
Proceedings, 2nd Stochastic Modeling Techniques and Data Analysis International Conference
5 - 8 June 2012 Chania Crete Greece

Application of Parametric Homogeneity of Variances Tests under Violation of Classical Assumption

Alisa A. Gorbunova and Boris Yu. Lemeshko

Novosibirsk State Technical University
Department of Applied Mathematics, Novosibirsk, Russia
(e-mail: gorbunova.alisa@gmail.com, Lemeshko@fpm.ami.nstu.ru)

Abstract. There are a good number of tests that are available for testing a hypothesis that samples come from populations with the same variance. It is well known that classical tests for comparing variances are very sensitive to departures from normality. However, they are more powerful than nonparametric ones. So, the new approach for testing hypotheses of variances homogeneity is proposed. Software for comparing variances using *parametric* tests (F-test, Cochran's, Bartlett's, Hartley's, Levene's, modified Levene's, Neyman-Pearson's, Z-variance, Overall-Woodward modified Z-variance and O'Brien tests) when samples are from any distribution (skewed, leptokurtic, platykurtic) has been developed. In this case the p-value is defined using a simulated empirical distribution in real-time testing of the hypothesis. Recommendations on choosing the most powerful test for a particular form of data distribution are given.
Keywords: homogeneity of variances, power, simulation study.

Introduction

Testing for equality of variances often attracts attention as preliminary to other analyses involving comparisons of means, such as an analysis of variance (ANOVA) or the t-test. Correct application of tests for means equality implies that variances are equal.

However, preliminary tests of variances equality used before applying a test of location are not recommended by some statisticians. Many authors (e.g., Zar[1]) stated that the tests presently available have such a poor performance that they are not really useful, with ANOVA being more robust to departures from homoscedasticity than can be detected using a test of homogeneity of variances, especially under non-normal conditions. But recent study by Legendre and Borcard[2] has showed that "heterogeneity of variances is *always* a problem in ANOVA, and is troublesome even in the most benign cases, i.e., when one of the variances is smaller than the others". So, there is a great need for a test that will correctly detect heterogeneity of variances before applying procedures for means comparison.

The homogeneity of variances tests per se are also of interest in a number of research areas. A variance could be considered as an indicator of uniformity, e.g. in the quality control of manufacturing processes, in agricultural

production systems or in the development of educational methods. Differences in variability of populations could be interesting to biologists, e.g. in the study of genetic diversity or mechanisms of adaption (Boos and Brownie[3]).

So, we want to know whether variances are equal, that is to test hypothesis of variances homogeneity. The null hypothesis for variances equality of m samples has the following form: $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_m^2$ and the alternative hypothesis is $H_1 : \sigma_i^2 \neq \sigma_j^2$, where the inequality holds at least for one pair of i, j.

To test hypothesis H_0 there are a good number of available tests both parametric and nonparametric. Moreover, there is considerable statistical literature on testing homogeneity of variances. Therefore, potential user of a test for equality of variances is faced with a confusing array of information concerning which test to use. And worse, this information is sometimes conflicting.

What are the problems when testing equality of variances? Primarily we should choose an appropriate test. Of course, we want to have a *robust* and *powerful* test. But it is well known that most parametric (classical) tests for comparing variances are extremely sensitive to the normality assumption. At the same time there are many nonparametric tests that do not depend on sample distribution. But in terms of power all parametric tests have an advantage; they are *always significantly* more powerful than nonparametric ones.

So, if you want to test the hypothesis of variances equality, you will have to choose between robustness and power. That is why we propose a new approach to testing homogeneity of variances that will help us to avoid problems with the validity of classical tests. In this case the p-value is calculated using a simulated empirical distribution in real-time testing the hypothesis. Then we only should know what test is the most powerful in the particular situation.

Thus, the purpose of this study is to:

- give a possibility to correctly apply *parametric* tests when the normality assumption may not be true;
- give recommendations on choosing the most powerful test for a particular form of data distribution;
- give recommendations on choosing the best test in terms of robustness in the case of using a software that do not provides simulation.

1 Description of tests studied

A great number of tests for the variances homogeneity have been proposed and examined in statistical literature. But so far, the most frequently cited and used methods have been the F-test, Bartlett's, Cochran's, Hartley's and Levene's tests. Below we give a description of these tests.

F-test. Let S_1^2 and S_2^2 denote the variance estimates of samples with sizes n_1 and n_2 respectively, the classical F-test utilizes the following test statistic:

$$F = S_1^2/S_2^2.$$

This test statistic has the F_{n_1-1,n_2-1}-distribution if the null hypothesis of variances equality is true. The null hypothesis is rejected if the statistic F is either too large or too small.

Bartlett's test was essentialy a generalization of the F-test to the several $k > 2$ populations case (Bartlett[4]). The test statistic B involves a comparison of the separate within-group sums-of-squares to the pooled within-group sum-of-squares: $B = (N-k)\ln S_p^2 - \sum_{i=1}^{k}(n_i-1)\ln S_i^2$, where $N = \sum_{i=1}^{k} n_i$, $S_p^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i-1)S_i^2$ is the pooled estimate for the variance. The correction factor C_B is calculated as: $C_B = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i-1} - \frac{1}{N-k}\right)$ and applied to B to obtain a corrected B_C statistic:

$$B_C = B/C_B.$$

If hypothesis H_0 is true and samples are normally distributed, the statistic B_C has approximately the χ_{k-1}^2 distribution.

Cochran's test. The test introduced by Cochran[5] was considerably easier to compute than the tests used at that time:

$$C = S_{max}^2/\left(S_1^2 + S_2^2 + \ldots + S_k^2\right).$$

Unfortunately, the distribution of Cochran's test statistic depends on the sample size. Tables of critical values for some combinations of the sample sizes n and the number of groups k have been presented by different authors. If the test statistic C is more than the critical value, the null hypothesis H_0 is rejected.

Cochran's test seems to be the best method to detect cases when the variance of one of the groups is much larger than that of the other groups.

Hartley's test. This test was proposed by Hartley[6] in 1950. It is well known as the "F-max" test and is very simple to calculate. Its test statistic is just a ratio between the largest and the smallest sample variances:

$$H = S_{\max}^2/S_{\min}^2.$$

It should be noted that Hartley's test resembles Cochran's test with a less optimal use of the information available. One can find in the literature tables of critical values created by Hartley. The tables evaluates the test statistic with degrees of freedom k and $n-1$ (if $n_1 = n_2 = \ldots = n_k = n$). The hypothesis H_0 should be rejected if the test statistic H is large.

Levene's test. In 1960, Levene[7] proposed using the one-way ANOVA F statistic on the variables $Z_{ij} = |X_{ij} - \bar{X}_i|$ as a method for testing equality of variances. The test statistic is given by:

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

where $\bar{X}_i$ is the estimated mean of the ith sample, $\bar{Z}_i$ is the mean of Z_{ij} for ith sample and $\bar{Z}$ is the overall mean of the Z_{ij}.

In some descriptions of this test it is said that the statistic L has a $F_{k-1,N-k}$-distribution. Actually, *distribution of Levene's test statistic is not Fisher's distribution*! If sample sizes are less than 40, the distribution of the statistic differs greatly from Fisher's one. We must take this into account when using this test.

Levene's test is less sensitive to departures from normality as compared to other classical tests. However, it is less powerful.

Modified Levene's test. Miller[8] showed that ANOVA on Levene's variables $|X_{ij} - \bar{X}_i|$ will be asymptotically incorrect if the population means are not equal to the population medians (essentially requiring symmetry). Brown and Forsythe[9] suggested using the sample median instead of the mean in computing Z_{ij} in the Levene's test statistic. That is $Z_{ij} = |X_{ij} - \tilde{X}_i|$, where $\tilde{X}_i$ is the median of the ith sample. This modification allows us to overcome the above problem by centering the variables.

For this study we have also chosen a group of tests that are referred to as the most powerful ones in recent publications. These are Neyman-Pearson's, Z-variance, Overall-Woodward modified Z-variance and O'Brien tests.

Neyman-Pearson's test. The test statistic is defined as the ratio of arithmetic mean to the geometric mean of variance estimates:

$$P = \left(\frac{1}{k} \sum_{i=1}^{k} S_i^2 \right) / \left(\prod_{i=1}^{k} S_i^2 \right)^{\frac{1}{k}}.$$

Th null hypothesis H_0 should be rejected, if $P > P_{\alpha,n,k}$, where $P_{\alpha,n,k}$ is a critical value of this test.

Z-variance test. A normal deviation transformation is used to obtain Z-score equivalents of the sample variance. The test statistic proposed by Overall and Woodward[10] is:

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

where $Z_i = \sqrt{(c_i(n_i-1)S_i^2)/MSE} - \sqrt{c_i(n_i-1)-c_i/2}$, $c_i = 2 + 1/n_i$, $MSE = \left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)^2\right)/(N-k)$, $N = \sum_{i=1}^{k} n_i$ - total sample size.

If samples are normally distributed and the null hypothesis is true, the statistic V does not depend on the sample size and has approximately the $F_{k-1,\infty}$-distribution.

Overall-Woodward modified Z-variance test. As other classical tests the Z-variance test is extremely sensitive to the normality assumption. So, Overall and Woodward[11] conducted a series of studies to determine a *c value* so that variances of Z_i would remain stable when samples are not normally distributed. Using regression they have found a *c value* based on the sample size and the kurtosis.

The new *c value* is calculated by: $c_i = 2.0\left(\frac{2.9+0.2/n_i}{\bar{K}}\right)^{1.6(n_i-1.8K+14.7)/n_i}$, where $\bar{K}$ is the mean of the kurtosis indices for all samples.

The kurtosis index used by Overall and Woodward is:
$K_i = \left(\sum_{j=1}^{n_i} G_{ij}^4\right)/(n_i-2)$, where $G_{ij} = \left(X_{ij}-\bar{X}_i\right)/\sqrt{S_i^2(n_i-1)/n_i}$.

Our study has shown that this test remains stable for distributions with different kurtosis indices. However, it is not true for skewness indices.

O'Brien test. O'Brien[12] has claimed that his test is a general method that does fairly well for behavioral science data. He also states that this ANOVA-based test is robust to departures from normality.

The O'Brien test statistic is calculated in the followwing way. First, every raw value X_{ij} is transformed using the formula:
$V_{ij} = \left((n_i-1.5)n_i(X_{ij}-\bar{X}_i)^2 - 0.5S_i^2(n_i-1)\right)/\left((n_i-1)(n_i-2)\right)$.

After this transformation the mean of V-values will be equal to the variance for original values, that is $\bar{V}_i = \left(\sum_{j=1}^{n_i} V_{ij}\right)/n_i = S_i^2$.

The O'Brien test statistic will be the F-value computed applying the usual ANOVA procedure on the transformed values V_{ij}. If the null hypothesis of equal variances is true, the test statistic has approximately the $F_{k-1,N-k}$-distribution.

2 Design of simulation study

All tests described above were compared in terms of robustness and power using Monte Carlo studies. The tests studied were (1) F-test, (2) Bartlett's, (3) Cochran's, (4) Hartley's (F-max), (5) Levene's, (6) modified Levene's, (7) Neyman-Pearson's, (8) Z-variance, (9) Overall-Woodward modified Z-variance and (10) O'Brien tests.

To know how the form of sample distribution influences the performance of tests we have taken five types of distributions of various forms: (1) normal, (2) leptokurtic, (3) platykurtic, (4) moderately skewed and (5) highly skewed.

We have used the exponential family of distributions with the density $De(\theta_0) = f(x;\theta_0,\theta_1,\theta_2) = \theta_0/\left(2\theta_1\Gamma(1/\theta_0)\right) exp\left(-\left(|x-\theta_2|/\theta_1\right)^{\theta_0}\right)$ to approximate symmetric distributions. The Laplace distribution ($De(1)$) and the distribution with $\theta_0 = 3$ ($De(3)$) were taken as leptokurtic and platykurtic distributions respectively.

The chi-squared distributions with 6 and 3 degrees of freedom (χ^2_6 and χ^2_3) have been chosen for moderately and highly skewed distributions respectively.

To investigate statistics distributions, to calculate percentage points and to estimate the power of tests we used statistical simulation methods and the software developed. Each test statistic was computed 1 000 000 times. Such a value gives a simulation accuracy of 0.001.

To estimate the tests power, we need to simulate statisitics distribution when the alternative hypothesis H_1 is true. For this purpose we set several competing hypotheses by manipulating the value of the parameter $r = \sigma^2_{max}/\sigma^2_{min}$. The larger r is, the more the corresponding populations depart from the hypothesis of equal variances, i.e. the distance between competing hypotheses is larger. A smaller distance makes it more difficult to detect differences in variances. We have considered different distances between competing hypotheses: small ($r = 1.1$), moderate ($r = 1.2$) and large ($r = 1.5$). According to Wludyka and Nelson[13] three basic variance configurations were studied: (1) $k-1$ variances are equal, the last variance is larger, (2) $k-1$ variances are equal, the last variance is smaller, (3) $k-2$ variances are equal, the first variance is smaller and the last variance is larger.

3 Simulation study results

The study of classical tests power has shown that F-test, Z-variance, Bartlett's, Cochran's, Hartley's, Neyman-Pearson's and O'Brien tests have *equal* and the highest power for two normal samples while the power of Levene's test is much less. This is also true for platykurtic distributions. But for leptokurtic and skewed distributions Levene's test is more powerful than the other procedures. Furthermore, the modified Levene's test outperformed the original test in this case.

Bartlett's, Cochran's, Hartley's, Levene's, Neyman-Pearson's, O'Brien, Z-variance and modified Z-variance tests can be applied when the number of samples $k > 2$. In such situations the power of these tests is different. If the normality assumption is true, these tests can be ordered according to the power decrease in the following way:

Cochran's $\succ$ O'Brien $\succ$ Z-variance $\succ$ Bartlett's, Neyman-Pearson's $\succ$ modified Z-variance $\succ$ Hartley's $\succ$ Levene's, modified Levene's.

The tests preference remains the same for platykurtic distributions. When samples are from leptokurtic or skewed distributions, this preference order changes. Now Levene's test has a greater power with the modified Levene's test being more powerful than the original one.

It has been mentioned earlier that Cochran's test is the best method to detect cases when the variance of one of the groups is much larger than that of the other groups (configuration (1) in this study). However, if the variance configuration differs from the aforenamed one, the power of Cochran's test decreases significantly. So, in such situations we should prefer O'Brien, Z-variance, Bartlett's or Neyman-Pearson's tests.

Based on the results obtained we have chosen Cochran's test and have compiled tables of upper percentage points for some non-normal symmetric distributions. These values can be used in situations when distribution from exponential family $De(\theta_0)$ with the appropriate parameter θ_0 is a good model for observable random variables.

The study of tests robustness has shown that Levene's, modified Levene's, Overall-Woodward modified Z-variance and O'Brien tests are the best ones.

Let us formulate recommendations on choosing the appropriate test for comparing variances taking into account all results obtained:

- If there is every reason to consider data distribution as symmetric with excess kurtosis equal or less than zero, i.e. mesokurtic or platykurtic, the best choice will be *O'Brien* or *Z-variance* tests. Here *Cochran's test* could be recommended only for situations when one variance is larger than others;
- If the data distribution is leptokurtic (excess kurtosis is positive, tails are heavy) or skewed, *modified Levene's test* should be chosen.

4 Software for testing hypotheses

It is impossible to develop distribution models for all distributions and sample sizes. So, we have developed the software that allows us to correctly apply tests for comparing variances when samples are from any distributions. We can choose any distribution from the list and simulate a distribution of the statistic. Also, we can set a required accuracy of simulation by defining the size of a statistics sample. Then a p-value is calculated using a simulated empirical distribution.

The simulation process is done using parallel computing, so the speed of simulation depends on the number of CPU cores and the required accuracy. It can be claimed that it takes not much time to make a *correct* decision when testing the hypothesis of equal variances.

Tables of critical values for Cochran's test and the latest version of the software can be obtained from the authors upon request.

Acknowledgements

This research was supported by the Russian Foundation for Basic Research (project no. 09-01-00056a), by the Federal Agency for Education within the framework of the analytical domestic target program "Development of the Scientific Potential of Higher Schools" and federal target program of the Ministry of Education and Science of the Russian Federation "Scientific and Scientific-Pedagogical Personnel of Innovative Russia".

References

1. Zar, J. H., *Biostatistical Analysis*, Fourth Edition: Upper Saddle River, N.J., Prentice Hall (1999).
2. Legendre, P., and Borcard, D., "Statistical Comparison of Univariate Tests of Homogeneity of Variances", *Journal of Statistical Computation and Simulation*, 514 (2008).
3. Boos, D.D., and Brownie, C., "Comparing Variances and Other Measures of Dispersion", *Statistical Science*, 19, No.4, 571–578 (2004).
4. Bartlett, M.S. "Properties of sufficiency of statistical tests", *Proc. Roy. Soc*, 160, 268–287 (1937).
5. Cochran, W.G.,"The distribution of the largest of a set of estimated variances as a fraction of their total", *Annals of Eugenics*, 11, 47–52 (1941).
6. Hartley, H.O.,"The maximum F-ratio as a short-cut test of heterogeneity of variance", *Biometrika*, 37, 308–312 (1950).
7. Levene, H. "Robust tests for the equality of variances", *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, 278–292 (1960).
8. Miller, R.G. "Jackknifing variances", *Ann. Math. Statist.*, 39, 567–582 (1968).
9. Brown, M.B., and Forsythe, A.B."Robust tests for the equality of variances", *J. Amer. Statist. Assoc.*, 69, 364–367 (1974).
10. Overall, J.E., and Woodward, J.A., "A simple test for heterogeneity of variance in complex factorial design", *Psychometrika*, 39, 311–318 (1974).
11. Overall, J.E., and Woodward, J.A., "A robust and powerful test for heterogeneity of variance", *University of Texas Medical Branch Psychometric Laboratory*, (1976).
12. O'Brien, R.G.,"Robust techniques for testing heterogeneity of variance effects in factorial designs", *Psychometrika*, 43, 327–342 (1978).
13. Wludyka, P.S., and Nelson, P.S., "Two non-parametric, analysis-of-means-type tests for homogeneity of variances", *Journal of Applied Statistics*, 26, No.2, 243–256 (1999).