PERFORMANCE OF TESTS FOR HOMOGENEITY OF VARIANCES FOR MORE THAN TWO SAMPLES

[1]KNAVOOT JIAMWATTANAPONG, [2]NISANAD INGADAPA

Department of General Education, Faculty of Liberal Arts Rajamangala University of Technology Rattanakosin, Thailand
E-mail: [1]knavoot.jia@rmutr.ac.th, nisanad.ing@rmutr.ac.th

Abstract - To test for homogeneity of variances plays a critical role in not only conducting the analysis of variance (ANOVA) but also monitoring the process in quality control. The performance of five tests including Cochran's test, Z-variance test, O'Brien test, Levene's test, and Modified Levene's Test was evaluated via a simulation study. Comparisons were made under the null hypothesis and the alternative with various difference levels of variances. Overall, Cochran's test and Z-variance test perform better than the other three tests and the two tests can be used in small sample sizes even five. Guidelines for selecting tests for homogeneity of variances and alternative methods to compare treatment means in ANOVA when the homogeneity of variances is violated are provided.

Keywords - Analysis of Variance, Equality of Variances, Heteroscedasticity, Test for Homogeneity of Variances.

I. INTRODUCTION

Testing for homogeneity of variances (HOV) is one of the most important techniques implemented for validating an assumption in the analysis of variance (ANOVA), the most frequently used statistical method in applied research [1]. When the variability is unequal across the range of values of a second variable, called heteroscedasticity, the results in ANOVA would be affected and might lead to a wrong conclusion on population means, particularly for the case of unequal sample sizes [2]-[4]. Testing for HOV is also beneficial to production monitoring in quality control. Due to its widespread applications, tests for HOV are commonly found in various fields such as science, engineering, and social sciences. Although there have been extensive studies on HOV, Cochran's test [5] is often used to detect the heterogeneity of variances when there is one group holding higher variance than the others. However, one major drawback of Cochran's test is that it is dependent on the sample size. Excluding Cochran's test, there are alternative tests of HOV, for example, Z-variance test [6], O'Brien test [7], Levene's test [8] and Modified Levene's test [9]. Many alternative methods for testing HOV have been proposed but are not available in common statistical packages, e.g. Statistical Package for Social Sciences (SPSS). Testing for HOV is a major concern and therefore it has attracted much attention from statisticians and scientists. The purpose of this study is to provide comparative information on the tests for equality of variances under normality when the sample sizes are equal.

II. FIVE HOV TESTS FOR MORE THAN TWO SAMPLES

Let $\left\{X_{ij}\right\}, i=1,2,...,k\,;\ j=1,2,...,n_i$ be random samples from populations with means μ_i and variances σ_i^2. The null hypothesis for testing the equality of variances is $H_0: \sigma_1^2=\sigma_2^2=...=\sigma_k^2$. Some tests reviewed here rely on the assumption of normal data whereas the others do not. Five tests for HOV include Cochran's test, Z-variance test, O'Brien test, Levene's test, and Modified Levene's Test. The exact table is available for Cochran's test and the asymptotic approximations are available for the others.

A. Cochran's Test

Cochran's test, or Cochran variance outlier test, is a ratio of the largest group variance to the sum of sample variances [5]. The null hypothesis of equality of variances is rejected when the test statistic exceeds the critical value obtained from the exact table, presented by Cochran (1941). The exact critical values are, however, available for the case of equal sample sizes and also depend on the sample size. It should be noted here that in the case of unequal sample sizes, either arithmetic mean or harmonic mean may be used to obtain the degrees of freedom. In this study, only the case of equal sample sizes was assessed.

Cochran's test statistic T_C is shown as

$$T_C = S^2_{\max} / \sum_{i=1}^{k} S_i^2 \ ,$$

where $S^2_{\max}$ is the largest group variance and S_i^2 is the ith sample variance.

B. Z-variance Test

Z-variance test, proposed by Overall and Woodward [6], is the test based on the idea presented by Fisher and Yates [10] but using the transformation of

Chi-square statistic S_i^2 into standard normal variate Z_i. Under normality and when the null hypothesis is true, the test statistic does not depend on the sample size and the approximated distribution is $F_{k-1,\infty}$. The test statistic T_Z is as follows.

$$T_Z = \sum_{i=1}^{k} Z_i^2 / (k-1),$$

where $Z_i = \sqrt{\frac{c_i(n_i-1)S_i^2}{MSE}} - \sqrt{c_i(n_i-1) - \frac{c_i}{2}}$,

$c_i = 2 + \frac{1}{n_i}$, $MSE = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2}{N-k}$,

and $N = \sum_{i=1}^{k} n_i$.

C. O'Brien's Test

O'Brien's test is formed by transforming of X_{ij} into V_{ij} as follows.

$$V_{ij} = \frac{n_i(n_i-1.5)(X_{ij} - \bar{X}_i)^2 - 0.5S_i^2(n_i-1)}{(n_i-1)(n_i-2)},$$

where $S_i^2 = \frac{\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}{n_i - 1}$.

The mean of the transformed values equals to the sample variance, i.e. $\bar{V}_i = \sum_{j=1}^{n_i} V_{ij} / n_i = S_i^2$. The O'Brien test statistic is obtained by using ordinary ANOVA but using V_{ij} instead of X_{ij}. The test performs well for behavioral science data [7].

D. Levene's Test

Levene (1960) proposed a test for equality of variances based on F-statistic in one-way ANOVA model. The values used to analyze are the deviate $Z_{ij} = \left| X_{ij} - \bar{X}_i \right|$. The Levene's test statistic T_L is as follows.

$$T_L = \frac{\sum_{i=1}^{k} n_i(\bar{Z}_i - \bar{Z})^2 / (k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{ij} - \bar{Z})^2 / (N-k)},$$

where $\bar{Z}_i$ denotes the mean of Z_{ij} from the ith sample and $\bar{Z}$ is the grand mean of Z_{ij}. Levene's test is robust to non-normal populations [9], [11].

E. Modified Levene's Test

One weakness of Levene's test is that the test does not perform well when the population mean and the population median are different [12]; in this case Brown and Forsythe [9] suggested to use the median instead of mean. The variate Z_{ij} is obtained from $Z_{ij} = \left| X_{ij} - \tilde{X}_i \right|$, where $\tilde{X}_i$ is median of the ith sample.

The Modified Levene's test statistic T_{ML} is as follows.

$$T_{ML} = \frac{\sum_{i=1}^{k} n_i(\bar{Z}_i - \bar{Z})^2 / (k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{ij} - \bar{Z})^2 / (N-k)}.$$

III. SIMULATION PROCEDURE

In this section, the performance of the tests was evaluated via a simulation study by considering the attained significance level and the empirical power of the tests. The condition settings are as follows: the number of samples $k = \{3,4,5,6\}$, the sample sizes $n_i = \{5,10\}$ and all group sizes are equal, the nominal significance levels $\alpha = \{0.05, 0.01\}$, and throughout the simulation study, the independent normal samples were generated with 10,000 iterations.

Under the null hypothesis that all group variances are equal, the attained significance level (ASL) was calculated as the proportion of the number of times the calculated test statistics falling inside the critical region in 10,000 times under the null hypothesis. Bradley's liberal criterion [13] was used to consider the ASL. When the ASLs of a test fall within the interval $(0.5\alpha, 1.5\alpha)$, the test is considered to be acceptable.

To obtain the empirical power, the simulation was conducted under the alternative hypothesis of heterogeneity of variances. In this case, one group of the k samples with larger variance was determined. The degree of the difference among variances, called difference levels of variances, was measured in a noncentrality parameter ϕ, presented by Games, P.A., Winkle, H.B., and Probert, D.A. [14], as

$$\phi^2 = \frac{\sum_{i=1}^{k} (\sigma_i^2 - \overline{\sigma^2})^2}{k\sigma_{\min}^2},$$

where σ_i^2 denotes the ith population variance, $\overline{\sigma^2}$, the mean of k population variances, and $\sigma_{\min}^2$, the lowest population variance. The difference levels were classified as low, medium, and high when the noncentrality parameters are $0 < \phi < 1.5$, $1.5 \leq \phi < 3.0$ and $\phi \geq 3.0$, respectively. The empirical power was computed as the proportion of the number of times the calculated test statistics falls inside the critical region to 10,000 iterations under the alternative hypothesis.

The condition settings under the alternative hypothesis were shown in Table I.

k	n_i	Difference Level of Variances	Ratio of variances	ϕ
3	5	Low	1.5:1:1	0.24
	5	Medium	5:1:1	1.89
	5	High	8:1:1	3.30
	10	Low	1.5:1:1	0.24
	10	Medium	5:1:1	1.89
	10	High	8:1:1	3.30
4	5	Low	1.5:1:1:1	0.22
	5	Medium	5:1:1:1	1.73
	5	High	8:1:1:1	3.03
	10	Low	1.5:1:1:1	0.22
	10	Medium	5:1:1:1	1.73
	10	High	8:1:1:1	3.03
5	5	Low	1.5:1:1:1:1	0.20
	5	Medium	5:1:1:1:1	1.60
	5	High	9:1:1:1:1	3.20
	10	Low	1.5:1:1:1:1	0.20
	10	Medium	5:1:1:1:1	1.60
	10	High	9:1:1:1:1	3.20
6	5	Low	1.5:1:1:1:1	0.19
	5	Medium	6:1:1:1:1	1.86
	5	High	9.5:1:1:1:1	3.17
	10	Low	1.5:1:1:1:1	0.19
	10	Medium	6:1:1:1:1	1.86
	10	High	9.5:1:1:1:1	3.17

Table I. Condition settings under the alternative hypothesis

IV. RESULTS

The performance of the tests was considered using both ASL and empirical power. ASLs of the five tests were shown in Table II and the power in Table III to V.

k	n_i	T_C	T_Z	T_O	T_L	T_{ML}
α = 0.05						
3	5	0.0497	0.0368	0.0264	0.0816	0.0042
	10	0.0514	0.0489	0.0359	0.0632	0.0329
4	5	0.0495	0.0386	0.0296	0.0923	0.0034
	10	0.0495	0.0430	0.0371	0.0660	0.0298
5	5	0.0487	0.0372	0.0336	0.0972	0.0029
	10	0.0464	0.0470	0.0354	0.0665	0.0271
6	5	0.0526	0.0369	0.0307	0.1021	0.0033
	10	0.0503	0.0466	0.0384	0.0697	0.0272

α = 0.01						
3	5	0.0080	0.0031	0.0032	0.0172	0.0000
	10	0.0102	0.0065	0.0054	0.0134	0.0048
4	5	0.0101	0.0048	0.0053	0.0234	0.0001
	10	0.0110	0.0082	0.0075	0.0164	0.0049
5	5	0.0100	0.0031	0.0054	0.0257	0.0003
	10	0.0127	0.0085	0.0077	0.0161	0.0096
6	5	0.0099	0.0042	0.0066	0.0303	0.0003
	10	0.0098	0.0079	0.0074	0.0173	0.0036

Table II. Attained significance levels of the five tests

At 0.05 nominal significance level, it is obvious that the test T_C performed better than any other tests. The ASLs of T_L were not close to the nominal alpha, obviously, they were all over. The ASLs of the other tests were lower the nominal value which in some cases of n_i = 10, the ASLs of T_Z were acceptable whereas the other tests were not. When the sample size was 10, the ASLs of the tests were closer to the nominal level. At 0.01 nominal level, the test T_C still outperformed and the second best was T_Z.

The empirical powers of the tests under low, medium and high difference levels of variances were shown, respectively, in Tables III to V.

k	n_i	T_C	T_Z	T_O	T_L	T_{ML}
α = 0.05						
3	5	0.1278	0.0962	0.0526	0.1361	0.0148
	10	0.2566	0.2204	0.1443	0.2051	0.1311
4	5	0.1390	0.1018	0.0063	0.1572	0.0134
	10	0.2663	0.2253	0.1650	0.2234	0.1359
5	5	0.1403	0.0974	0.0660	0.1646	0.0113
	10	0.2807	0.2306	0.1800	0.2277	0.1328
6	5	0.1460	0.0969	0.0690	0.1744	0.0102
	10	0.2804	0.2250	0.1828	0.2307	0.1277
α = 0.01						
3	5	0.0345	0.0122	0.0100	0.0437	0.0000
	10	0.1038	0.0719	0.0368	0.0734	0.0353
4	5	0.0474	0.0210	0.0148	0.0551	0.0006
	10	0.1239	0.0813	0.0528	0.0785	0.0425
5	5	0.0511	0.0213	0.0177	0.0524	0.0006
	10	0.1294	0.0870	0.0615	0.0943	0.0422
6	5	0.0521	0.0247	0.0234	0.0670	0.0020
	10	0.1391	0.0931	0.0684	0.0940	0.0419

Table III. Empirical powers of the tests under low difference level of variances

When the variance of one group was slightly different from the others, see Table III, the empirical powers of all of the tests were unsatisfied. As the sample size was increased from 5 to 10, the powers of all tests were higher. The powers showed similar results at the nominal levels were either 0.05 or 0.01. When the variance of one group was moderately different from the others, see Table IV, at 0.05 level, the test T_C outperformed and the tests T_Z and T_L were the next best. The tests T_{ML} and T_O did not perform well for small sample size of 5 but they all performed better as the sample size was up to 10. At 0.01 nominal level, the test T_C performed the best and the second best was the test T_Z. The other three tests were acceptable only when the sample size was 10, not 5.

From Table V, when the different level of variance was high, it is obvious that the powers of the tests T_C, T_Z and T_L were high whereas the other two tests T_{ML} and T_O performed well only when the sample size was 10. The tests showed similar results at both the nominal levels 0.05 and 0.01.

k	n_i	T_C	T_Z	T_O	T_L	T_{ML}
α = 0.05						
3	5	0.9104	0.8774	0.2884	0.7030	0.3250
	10	0.9977	0.9971	0.8327	0.9796	0.9512
4	5	0.9324	0.9021	0.3724	0.7854	0.4170
	10	0.9986	0.9979	0.8980	0.9875	0.9706
5	5	0.9401	0.9132	0.4314	0.8221	0.4748
	10	0.9995	0.9985	0.9304	0.9929	0.9796
6	5	0.9708	0.9539	0.5156	0.9041	0.6082
	10	0.9997	0.9997	0.9603	0.9973	0.9940
α = 0.01						
3	5	0.8023	0.6856	0.1069	0.4091	0.0635
	10	0.9932	0.9893	0.5158	0.9052	0.7984
4	5	0.8680	0.7844	0.1751	0.5249	0.1335
	10	0.9963	0.9937	0.6719	0.9474	0.8787
5	5	0.8880	0.8204	0.2382	0.6025	0.1999
	10	0.9973	0.9951	0.7662	0.9690	0.9166
6	5	0.9446	0.9069	0.3094	0.7289	0.3282
	10	0.9595	0.9990	0.8535	0.9908	0.9696

Table IV. Empirical powers of the tests under medium difference level of variances

k	n_i	T_C	T_Z	T_O	T_L	T_{ML}
α = 0.05						
3	5	0.9827	0.9753	0.3349	0.8470	0.4857
	10	0.9999	0.9998	0.8757	0.9981	0.9923
4	5	0.9861	0.9789	0.4289	0.9161	0.6150
	10	0.9999	0.9999	0.9250	0.9991	0.9970
5	5	0.9408	0.9116	0.4296	0.8225	0.4763
	10	0.9995	0.9985	0.9311	0.9909	0.9782
6	5	0.9939	0.9899	0.5501	0.9722	0.7603
	10	1.0000	1.0000	0.9698	1.0000	0.9992
α = 0.01						
3	5	0.9523	0.9087	0.1357	0.5352	0.1183
	10	0.9999	0.9998	0.5685	0.9737	0.9226
4	5	0.9686	0.9461	0.2144	0.6635	0.2523
	10	0.9999	0.9999	0.7330	0.9924	0.9717
5	5	0.9860	0.9750	0.2983	0.7963	0.3919
	10	1.0000	1.0000	0.8244	0.9981	0.9886
6	5	0.9895	0.9821	0.3480	0.8686	0.4963
	10	0.9998	0.9997	0.8781	0.9989	0.9938

Table V. Empirical powers of the tests under high difference level of variances

V. CONCLUSIONS AND RECOMMENDATIONS

In this study, the variance of one group was set to be higher than the others to examine the performance of five tests for homogeneity of variances including Cochran's test, Z-variance test, O'Brien's test, Levene's test, and Modified Levene's test. Overall, all of the five tests are unable to detect low different level of variances. However, when the difference levels are medium or high, all of the tests perform better. Cochran's test and Z-variance test performed better than the others across all the conditions studied. These two tests can be used for small sample size even 5, and when the sample size is 10, these two tests perform better. The advantage of Cochran's test for detecting one higher variance group is confirmed by the results, followed by the Z-variance test. Excluding Cochran's test and Z-variance test, the other three tests can be ordered as Levene's test, Modified's Levene's test, and O'Brien's test. It is worth here to give some recommendations for those who have to examine the

equality of variances before conducting further analyses such as ANOVA. When the tests which are not sensitive to detect the heterogeneity of variances such as Levene's test, Modified Levene's test, and O'Brien's test are used, it may not be found to be significant; this can lead to incorrect results in ANOVA. In this case, the methods such as Cochran's test and Z-variance test should be considered because they can better detect the difference of variances. The alternative methods for ANOVA procedure such as Brown-Forsythe's test [9] and Welch's test [15] are likely to be used when the assumption of homogeneity of variances is not met. Furthermore, when a violation of assumption occurs, a non-parametric equivalent of the analysis is likely to be conducted.

ACKNOWLEDGEMENT

This study was financially supported by Rajamangala University of Technology Rattanakosin (RMUTR).

REFERENCES

[1] S. L. Jan, and G. Shieh, "Determining sample sizes for precise contrast analysis with heterogeneous variances," Journal of Educational and Behavioral Statistics, vol. 39(2), pp. 91-116, 2014.

[2] G. E. Box, "Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification," The annals of mathematical statistics, vol. 25(2), pp. 290-302, 1954.

[3] T. A. Bishop, and E. J. Dudewicz, "Exact analysis of variance with unequal variances: test procedures and tables," Technometrics, vol. 20(4), pp. 419-430, 1978.

[4] S. Y. Chen, and H. J. Chen, "Single-stage analysis of variance under heteroscedasticity," Communications in Statistics-simulation and computation, vol. 27(3), pp. 641-666, 1998.

[5] W. G. Cochran, "The distribution of the largest of a set of estimated variances as a fraction of their total," Annals of Eugenics, vol. 11(1), pp. 47-52, 1941.

[6] J. E. Overall, and J. A. Woodward, "A simple test for heterogeneity of variance in complex factorial designs," Psychometrika, vol. 39(3), pp. 311-318, 1974.

[7] R. G. O'Brien, "Robust techniques for testing heterogeneity of variance effects in factorial designs," Psychometrika, vol. 43(3), pp. 327-342, 1978.

[8] H. Levene, "Robust tests for equality of variances," Contributions to probability and statistics: essays in honor of Harold Hotelling, Stanford University Press, Palo Alto, CA. pp. 278–292, 1960.

[9] M. B. Brown, and A. B. Forsythe, "Robust tests for the equality of variances," Journal of the American Statistical Association, vol. 69(346), pp. 364-367, 1974.

[10] R. A. Fisher, and F. Yates, "Statistical tables for biological, agricultural and medical research," Statistical tables for biological, agricultural and medical research, 6th ed, 1963.

[11] R. R. Fellers, "The effects of nonnormality and sample size on the robustness of tests of homogeneity of variance," Paper presented at the meeting of the Northeast Educational Research Association, 1972.

[12] R. G. Miller, "Jackknifing variances," The Annals of Mathematical Statistics, vol. 39(2), pp. 567-582, 1968.

[13] J. V. Bradley, "Robustness?," British Journal of Mathematical and Statistical Psychology, vol. 31(2), pp. 144-152, 1978.

[14] P. A. Games, H. B. Winkler, and D. A. Probert, "Robust tests for homogeneity of variance," Educational and Psychological Measurement, vol. 32(4), pp. 887-909, 1972.

[15] B. L. Welch, "On the comparison of several mean values: an alternative approach," Biometrika, vol. 38(3/4), pp. 330-336, 1951.

★★★