Application of Variance Homogeneity Tests Under Violation of Normality Assumption

Alisa A. Gorbunova, Boris Yu. Lemeshko
Novosibirsk State Technical University
Novosibirsk, Russia
e-mail: gorbunova.alisa@gmail.com

Abstract

Classical tests for homogeneity of variances (Fisher's, Bartlett's, Cochran's, Hartley's, Neyman-Pearson's, Levene's, modified Levene's, Z-variance, Overall-Woodward modified Z-variance, O'Brien tests) and nonparametric tests (Ansari-Bradley's, Mood's, Siegel-Tukey's, Capon's and Klotz's tests) have been considered. Distributions of classical tests statistics have been investigated under violation of assumption that samples are normally distributed. The comparative analysis of power of classical tests with power of nonparametric tests has been carried out. Tables of percentage points for Cochran's test have been made for distributions which are different from normal. Software, that allows us to apply tests correctly, has been developed.

Keywords: homogeneity of variance test, power of test.

Introduction

Testing for samples homogeneity is frequently of interest in a number of research areas. The question can be about homogeneity of samples distributions, population means or variances. Of course, conclusions in full measure can be made in the first case. However, researcher can be interested in possible deviations in the sample mean values or differences in variances of measurements.

One of the basic assumptions to formulate classical tests for comparing variances is normal distribution of samples. It is well known, that classical tests are very sensitive to departures from normality. Therefore, the application of classical criteria always involves the question of how valid the obtained results are in this particular situation.

In this work classical Bartlett's, Cochran's, Fisher's, Hartley's, O'Brien, Neyman-Pearson's, Levene's, modified Levene's, Z-variance, Overall-Woodward modified Z-variance tests [1, 2] are compared, nonparametric (rank) Ansari-Bradley's, Mood's, Siegel-Tukey's, Capon's and Klotz's tests [1] are considered.

The purpose of our study was to:

- investigate distributions of the statistics for several tests when samples are not normally distributed;
- make a comparative analysis of the criteria power for concrete competing hypotheses;

- give a possibility to apply classical tests when the normality assumption may not be true.

A null hypothesis of equal variances for m samples is $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2$ and the alternative hypothesis is $H_1 : \sigma_i^2 \neq \sigma_j^2$, where the inequality holds at least for one pair of subscripts i, j.

Statistical simulation methods and developed software were used to investigate statistics distributions, to calculate percentage points and to estimate tests power for different competing hypotheses. Each test statistic was computed $N = 10^6$ times. In this case an absolute value of the difference between the true law of statistics distribution and a simulated empirical distribution does not exceed 10^{-3}.

Distributions of the statistics were investigated using various distributions, in particular, in the case when simulated samples are in the family of distributions with the density:

$$De(\theta_0) = f(x; \theta_0, \theta_1, \theta_2) = \frac{\theta_0}{2\theta_1 \Gamma(1/\theta_0)} exp\left(-\left(\frac{|x-\theta_2|}{\theta_1}\right)^{\theta_0}\right) \tag{1}$$

using different values of the shape parameter θ_0. This family can be a good model for error distributions of many measuring systems. Special cases of the family $De(\theta_0)$ are the Laplace ($\theta_0 = 1$) and the normal ($\theta_0 = 2$) distributions. This family makes it possible to set various symmetric distributions that differ from the normal distribution. That is a smaller value of the shape parameter θ_0 leads to a "heavier" tails of the distribution.

We also consider chi-square distributions ($df = 6, df = 5$) to approximate skewed distributions where the chi-square distribution with 6 degrees of freedom is less skewed than the one with 5 degrees of freedom.

In the comparative analysis of the tests power we consider the competing hypotheses of the form $H_1 : \sigma_2 = d\sigma_1$ ($d \neq 1$). Some tests can be applied when number of samples is more then two. For these tests we consider hypotheses when different number of samples have another variance.

1 Classical tests of variances homogeneity

1.1 Bartlett's test

The test statistic is:

$$T = M\left(1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{\nu_i} - \frac{1}{N}\right)\right)^{-1}, \tag{2}$$

where $M = N \ln\left(\frac{1}{N}\sum\limits_{i=1}^{k}\nu_i S_i^2\right) - \sum\limits_{i=1}^{k}\nu_i \ln S_i^2$, k - the number of samples, $\nu_i = n_i - 1$, n_i - sample size of ith sample, $N = \sum\limits_{i=1}^{k}\nu_i$, S_i^2 - the unbiased estimate of variance for the ith sample.

If hypothesis H_0 is true, all $\nu_i > 3$ and samples are normally distributed, the statistic (2) is almost independent of the sample size and has approximately χ^2_{k-1} distribution. If samples are

not from a normal population, the distribution of the statistic depends on the sample size and differs from χ^2_{k-1}.

1.2 Cochran's test

The Cochran's test is defined as following:

$$C = \frac{S^2_{\max}}{S_1^2 + S_2^2 + \cdots + S_k^2}, \tag{3}$$

where $S^2_{\max} = \max(S_1^2, S_2^2, \ldots, S_k^2)$, S_i^2 - the unbiased estimate of variance for the ith sample, k - the number of samples.

Distribution of Cochran's test statistic depends on the sample size. The reference litetature gives tables with percentage points for limited number of values n, that are used in hypothesis testing. If the test statistic (3) exceeds the critical value, the null hypothesis is rejected.

1.3 Fisher's test

Fisher's test is used to test hypothesis of variances homogeneity for *two* samples with sample sizes n_1 and n_2. The test statistic has a simple form:

$$F = \frac{S_1^2}{S_2^2}, \tag{4}$$

where S_1^2 and S_2^2 - the unbiased sample variances.

If samples are normally distributed and hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ is true, statistic (4) has the F_{n_1-1,n_2-1}-distribution.

1.4 Hartley's test

Hartley's test is very simple to calculate. Its test statistic is just a ratio between the largest sample variance and the smallest:

$$H = \frac{S^2_{\max}}{S^2_{\min}} \tag{5}$$

where $S^2_{\max} = \max(S_1^2, S_2^2, \ldots, S_k^2)$, $S^2_{\min} = \min(S_1^2, S_2^2, \ldots, S_k^2)$, S_i^2 - the unbiased estimate of variance for the ith sample, k - the number of samples.

One can find in literature table of values created by Hartley. This table evaluates the test statistic with degrees of freedom k and $n-1$ (if $n_1 = n_2 = \cdots = n_k = n$). Reject H_0 if the test statistic (5) is more than critical value, otherwise do not reject H_0.

1.5 Neyman-Pearson's test

The test statistic is defined as ratio between arithmetic mean and geometric mean of variance estimates:

$$P = \frac{\frac{1}{k}\sum_{i=1}^{k} S_i^2}{\left(\prod_{i=1}^{k} S_i^2\right)^{\frac{1}{k}}}, \tag{6}$$

where k - the number of samples.

1.6 Levene's test

If X_{ij}'s represent the raw scores, Levene's test statistic is defined as:

$$L = \frac{(N-k)\sum_{i=1}^{k} n_i \left(\bar{Z}_i - \bar{Z}\right)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(Z_{ij} - \bar{Z}_i\right)^2}, \tag{7}$$

where k - the number of samples, n_i - sample size of ith sample, $N = \sum_{i=1}^{k} n_i$ - total sample size, $Z_{ij} = |X_{ij} - \bar{X}_i|$, $\bar{X}_i$ - the mean of the ith sample, $\bar{Z}_i$ - the mean of Z_{ij} for ith sample, $\bar{Z}$ - the overall mean of the Z_{ij}.

In some descriptions of this test it is said that statistic (7) 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 grom Fisher's. We must take it into account when using this test.

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

1.7 Modified Levene's test

The modified Levene's test is nearly identical to the original Levene's test. Brown and Forsythe suggested using the sample median instead of the mean in computing Z_{ij}. That is $Z_{ij} = |X_{ij} - \tilde{X}_i|$, where $\tilde{X}_i$ - the median of the ith sample.

This test is more robust than original Levene's test.

1.8 Z-variance test

The test statistic is:

$$V = \frac{\sum\limits_{i=1}^{k} Z_i^2}{k-1}, \tag{8}$$

where $Z_i = \sqrt{\frac{c_i(n_i-1)S_i^2}{MSE}} - \sqrt{c_i(n_i-1) - \frac{c_i}{2}}, MSE = \frac{\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)^2}{N-k}$, k - the number of samples, $c_i = 2 + \frac{1}{n_i}$, n_i - sample size for ith sample, S_i^2 - the unbiased estimate of variance for ith sample, $N = \sum\limits_{i=1}^{k} n_i$ - total sample size, X_{ij} - jth observation in ith sample, $\bar{X}_i$ - the mean of ith sample.

If samples are normally distributed and null hypothesis is true, statistic (8) does not depend on sample size and has approximately $F_{k-1,\infty}$-distribution.

1.9 Overall-Woodward modified Z-variance test

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

The new c value is evaluated as following:

$$c_i = 2.0\left(\frac{2.9 + \frac{0.2}{n_i}}{\bar{K}}\right)^{\frac{1.6(n_i-1.8K+14.7)}{n_i}},$$

where n_i - sample size of the ith sample, $\bar{K}$ - the mean of the kurtosis indices for all samples.

The index of kurtosis is $K_i = \frac{\sum\limits_{j=1}^{n_i} G_{ij}^4}{n_i-2}$, where $G_{ij} = \frac{X_{ij}-\bar{X}_i}{\sqrt{\frac{n_i-1}{n_i}S_i^2}}$.

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

1.10 O'Brien test

Every raw score X_{ij} is transformed using the following formula:

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

where n_i - sample size for ith sample, $\bar{X}_i$ - the mean of ith sample, S_i^2 - the unbiased estimate of variance for ith sample.

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

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

2 Nonparametric (rank) tests

Nonparametric analogues of variance homogeneity tests are used to test hypotheses that *two* samples with sample sizes n and m are from population with equal dispersion characteristics. To calculate test statistic we use ranks instead of sample values.

2.1 Ansari-Bradley's test

Tha Ansari-Bradley's test statistic is:

$$A = \sum_{i=1}^{m} \left(\frac{m+n+1}{2} - \left| R_i - \frac{m+n+1}{2} \right| \right), \tag{9}$$

where m, n - sample sizes ($m \leq n$), R_i - rank of ith value of sample with sample size m in general variational row.

Discreteness of statistics distribution can be practically neglected when $m, n > 40$.

2.2 Mood's test

The test statistic is defined as following:

$$M = \sum_{i=1}^{m} \left(R_i - \frac{n+m+1}{2} \right)^2, \tag{10}$$

where m, n - sample sizes ($m \leq n$), R_i - rank of ith value of sample with sample size m in general variational row.

Discreteness of statistics distribution can be practically neglected when $m, n > 20$.

2.3 Siegel-Tukey's test

The general variational row $X_1 \leq X_2 \leq \cdots \leq X_N$ ($N = n + m$) is transformed into sequence:

$$X_1, X_N, X_{N-1}, X_2, X_3, X_{N-2}, X_{N-3}, X_4, X_5, \ldots$$

When $m \leq n$ the test statistic is defined as:

$$W = \sum_{i=1}^{m} R_i, \tag{11}$$

where R_i - rank of ith value of sample with sample size m in transformed row.

Discreteness of statistics distribution can be practically neglected when $m, n > 30$.

2.4 Capon's test

Capon's test statistic is:

$$K = \sum_{i=1}^{m} a_{m+n}(R_i), \tag{12}$$

where m, n - sample sizes ($m \leq n$), R_i - rank of ith value of sample with sample size m in general variational row, $a_i(j)$ - the mean value of square of jth order statistic in sample with sample size i from standard normal distribution.

2.5 Klotz's test

The test statistic is defined as:

$$Q = \sum_{i=1}^{m} u^2_{\frac{R_i}{m+n+1}}, \tag{13}$$

where m, n - sample sizes ($m \leq n$), R_i - rank of ith value of sample with sample size m in general variational row, u_γ - γ-quantile of standard normal distribution.

3 Comparative analysis of power

At the given probability of a type I error α (to reject the null hypothesis when it is true) it is possible to judge about the advantages of the test by the value of power $1 - \beta$, where β - probability of type II error (not to reject the null hypothesis when alternative is true).

The study of power of classical tests for several competing hypotheses $H_1 : \sigma_2 = d\sigma_1$ ($d \neq 1$) has shown that Bartlett's, Cochran's, Hartley's, Fisher's, Neyman-Pearson's and Z-variance tests have equal power for two normal samples and Levene's test power is much less in this case.

As for non-normal distributions, for example, family of distributions with density (1), Bartlett's, Cochran's, Hartley's, Fisher's, Neyman-Pearson's and Z-variance tests remain equal in power, and Levene's test power is also much less. However, for heavy-tailed (for example, the Laplace distribution) and skewed distributions Levene's test is more powerful than the others. Furthermore 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 number of samples $k > 2$. In such situations the power

of these tests is different. If $k > 2$ and normality assumption is true, these tests can be ordered according to the decrease of power in the following way:

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

The preference order also remains in case of violation of a normality assumption. When samples are from heavy-tailed or skewed distributions, this preference order changes. For example, in the case of the Laplace distribution Levene's test has a greater power. Also modified Levene's test is more powerful than the original one.

However if number of samples with smaller value of variance is less than number of samples with greater value, power of Cochran's test significantly goes down. So in this case we should prefer O'Brien, Z-variance, Bartlett's or Neyman-Pearson's test.

The study of the nonparametric criteria power has shown that Mood's test power is the highest. And other nonparametric tests, as Siegel-Tukey's, Ansari-Bradley's, Capon's and Klotz's have practically equal power. But for skewed distributions all nonparametric tests are biased (power of test is less than significance level).

4 Cochran's test for non-normal distributions

The main and valid reason for using nonparametric tests is based on the fact that these test statistics are distribution-free. But this is true if both samples are from the same population. If samples are not identically distributed, *nonparametric tests depend on both sample laws and even the order in which these laws are used*.

Also classical tests have a great advantage in power over nonparametric tests. This advantage remains when samples are not normally distributed. Therefore, there is every reason to study distributions of *classical* criteria for testing variances homogeneity. To study distributions means to develop distribution models or tables of percentage points. It should be done for non-normal distributions mostly used in practice. Among the tests studied Cochran's test seems to be the most suitable for this purpose.

Tables of upper percentage points (1%, 5%, 10%) for Cochran's test were made using statistical simulation for the number of samples $m = 2 \div 5$ when simulated samples were taken from an exponential family of distributions (1) with shape parameter $\theta_0 = 1, 2, 3, 4, 5$. The results obtained can be used in situations when distribution from an exponential family (1) with an appropriate parameter θ_0 is a good model for the observed variables. Computed percentage points expand possibilities to apply Cochran's test.

5 Software for testing hypotheses

It is impossible to develop distribution models for all distributions and sample sizes. So we have developed 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 any size of simulated samples of statistics according to required precision.

Then we define a p-value using simulated statistic distribution. Simulation process is done using parallel computing, so speed of simulation depends on number of CPU cores and takes not much time to make correct decision when testing the hypothesis of equal variances.

References

[1] Kobzar A.I. (2006). *Applied mathematical statistics. For engineers and scientists* (in Russian). FIZMATLIT Publishing House, Moscow.

[2] Lee H.B., Katz G.S., Restori A.F. (2010). A Monte Carlo study of seven homogeneity of variance tests. *Journal of Mathematics and Statistics*. Vol. **6(3)**, pp. 359-366.