Determine Robust Procedure for Testing Variance Equality Using Type I Error Rate and Power

Patricka O. Adebayo[1*] and I. Ahmed[2]

[1]Department of Statistics, Phoenix University Agwada, Nasarawa State, Nigeria.
[2]Department of Statistics, Nasarawa State University Keffi, Nasarawa State, Nigeria.

E-mail: adebayo.patrick@phoenixuniversity.edu.ng*

Telephone: +2349019964334*
+2348033627847

ABSTRACT

Comparisons were made between seven of the many procedures used to determine variance homogeneity. The seven tests that have been chosen are the Bartlett's test, the Levene's test (mean), the Levene's test (median), the Levene's test (trimmed mean), the O'Brien test, the Cochran test, and the Fligner's test. Data were simulated to compare the seven procedures using different distributions (Normal, Beta, and Uniform), sample sizes (5, 10, 50, and 100), equal samples (n_1=n_2=...=n_k), and five (5) levels (i.e., k = 5). The power of the test and type I error rate were used to compare the selected procedures at a significant level of 0.05. The findings demonstrated that the Fligner technique is superior to all other procedures when the dataset is normally distributed, whereas the Bartlett procedure is superior regardless of sample size when the dataset is not normally distributed.

(Keywords: homogeneity, type 1 error rate, power of the test, robust, normal, non-normal)

INTRODUCTION

Choosing whether sample differences in central tendency reflect real differences in parent populations is a crucial topic in applied research. The most effective method for examining this phenomenon's hypotheses is the analysis of variance (ANOVA) and students t test, provided that the assumptions of normality, homogeneity of variance, and independent of errors are met. Any assumption that proves false may reduce the test's usefulness and result in faulty or incorrect conclusions (Cochran, 1947; Bodhisuwan, 1991).

Homogeneity of variance is a condition in which the variances of the observations within each group are equal (the absence of which is known as heteroscedasticity). Therefore, it is crucial for researchers to verify this assumption before doing an analysis of variance (ANOVA) and t test to ensure that the homogeneity of group variances assumptions are valid.

There is a large body of statistical literature that addresses the numerous methods that have been proposed for determining the homogeneity of variances. Conover et al. (1981) provide an extensive study on tests for homogeneity of variances. To assess the robustness of tests at nominal significance levels, many tests have been investigated and simulated. The F test (two samples), Bartlett's (1937), and Levene's (1960) tests are the ones that have drawn the greatest attention. It is well known that the normality assumption has a significant impact on the F test (two samples) (Markowski and Markowski 1990). According to Conover et al. (1981) and Lim and Loh (1996), Bartlett's test is incredibly weak against non-normality. Conover et al. (1981) and Lim and Loh (1996) both employed the kurtosis correction for Bartlett's test that Layard (1973) recommended. The modified Bartlett's test shows considerable improvement, but it is still not very reliable.

The homogeneity of variance assumption in an ANOVA process states that treatment variances are equal. This is:

$$H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2,$$

where k represents the number of study groups that were compared. According to Game et al (1979), small deviations from the assumption of

equal variances may not have a significant impact on the outcomes of an ANOVA, but researchers may be concerned about significant departures from the assumption of homogeneity of variance. As a result, a critical step in ANOVA analysis is constantly evaluating the homogeneity of variance. The assumption of homogeneity of variance should therefore be tested using a variety of statistical techniques that are advised in the literature. In order to verify the ANOVA assumptions, this study will concentrate on the following statistics: Bartlett's test, Levene's test (mean, median, and trimmed mean), O'Brien test, Cochran test, and Fligner's test.

Bartlett's Test

A pooled estimate of variance (across all groups) is compared to the sum of the logarithms of the variances of the individual groups in Bartlett's test of the null hypothesis of equality of group variances. The test statistic can be found in:

$$M = v\log s^2 - \sum_{i=1}^{k} v_i \log s_i^2 \ , where$$

$$v_i = (n_i - 1),$$

$$v = \sum_{i=1}^{k} v_i \ ,$$

$$s_i^2 = \frac{1}{(n_i - 1)} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 , and$$

$$s^2 = \frac{1}{v} \sum_{i=1}^{k} v_i s_i^2$$

To evaluate the significance of M,the value of $X^2 = \frac{M}{C}$ (that is chi – squared = $\frac{M}{C}$) where:

$$C = 1 + \frac{1}{3(k-1)} \left(\left(\sum_{i=1}^{k} \frac{1}{v_i} \right) - \frac{1}{v} \right)$$

can be likened to the $k-1$ degree of freedom Chi-square distribution. ((X_{k-1}^2). Other comparable tests can be employed in those circumstances because Bartlett's test is known to be sensitive to non-normality.

O'Brien Test

O'Brien (1979, 1981) Just like Levene's test, O'Brien's test of homogeneity of group differences is performed on a transformation of group data using ANOVA. To perform this test, we transform the following data u_{ij} where x_{ij} represents the i[th] element of the j^{th} group:

$$u_{ij} = \frac{n_j (n_j - 1.5)(x_{ij} - \bar{x}_j)^2 - \frac{SS_j}{2}}{(n_j - 1)(n_j - 2)}$$

Where

$$SS_j = \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

That is the sum of the square deviation from the j^{th} group mean and $n_j$is the number of the elements in the j^{th} group. But if all the experimental groups have the same sample size, the above formula can be simplified as:

$$u_{ij} = \frac{n(n - 1.5)(x_{ij} - \bar{x}_j)^2 - \frac{SS_j}{2}}{(n-1)(n-2)}$$

Cochran's C Test Statistic

For example, in a linear regression model, the C test has been employed as an alternative to Bartlett's, Levene's, and Brown-Forsythe's tests to assess homoscedasticity (literally, same variance). This Cochran's C test should not be confused with the Cochran's Q test, which is used to analyze two-way randomized block designs with several treatments in an experimental design. We would anticipate that since the Cochran's C test is derived from the F-test, it would be sensitive to outliers. The highest variance in the data set and the total variance have a simple estimating equation that goes like this:

$$C = \frac{S_{j,L}^2}{\sum_{i=1}^{N} s_i^2}$$

where N is the number of sample groups included in the data set, and $s_{j.L}$ is the biggest standard deviation in the data series j within the

data set. s_i is also the standard deviation of the data series with $1 \leq i \leq N$. By taking into account variables like N (number of groups/subjects), n (number of replicates in each group), α (level of significance desired), and a F_c value that can be obtained from the F distribution table or derived from computer software for the F function, the Cochran's upper critical value C_{UL} is obtained as follows:

$$C_{UL}(\alpha,n,N) = \left[1 + \frac{N-1}{F_c\left(\frac{\alpha}{N},(n-1),(N-1)(n-1)\right)}\right]^{-1}$$

Levene Test

A one-way analysis of variance on the absolute deviations of observations from their group medians is essentially what this test is. The test statistic for this test is, hence:

$$L_{BF} = \frac{\frac{\sum_{i=1}^{k} n_i(\bar{Z}_{i.} - \bar{Z}_{..})^2}{(k-1)}}{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Z}_{ij} - \bar{Z}_{i.})^2}{(N-k)}}$$

Where Z_{ij} can have one of the following three definitions:

a. $Z_{ij} = \left|X_{ij} - \bar{X}_{i.}\right|$ Where $\bar{X}_{i.}$ is the mean of the ith subgroup.

b. $Z_{ij} = \left|X_{ij} - \bar{X}_{i.}\right|$ Where $\bar{X}_{i.}$ is the median of the ith subgroup

c. $Z_{ij} = \left|X_{ij} - \bar{X}_{i.}'\right|$ Where $\bar{X}_{i.}'$ is the 10% trimmed mean of the ith subgroup

$\bar{Z}_{i.}$ are the group means of the Z_{ij} and $Z_{..}$ is the overall mean of the Z_{ij}.

Levene's test's robustness and power are determined by the three options for determining Z_{ij}. When we talk about robustness, we imply the test's capacity to avoid misidentifying unequal variances when the underlying data are not normally distributed, and the variables are equal. The test's power is its capacity to identify unequal variances when they exist. Additionally, it also takes advantage of the fact that ANOVA procedure is fairly robust against nonnormality. If $L_{BF} > F_{\alpha,k-1,N-k}$, this test rejects the hypothesis that the k variances are equal (at a significance level of α).

The Levene test has been shown to be a robust and powerful test in many simulation studies (Brown and Forsythe, 1974; Conover, Johnson and Johnson, 1981; Lim and Loh, 1996; Shoemaker, 2003) particularly for asymmetric or skewed distributions.

Modified Fligner-Killeen Test (F-K:med χ^2)

Conover, *et al*. (1981) suggested modifying the Fligner-Killeen test (Fligner and Killeen, 1976) by using the ranks of $\left|X_{ij} - \bar{X}_i\right|$, R_{ij}, where $\bar{X}_i$ is the median of the ith group, and assigning increasing scores of:

$$a_{N,R_{ij}} = a_{N,i} = \Phi^{-1}\left(\frac{1}{2} + \frac{l}{2(N+1)}\right)$$

based on those ranks, where $\Phi(x)$ is the cdf of a standard normal distribution. The chi-squared test is created based on the statistics from these results.

$$X^2 = \sum_{i=1}^{k} \frac{n_i(\bar{A}_i - \bar{a})^2}{V^2}$$

and

$$V^2 = \sum_{i=1}^{N} \frac{\left(a_{NJ} - \bar{a}\right)^2}{(N-1)}.$$

Where $\bar{A}_j$is the mean score for the jth sample.

$\bar{a}$ is the overall mean score, that is $\bar{a} = \frac{1}{N}\sum_{i=1}^{N} a_{N,j}$

The distribution of statistic X^2 has an asymptotic value of $\chi^2_{k-1}$where k is the total number of variances compared. If $X^2 > \chi^2_{\alpha,k-1}$, then this test rejects the null hypothesis that k variances are identical (at a significance level of α).

Simulation Study

To compare the type I error rate and power of the tests of the chosen procedures, namely: Bartlett's test, Levene's test (mean, median, and trimmed mean), O'Brien test, Cochran test, and Fligner's test, the stimulation research will be carried out using the R program. The simulations are performed when:

- there is equal variance $(\Sigma_1 = \Sigma_2 = \ldots = \Sigma_j)$
- the $H_o\ (\mu_1 = \mu_2 = \ldots = \mu_j)$ is true
- the data are normal and non-normal
- sample size are the same $(n_1 = n_2 = \ldots = n_k)$.
- significant level α are 0.05.

All of these were repeated 1,000 times, and the outcomes were shown in tabular formats.

Table 1: Four Factors used for Evaluation Criteria in Stimulation.

Distributions	Significant Level	Nature of Sample	Levels
Normal	$\alpha = 0.05$	$n_1 = n_2 = \ldots = n_k$	5
Beta	$\alpha = 0.05$	$n_1 = n_2 = \ldots = n_k$	5
Gamma	$\alpha = 0.05$	$n_1 = n_2 = \ldots = n_k$	5
Uniform	$\alpha = 0.05$	$n_1 = n_2 = \ldots = n_k$	5

Data Generation

The dataset used for the procedure under consideration included the normal, beta, gamma, and uniform distributions. It was replicated one thousand times (1000) using the R package when the sample sizes were equal (n_1=n_2= ... 〖=n〗 _k), and the significant level was set at 0.05 when there was no statistically significant difference in the variance.

RESULTS

All the procedures considered for the comparison are Bartlett's test, Levene's test (mean, median and trimmed mean), O'Brien test, Cochran test and Fligner's test, respectively.

Table 2: Simulation Result on Type I Error Rate when Sample Size are Equal (Normal Distribution).

Sample	Bartllet	Fligner	L.mean	L.med	L.trim	O'Brien	Cochran
Level = 5 rnorm(n1,3,8), rnorm(n2,3,8), rnorm(n3,3,8) , rnorm(n3,3,8), rnorm(n3,3,8)							
5	0.035	0.000	0.094	0.003	0.045	0.034	0.041
10	0.047	0.024	0.068	0.031	0.054	0.055	0.049
50	0.054	0.038	0.047	0.039	0.047	0.044	0.049
100	0.050	0.041	0.049	0.041	0.046	0.045	0.048

Table 3: Simulation Result on Power of the Test when Sample Size are Equal (Normal Distribution).

Sample	Bartllet	Fligner	L.mean	L.med	L.trim	O'Brien	Cochran
Level = 5 rnorm(n1,3,8), rnorm(n2,3,8), rnorm(n3,3,8) , rnorm(n3,3,8) , rnorm(n3,3,8)							
5	0.953	0.914	0.974	0.908	0.953	0.947	0.824
10	0.953	0.949	0.964	0.943	0.955	0.959	0.804
50	0.959	0.955	0.957	0.950	0.956	0.953	0.785
100	0.946	0.936	0.934	0.933	0.936	0.934	0.769

Table 4: Simulation Result on Type I Error Rate when Sample Size are Equal (Gamma Distribution).

Sample	Bartllet	Fligner	L.mean	L.med	L.trim	O'Brien	Cochran
Level = 5 rgamma(n, 1.5, 0.5), rgamma(n, 1.5, 0.5), rgamma(n, 1.5, 0.5), rgamma(n, 1.5, 0.5), rgamma(n, 1.5, 0.5)							
5	0.223	0.000	0.216	0.013	0.098	0.054	0.221
10	0.347	0.059	0.191	0.039	0.070	0.058	0.286
50	0.485	0.089	0.186	0.049	0.074	0.052	0.388
100	0.480	0.081	0.154	0.038	0.053	0.041	0.364

Table 5: Simulation Result on Power of the Test when Sample Size are Equal (Gamma Distribution).

Sample	Bartllet	Fligner	L.mean	L.med	L.trim	O'Brien	Cochran
Level = 5 rgamma(n, 1.5, 0.5), rgamma(n, 1.5, 0.5), rgamma(n, 1.5, 0.5), rgamma(n, 1.5, 0.5), rgamma(n, 1.5, 0.5)							
5	0.980	0.910	0.989	0.936	0.969	0.961	0.910
10	0.989	0.966	0.987	0.955	0.968	0.962	0.951
50	0.993	0.965	0.979	0.955	0.958	0.957	0.962
100	0.992	0.964	0.974	0.944	0.953	0.945	0.954

Table 6: Simulation Result on Type I Error Rate when Sample Size are Equal (Beta Distribution).

Sample	Bartllet	Fligner	L.mean	L.med	L.trim	O'Brien	Cochran
Level = 5 rbeta(n, 0.7, 1.5), rbeta(n, 0.7, 1.5), rbeta(n, 0.7, 1.5), rbeta(n, 0.7, 1.5), rbeta(n, 0.7, 1.5)							
5	0.060	0.000	0.181	0.003	0.091	0.020	0.021
10	0.040	0.041	0.144	0.045	0.076	0.075	0.016
50	0.017	0.078	0.136	0.065	0.086	0.070	0.013
100	0.007	0.083	0.124	0.042	0.072	0.046	0.008

Table 7: Simulation Result on Power of the Test when Sample Size are Equal (Beta Distribution).

Sample	Bartllet	Fligner	L.mean	L.med	L.trim	O'Brien	Cochran
Level = 5 rbeta(n, 0.7, 1.5), rbeta(n, 0.7, 1.5), rbeta(n, 0.7, 1.5), rbeta(n, 0.7, 1.5), rbeta(n, 0.7, 1.5)							
5	0.937	0.892	0.972	0.880	0.950	0.927	0.775
10	0.913	0.930	0.975	0.923	0.957	0.948	0.703
50	0.901	0.945	0.970	0.940	0.958	0.944	0.653
100	0.911	0.961	0.971	0.947	0.958	0.953	0.649

Table 8: Simulation Result on Type I Error Rate when Sample Size are Equal (Uniform Distribution).

Sample	Bartllet	Fligner	L.mean	L.med	L.trim	O'Brien	Cochran
Level = 5 runif(n,3,8), runif(n,3,8), runif(n,3,8) , runif(n,3,8), runif((n,3,8)							
5	0.012	0.000	0.089	0.000	0.040	0.015	0.004
10	0.003	0.011	0.065	0.019	0.042	0.050	0.000
50	0.000	0.035	0.057	0.030	0.046	0.040	0.001
100	0.000	0.038	0.052	0.038	0.049	0.044	0.000

Table 9: Simulation Result on Power of the Test when Sample Size are Equal (Uniform Distribution).

Sample	Bartllet	Fligner	L.mean	L.med	L.trim	O'Brien	Cochran
Level = 5 runif(n,3,8), runif(n,3,8), runif(n,3,8) , runif(n,3,8), runif((n,3,8)							
5	0.847	0.854	0.952	0.836	0.932	0.906	0.606
10	0.837	0.942	0.961	0.923	0.956	0.942	0.516
50	0.797	0.943	0.959	0.948	0.961	0.953	0.441
100	0.753	0.947	0.944	0.933	0.937	0.934	0.397

DISCUSSION OF RESULTS

When the sample size is minimal (say, 5) in all the groups in Table 2, 4, and 6, the Fligner test outperformed all other procedures in terms of type I error rate for all the datasets evaluated (Normal, Gamma, beta and uniform distribution). However, when the dataset is normally distributed, the Fligner test outperforms every other method in terms of type I error rate for all sample sizes (5, 10, 50, and100).

In terms of type I error rate, using a normal distribution from Table 2, the Levene test (mean) fared poorly when the sample size was small and competed well with other procedures as the sample size increased. However, compared to the other tests from Tables 4, 5, 6, and 7, the Levene test (mean) has the highest type I error rate and power when the distribution is not normally distributed (beta, gamma, and uniform).

From Tables 4, 5, 6, and 7, it can be shown that the Cochran test and the Bartlett test compete with one another in terms of type I error rate when the dataset is not normal (i.e., uniform and beta distribution). However, the Bartlett test performs better than the Cochran test in terms of test power.

With the exception of the Cochran test, all procedures considered outperformed one another in terms of test power when the dataset was normal and not normally distributed.

CONCLUSION

In conclusion, the Fligner test is the best procedure when the data is normally distributed for both small and large sample sizes, and the Bartlett test is the procedure to employ when the data is not normally distributed.

REFERENCES

1. Bartlett, M.S. 1937. "Properties of Sufficiency and Statistical Tests". *Proc. Roy. Soc. Ser. A*. 160: 268–282.

2. Bodhisuwan, W. 1991. "A Comparison of the Test Statistics for Homogeneity of Variances". MS Thesis in Statistics. Faculty of Graduate Studies, Chulalongkorn University: Bangkok, Thailand.

3. Brown, M.B and A.B. Forsythe. 1974. "Robust Tests for the Equality of Variances". *JASA*. June; 69: 364- 7.

4. Cochran, W.G. 1947. "Some Consequences when the Assumptions for the Analysis of Variance are not Satisfied". *Biometics*, 3: 22-38.

5. Conover, W.J., M.E. Johnson, and M.M. Johnson. 1981. "A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data". *Technometrics*. 23: 351–361.

6. Fligner, M.A. and T.J. Killeen. 1976. "Distribution-Free Two-Sample Tests for Scale". *J. Amer. Statist. Assoc*. 71: 210- 213.

7. Games, P.A., H.J. Keselman, and J.J. Clinch. 1979. "Tests for Homogeneity of Variance in Factorial Designs". *Psychological Bulletin*. 86: 978–984.

8. Glass, G.V. and J.C. Stanley. 1970. *Statistical Methods in Education and Psychology*. Prentice-Hall: Englewood Cliffs, NJ.

9. Layard, M.W.J. 1973. "Robust Large-Sample Tests for Homogeneity of Variance". *J. Amer. Statist. Assoc*. 68: 195-198.

10. Levene, H. 1960. "Robust Test for the Equality of Variances". In: Olkin, I. (Ed.). *Contributions to Probability and Statistics*. Standford University Press: Palo Alto, CA.

11. Levene, H. 1949. “On a Matching Problem arising in Genetics”. *Ann. Math. Statist*. 20: 91–94.

12. Levene, H. 1953. “Genetic Equilibrium when more than one Ecological Niche is Available”. *American Naturalist*. 87: 331–333.

13. Lim, T.S. and W.Y. Loh. 1996. “A Comparison of Tests of Equality of Variances”. *Comput. Statist. Data Anal*. 22: 287–301.

14. Markowski, C.A. and E.P. Markowski. 1990. “Conditions for the Effectiveness of a Preliminary Test of Variance”. *Am. Stat.* 44: 322–326.

15. Martin, C.G. and P.A. Games. 1977. “ANOVA Tests for Homogeneity of Variance: Non-Normality and Unequal Samples”. *Journal of Educational Statistics*. 2: 187–206.

16. Miller, R.G. 1968. “Jackkniffiing Variances”. *Annals of Mathematical Statistics*. 39: 567–582.

17. O’Brien, R.G. 1979. “A General ANOVA Method for Robust Test of Additive Models for Variance”. *Journal of the American Statistical Association*. 74: 877–880.

18. O’Brien, R.G. 1981. “A Simple Test for Variance Effects in Experimental Designs”. *Psychological Bulletin*. 89: 570–574.

19. Shoemaker, L.H. 2003. “Fixing the F Test for Equal Variances”. *Amer. Statist.* 57: 105– 114.

SUGGESTED CITATION

Adebayo, P.O. and I. Ahmed. 2024. “Determine Robust Procedure for Testing Variance Equality Using Type I Error Rate and Power”. *Pacific Journal of Science and Technology*. 25(1): 24-30.