only in condition balance sample size with unequal variance. The classical t-test found to be

robust only under balanced sample size with equal variance. Other than that, the test fails to control the Type I error rates. Between the *AG_ATM* and the *AG_AWM*, there is a huge gap where the *AG_ATM* has an excellent performance in controlling the Type I error rates but not to the *AG_AWM*. The *AG_AWM* has the worst performance among all with only robust in the condition of unbalanced sample size with equal variance.

For skewed with heavy-tailed distribution, the original AG test can be considered robust where it is capable of controlling Type I error rate in all conditions. The results for *AG_ATM* are as good as well. The classical *t*-test maintains its robustness under the condition of balanced sample size. However, the *AG_AWM* is still the worst, with only one robust condition.

Table II
Type I error rates

Distribution	Sample size	Variance	*AG_ATM*	*AG_AWM*	*AG*	*t*-test
Normal ($g = 0$; $h = 0$)	20, 20	1,1	0.0536**	0.0910	0.0508**	0.052**
		1, 36	0.0656*	0.0990	0.0562*	0.0710*
	15, 25	1,1	0.0484**	0.0692*	0.0468**	0.0540**
		1,36	0.0626*	0.0898	0.0560*	0.0270*
		36,1	0.0632*	0.0674*	0.0478**	0.1290
Symmetry heavy -tailed ($g = 0$; $h = 0.5$)	20, 20	1,1	0.0414*	0.1284	0.0336*	0.0300*
		1, 36	0.0484**	0.1912	0.0340*	0.0490**
	15, 25	1,1	0.0342*	0.0738*	0.0284*	0.0410*
		1,36	0.0424*	0.1790	0.0956	0.0130
		36,1	0.0432*	0.0976	0.0554*	0.1110
Skewed normal tailed ($g = 0.5$; $h = 0$)	20, 20	1,1	0.0476**	0.0874	0.0450**	0.0540**
		1, 36	0.0692*	0.1272	0.0772	0.0890
	15, 25	1,1	0.0452**	0.0612*	0.0476**	0.0440*
		1,36	0.0644*	0.1126	0.0394*	0.0340*
		36,1	0.0492**	0.0804	0.0296*	0.1540
Skewed and heavy-tailed ($g = 0.5$; $h = 0.5$)	20, 20	1,1	0.0328*	0.1040	0.0264*	0.0340*
		1, 36	0.0684*	0.4108	0.0360*	0.0560*
	15, 25	1,1	0.0270*	0.0580*	0.0286*	0.0330*
		1,36	0.0658*	0.3848	0.0458**	0.0140
		36,1	0.0416*	0.2558	0.0288*	0.1020
Number of conditions			* * 6 * 14	* * 0 * 5	** 6 *12	** 4 * 9

5. Conclusions

The classical tests for independent groups such as *t*-test are usually the favorite among practitioners when dealing with two groups case. This method is powerful under perfect data condition; that is when the distribution is normal with equal variance. However, the perfect data condition is hardly achieved in this real world. Real life data come in various conditions such as nonnormal data distribution or unequal variance or the worst condition that is when both occur simultaneously.

Constrained by the assumptions of normal distribution and equal variance, the classical *t*-test has limited usage. For those who are aware of the problems, the selection of the test will be done with precaution. Realizing the importance of using the right test statistics, this study proposes a modified robust test statistics as an alternative to the *t*-test. The *AG* test was

proved to be a better alternative of *t*-test when dealing with heterogeneous variance. Nevertheless, this test still it needs improvement when it cannot handle nonnormality. Therefore this paper introduces adaptive trimmed mean and adaptive *Winsorized* mean as an alternative to the usual mean in the *AG* test. This modification produces two new test statistics denoted as *AG_ATM* and *AG_AWM* for *AG* test using adaptive trimmed mean and AG test using adaptive *Winsorized* mean respectively.

Generally, the results of the investigation show that the modifications done on the *AG* test has improved the performance greatly, especially under nonnormal conditions. However, between the adaptive trimmed mean and the adaptive *Winsorized* mean, there is a huge difference in terms of controlling the Type I error rates. The adaptive trimmed mean in the *AG_ATM* test has increased the robustness of the *AG* test, whereby it is robust in all considered conditions. However, this is not the case for *AG_AWM* where the tests are not robust under most conditions. Without modification, the original *AG* test shows nonrobustness in two conditions only, one under symmetric heavy-tailed and another one is under skewed heavy-tailed. The classical *t*-test still maintains its robustness across different data distribution as long as the data has balance sample size with equal variance.

6. Acknowledgments

The authors would like to acknowledge the work that led to this paper, which was fully funded by the Malaysia Ministry of Education via F**undamental Research Grant Scheme (FRGS) and also to the** Universiti Utara Malaysia for supporting this research.

References

[1] Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and nonnormality. *Educational and Psychological Measurement, 58*(3), 409-429.

[2] 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.

[3] 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), 271-287.

[4] Myers, L. (1998). Comparability of the James' second-order approximation test and the Alexander and Govern A statistic for non-normal heteroscedastic data. *Journal of Statistical Computational Simulation, 60*, 207-222.

[5] Reed III, J. F & Stark, D. B, Hinge estimator of location: Robust to asymmetry. *Computer methods and programming in biomedicine,* 1996, 49, 11-17. [9] Bradley, J. V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology(31), 144-152.

[6] Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (Second ed.): California: Academic Press.

[7] Keselman, H. J. (1976). A power investigation of the Tukey multiple comparison statistics. *Educational and Psychological Measurement, 36*(97), 97-104.

[8] Wilcox, R. R & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, Vol 8, No. 3, 254 – 274.

[9] Keselman, H. J., Wilcox, R. R., Lix, L. M., Algina, J., & Fradette, K. (2007). Adaptive robust estimation and testing. *British Journal of Mathematical and Statistical Psychology, 60*, 267-293.