Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

04-12

Testing Equality of Variances for Multiple Univariate Normal Populations

David Allingham and J.C.W. Rayner

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Testing Equality of Variances for Multiple Univariate Normal Populations

David Allingham
Centre for Computer-Assisted Research Mathematics and its Applications
School of Mathematical and Physical Sciences,
University of Newcastle, NSW 2308, Australia
David.Allingham@newcastle.edu.au

and

J.C.W. Rayner
Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522, Australia *and*
School of Mathematical and Physical Sciences, University of Newcastle, NSW 2308, Australia
John.Rayner@newcastle.edu.au

Abstract

To test for equality of variances in independent random samples from multiple univariate normal populations, the test of first choice would usually be the likelihood ratio test, the Bartlett test. This test is known to be powerful when normality can be assumed.

Here two Wald tests of equality of variances are derived. The first test compares every variance with every other variance and was announced in Mather and Rayner (2002), but no proof was given there. The second test is derived from a quite different model using orthogonal contrasts, but is identical to the first. This second test statistic is similar to one given in Rippon and Rayner (2010), for which no empirical assessment has been given. These tests are compared with the Bartlett test in size and power.

The Bartlett test is known to be non-robust to the normality assumption, as is the orthogonal contrasts test. To deal with this difficulty an analogue of the new test is given. An indicative empirical assessment shows that it is more robust that the Bartlett test and competitive with the Levene test in its robustness to fat-tailed distributions. Moreover it is a Wald test and has good power properties in large samples. Advice is given on how to implement the new test.

Key Words: Bartlett's test; Levene test; Orthogonal and non-orthogonal contrasts; Wald tests.

1. Introduction

Suppose we have m independent random samples, with the jth being of size n_j and being from a normal $N(\mu_j, \sigma_j^2)$ population, for $j = 1, \ldots, m$. The total sample size is $n = n_1 + \ldots + n_m$. We seek to test equality of variances: H: $\sigma_1^2 = \ldots = \sigma_m^2 = \sigma^2$, say, against the alternative K: not H. Probably the most popular choices for testing H against K are the parametric Bartlett test, which is known to be not robust to departures from normality, and the nonparametric Levene test, which is known to be robust but is less powerful than Bartlett's test when the data are approximately normal.

In the case of just two independent random samples the likelihood ratio test is equivalent to the F test based on the quotient of the (unbiased) sample variances. Using the Wald test, Rayner (1997) derived a test based on the difference of the sample variances. This test, and a robust analogue of it, were discussed in Allingham and Rayner (2011).

For the *m*-sample problem Mather and Rayner (2002) gave, without proof, a new test statistic. Here a relatively succinct derivation is given. Using the same approach a new competitor test, based on orthogonal contrasts, is derived. It is not immediately obvious, but the two tests are identical.

In a small empirical study we show that for the configurations chosen here the orthogonal contrasts test is slightly inferior to, but certainly competitive with, the Bartlett test in both size and power. This is consistent with the fact that likelihood ratio and Wald tests are asymptotically equivalent.

In the case $m = 2$, Allingham and Rayner (2011) showed that the test derived here is, like the Bartlett test, sensitive to departures from normality. However it is possible to give an analogue of the test that is robust to non-normality. We argue that this test is a Wald test and hence has good power in large samples. A similar analogue of the Bartlett test is not apparent.

For convenience the Bartlett test statistic B is given here. Define

$$B = \frac{(n-m)\log S^2 - \sum_{j=1}^{m}(n_j - 1)\log S_j^2}{1 + \frac{1}{3(m-1)}\sum_{j=1}^{m}\left(\frac{1}{n_j - 1} - \frac{1}{n-m}\right)},$$

in which S_j^2 is the unbiased sample variance from the jth sample, $j = 1, \ldots, m$ and $S^2 = \sum_j (n_j - 1)S_j^2/(n-m)$ is the pooled sample variance. The test statistic B has asymptotic distribution χ^2_{m-1}. Common practice when normality is in doubt is to use Levene's test. This is based on the ANOVA F test applied to the sample residuals. There are different versions of Levene's test using different definitions of residual. The version employed here uses the group means, $\left|X_{ij} - \overline{X}_{i.}\right|$, in an obvious notation. The distribution of the test statistic, L, say, is approximately $F_{m-1, n_1+\ldots+n_m-m}$.

In Sections 2 and 3 derivations of the Mather and Rayner (2002) and the orthogonal contrasts tests are given. A brief empirical assessment of these tests is given in Section 4. A robust analogue of the orthogonal contrasts test is given in Section 5. An example is given in Section 6, while section 7 gives by a brief conclusion, including advice on how to implement the recommended robust test.

2. The Mather-Rayner *m*-Sample Test

To test H against the alternative K first define an $(m - 1) \times m$ contrast matrix C. Essentially, the rows of C are used to define contrasts between the population variances σ_j^2. These contrasts must be specified before sighting the data. If we define

$\phi = (\sigma_1^2, \ldots, \sigma_m^2)^T$, then the null hypothesis of equality of variances is equivalent to $\theta = C\phi = 0$. In this section we take $(C)_{ii} = 1$ and $(C)_{i(i+1)} = -1$, $i = 1, \ldots, m-1$, with all other elements zero. The contrasts are between successive variances: $\sigma_1^2 - \sigma_2^2$, $\sigma_2^2 - \sigma_3^2, \ldots, \sigma_{m-1}^2 - \sigma_m^2$. Clearly there are many other possible choices of contrast matrix.

If, as above, S_j^2 is the unbiased sample variance from the jth sample, then put $\hat{\phi} = (S_j^2)$ and $\hat{\theta} = C\hat{\phi}$. A Wald test of H against K may be based on $\hat{\theta}^T \hat{\text{cov}}^{-1}(\hat{\theta})\,\hat{\theta}$. Using the Rao-Blackwell theorem as in Rayner (1997), the covariance matrix of $\hat{\phi}$, $\text{cov}(\hat{\phi}) = \text{diag}(2\sigma_j^4/(n_j - 1))$, may be optimally estimated by $D = \text{diag}(d_j)$, where $d_j = 2S_j^4/(n_j + 1)$ for $j = 1, \ldots, m$. Hence the covariance matrix of $\hat{\theta}$ may be optimally estimated by $C\,\text{diag}(d_j)\,C^T$, which for this choice of C is

$$\begin{pmatrix} d_1 + d_2 & -d_2 & \cdots & 0 \\ -d_2 & d_2 + d_3 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{m-1} + d_m \end{pmatrix}.$$

Note that only variances and adjacent elements of $\hat{\theta}$ have non-zero covariance.

Mather and Rayner (2002) gave the test statistic in terms of contrasts between all the sample variances:

$$T_{MR} = \frac{\sum_{r=1}^{m-1}\sum_{s>r}^{m}(n_r + 1)(n_s + 1)(S_r^2 - S_s^2)^2/(S_r^4 S_s^4)}{2\sum_{r=1}^{m}(n_r + 1)/S_r^4}. \quad (2.1)$$

They noted that the derivation required the inversion of the tri-diagonal matrix, and in fact the inverse was calculated for some small values of m, an inverse guessed and the result proven by a lengthy induction. For the two-sample problem a simple calculation shows that $\hat{\theta} = S_1^2 - S_2^2$ and $\text{var}(\hat{\theta}) = d_1 + d_2$, agreeing with the result in Rayner (1997). A relatively succinct derivation of T_{MR} is given in the Appendix.

3 A Wald Test Using Orthogonal Contrasts

The complexity of the derivation of the Wald test can be reduced by requiring that the rows of C be orthogonal. First, define $C^* = (C^T|u)^T$ to be an $m \times m$ orthogonal matrix. Should the null hypothesis be true, the common variance could be estimated by the pooled sample variance, $S^2 = \sum_j (n_j - 1)S_j^2/(n - m)$. Put $w_j = (n_j - 1)/(n - m)$ for $j = 1, \ldots, m$ and note that $\sum_j w_j = 1$, since $\sum_j (n_j - 1) = n - m$. Now, define $\sigma^2 = \sum_j w_j \sigma_j^2$, $\phi = (\sigma_j^2 \sqrt{w_j})$, $u = (\sqrt{w_j})$ and $C = I_m - uu^T$. Note that C is symmetric, idempotent, orthogonal and not of full rank. Its only specification so far, and ultimately, is that its rows are orthogonal to u.

Put $\theta = \mathrm{C}\phi = ((\sigma_j^2 - \sigma^2)\sqrt{w_j})$, which is zero if and only if $\sigma_j^2 = \sigma^2$ for all j. Hence we again wish to test H: $\theta = 0$ against K: $\theta \neq 0$, albeit for a slightly different θ than in the previous section. Here, $\hat{\theta} = ((S_j^2 - S^2)\sqrt{w_j})$ is an unbiased estimator of θ, and is asymptotically equivalent to the maximum likelihood estimator of θ. The covariance matrix of $\hat{\theta}$ is estimated by $\hat{\mathrm{cov}}(\hat{\theta})$ = CDC, where now D = diag($d_j\, w_j$) with the d_j as in the previous section. To find the inverse of $\hat{\mathrm{cov}}(\hat{\theta})$ a routine lemma is needed. The proof is omitted.

Lemma 3.1. If $\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} E & F \\ G & H \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$ and if E, F, G, H are known with H invertible, then $\mathrm{A}^{-1} = \mathrm{E} - \mathrm{F}\mathrm{H}^{-1}\mathrm{G}$.

The lemma is applied with A = $\hat{\mathrm{cov}}(\hat{\theta})$ and

$$\begin{pmatrix} E & F \\ G & H \end{pmatrix} = \mathrm{C}^{*T-1}\mathrm{D}^{-1}\mathrm{C}^{*-1} = \mathrm{C}^{*}\mathrm{D}^{-1}\mathrm{C}^{*T} = \begin{pmatrix} CD^{-1}C^T & CD^{-1}u \\ u^T D^{-1}C^T & u^T D^{-1}u \end{pmatrix},$$

so that $\mathrm{A}^{-1} = \mathrm{C}\mathrm{D}^{-1}\mathrm{C}^T - (\mathrm{C}\mathrm{D}^{-1}u)(u^T\mathrm{D}^{-1}\mathrm{C}^T)/(u^T\mathrm{D}^{-1}u)$.

A Wald test of H: $\theta = 0$ against K: $\theta \neq 0$ can be based on $T_{\mathrm{OC}} = \hat{\theta}^T\, \hat{\mathrm{cov}}^{-1}(\hat{\theta})\, \hat{\theta}$. Since $\phi = (\sigma_j^2\sqrt{w_j})$, $\hat{\phi} = (S_j^2\sqrt{w_j})$. Substituting gives

$$T_{\mathrm{OC}} = \hat{\phi}^T \mathrm{C}^T (\mathrm{C}\mathrm{D}^{-1}\mathrm{C}^T)\, \mathrm{C}\hat{\phi} - (u^T\mathrm{D}^{-1}\mathrm{C}^T\mathrm{C}\hat{\phi})^2/(u^T\mathrm{D}^{-1}u).$$

Now, since C* is orthogonal,

$$\mathrm{I}_m = \mathrm{C}^{*T}\,\mathrm{C}^{*} = \begin{pmatrix} C^T & u \end{pmatrix} \begin{pmatrix} C \\ u^T \end{pmatrix} = \mathrm{C}^T\mathrm{C} + u\,u^T,$$

giving $\mathrm{C}^T\mathrm{C} = \mathrm{I}_m - uu^T$. With $u = (\sqrt{w_1}, \ldots, \sqrt{w_m})^T$ it follows that $u^T\hat{\phi} = \sum_j w_j S_j^2 = S^2$ and $(\mathrm{I}_m - uu^T)\hat{\phi} = (\sqrt{w_j}(S_j^2 - S^2)) = \hat{\phi}_C$, say. Thus

$$\begin{aligned} T_{\mathrm{OC}} &= \hat{\phi}^T(\mathrm{I} - uu^T)\mathrm{D}^{-1}(\mathrm{I}_m - uu^T)\hat{\phi} - \{u^T\mathrm{D}^{-1}(\mathrm{I}_m - uu^T)\hat{\phi}\}^2/(u^T\mathrm{D}^{-1}u) = \\ &= \hat{\phi}_C^T\, \mathrm{D}^{-1}\, \hat{\phi}_C - (u^T\mathrm{D}^{-1}\hat{\phi}_C)^2/(u^T\mathrm{D}^{-1}u) \\ &= \sum_{j=1}^m \lambda_j (S_j^2 - S^2)^2 - \{\sum_{j=1}^m \lambda_j (S_j^2 - S^2)\}^2 / \{\sum_{j=1}^m \lambda_j\} \qquad (3.1) \end{aligned}$$

in which $\lambda_j = 1/d_j$ for $j = 1, \ldots, m$. It is not immediately obvious, but T_{OC} and T_{MR} are identical. We do not show that here.

In Rippon and Rayner (2010) orthogonal contrasts and the Moore-Penrose inverse were used to derive a Wald test based on

$$\sum_{j=1}^{m} \lambda_j (S_j^2 - S^2)^2 = T_{MP} \text{ say.}$$

Clearly T_{MP} is just T_{OC} without the final 'correction term' that can be expected to be small when the null hypothesis is true.

Note that under the null hypothesis of equality of variances, the test statistics T_{MR}, T_{MP} and T_{OC} all have asymptotic distribution χ^2_{m-1}. Moreover, these three test statistics are invariant under transformations $Y_{jk} = a(X_{jk} - b_j)$, for constants a, b_j, for $j = 1, ..., m$.

4 Empirical Assessment

Throughout this assessment we take samples of the same size, N, in each of the m populations. Thus, the total sample size, n, is given by $n = mN$.

The test sizes of the competing tests were compared for a number of combinations of the parameters involved. Figure 1 shows the proportion of rejections in 100,000 Monte Carlo simulations for nominal 5% level tests for varying sample sizes N, each population being normal with mean 0 and variance 1. The left panel compares $m = 4$ populations, the right panel $m = 8$ populations, for the tests based on B, T_{OC} and T_{MP}. The test based on T_{MP} is clearly not competitive. The form of B used here (given in Section 1) has been adjusted from the likelihood ratio test statistic to improve its type 1 error rate, and that is reflected in its excellent adherence to the nominal level in Figure 1. Interestingly the test based on T_{OC} performs far worse for $m = 8$ than it does for $m = 4$.

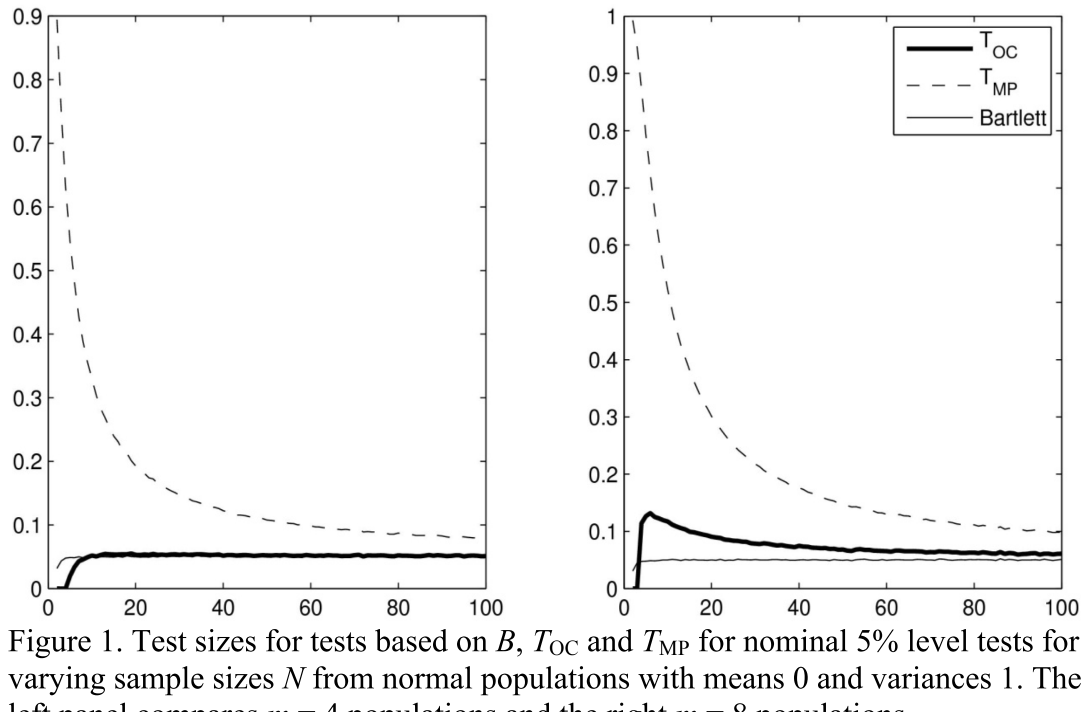

Figure 1. Test sizes for tests based on B, T_{OC} and T_{MP} for nominal 5% level tests for varying sample sizes N from normal populations with means 0 and variances 1. The left panel compares $m = 4$ populations and the right $m = 8$ populations.

The powers of the tests based on B and T_{OC} are compared in Figure 2. Critical values have been estimated so that all are 5% level tests. The test based on T_{MP} was excluded because, as noted above, in terms of the agreement between the achieved and nominal levels, it is not competitive. Each point is the proportion of rejections in 100,000 sets of data, with samples sizes of N = 10, 20 and 50 for each of m = 4 populations from the normal distributions with means 0 and population variances 1, 1 + $\Delta\sigma^2$, 1 + 2 $\Delta\sigma^2$, 1 + 3 $\Delta\sigma^2$, where $\Delta\sigma^2$ varies between 0 and 1 in steps of 0.001. Thus, at the LHS of the plot, the variances are equal and at the RHS they are 1, 2, 3 and 4.

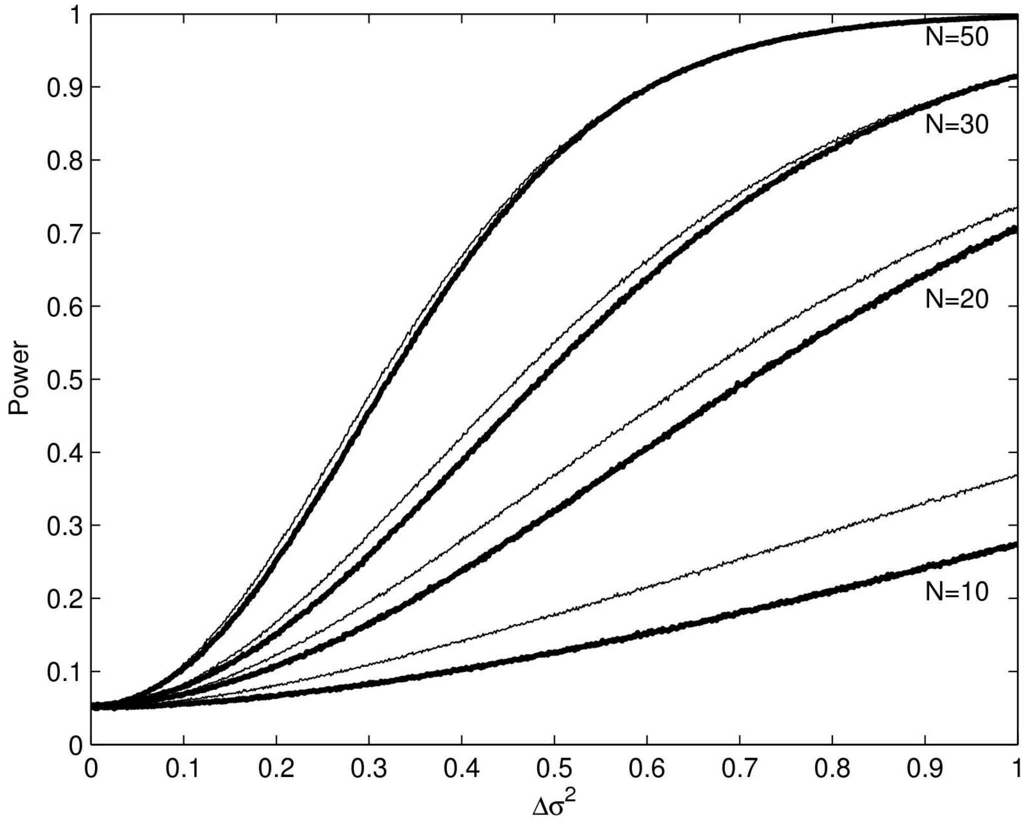

Figure 2. Power functions for the Bartlett test (thin line) and the test based on T_{OC} (thick line), based on four normal populations, a significance level of 5% and sample sizes from each population of N = 10, 20, 30 and 50.

As N increases, the powers of the two tests approach each other. This is as expected since both are asymptotically optimal tests. The parameter space is clearly multidimensional, and here we have chosen a very particular subset. In this space, for smaller sample sizes, the power of the Bartlett test is slightly superior to the test based on T_{OC}. However, across the entire sample space it could be expected that they will perform similarly, with sometimes one, and sometimes the other, being slightly superior.

5. A Robust Form of the Orthogonal Contrasts Test

In the discussion so far we have introduced tests based on T_{MR}, T_{OC} and T_{MP}. The test based on the Moore-Penrose test statistic, T_{MP}, has poor adherence to its nominal test size, and the test statistics T_{MR} and T_{OC} are, in fact, identical. The use of the test based on T_{OC} may be compromised if, like the Bartlett test, it is not robust to departures from the normality assumption. In fact, as was shown in Allingham and Rayner (2011) for the case $m = 2$ and as we shall see in Figure 3 for $m = 4$ and $m = 8$, the agreement between nominal and actual test sizes for the test based on T_{OC} is not good enough for practical use when sampling from a fat-tailed distribution.

We therefore propose an analogue of T_{OC}. In deriving that test, the variances var(S_j^2) were estimated optimally using the Rao-Blackwell theorem. The estimation depends very strongly on the assumption of normality. If normality is in doubt then var(S_j^2) can be estimated using results given, for example, in Stuart and Ord (1994). For a random sample $Y_1, ..., Y_n$ with population and sample central moments μ_r and $m_r = \sum_{j=1}^{n}(Y_j - \bar{Y})^r / n$, $r = 2, 3, ...$ respectively, Stuart and Ord (1994) gave

$$E[m_r] = \mu_r + O(n^{-1}) \text{ and } \mathrm{var}(m_2) = (\mu_4 - \mu_2^2)/n + O(n^{-2}).$$

Applying Stuart and Ord (1994, 10.5), μ_2^2 may be estimated to $O(n^{-1})$ by m_2^2, or, equivalently, by $n m_2^2/(n - 1) = S^4$, where S^2 is the unbiased sample variance. It follows that, to order $O(n^{-2})$, var(m_2) may be estimated by $(m_4 - m_2^2)/n$. Therefore instead of estimating var(S_j^2) by $2S_j^4/(n_j + 1)$, we propose using $(m_{j4} - S_j^4)/n_j$. Thus, in the formula for T_{OC} given in (3.1), replace $\lambda_j = (n_j + 1)/(2S_j^4)$ by $\lambda_j = n_j/(m_{j4} - S_j^4)$. This test statistic will be denoted by T_{AR}.

A preliminary study found that agreement of the nominal and actual test size for T_{OC} was not good enough for practical use (see Figure 3). The Bartlett test statistic, however, has been adjusted to improve this agreement, and so a comparison between B, T_{OC} and T_{AR} does not compare like with like. There are various options for such an improvement, and here we have opted to adjust the 5% critical points. Under the assumption of normality and under the null hypothesis of equality of variances, we estimated the 5% critical points for N between 10 and 50, inclusive, using 100,000 simulations of each sample size for $m = 4$ and $m = 8$ populations. Using standard curve fitting techniques, we found that these were well approximated over that range by

$$c(4, N, 0.05) = \chi^2_{3;0.05}\,(-0.714 - 28.082N^{-0.5} - 141.396N^{-1} + 280.374N^{-1.5}) \text{ for } m = 4 \text{ and}$$
$$c(8, N, 0.05) = \chi^2_{7;0.05}\,(-1.634 - 42.679N^{-0.5} - 213.485N^{-1} + 415.522N^{-1.5}) \text{ for } m = 8.$$

Using these critical values for $m = 4$ and $m = 8$ and for $10 \le N \le 50$ the actual test sizes vary between 0.046 and 0.054.

Use of the asymptotic 5% values $\chi^2_{m-1;0.05}$ required very large sample sizes, depending on m and the nominal level α. Their use cannot be recommended.

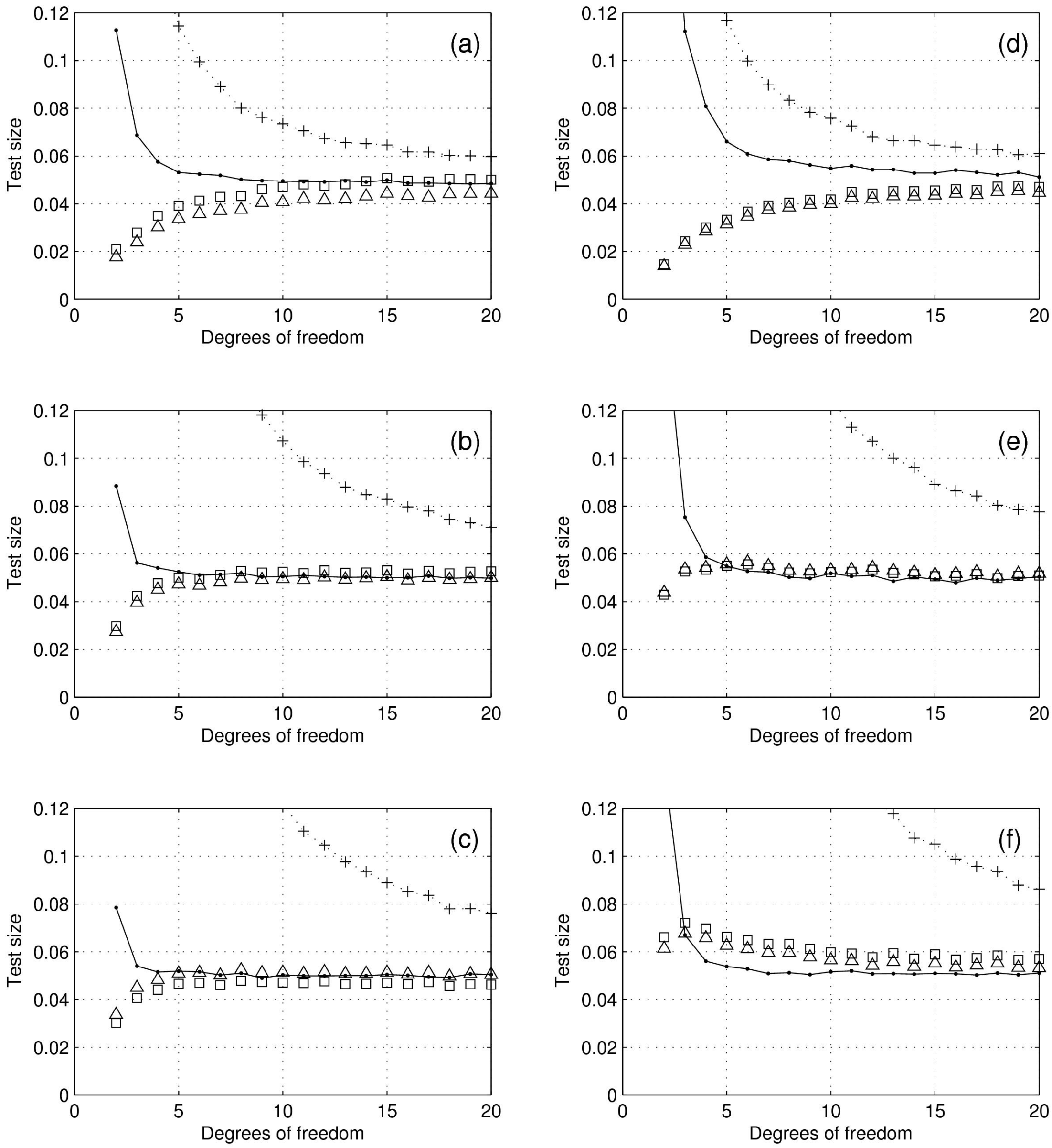

Figure 3. Test sizes for the tests based on L (solid line and dots), T_{OC} (+ and dots), T_{AR} using critical values estimated under normality (triangles) and T_{AR} using four term Bartlett correction (squares). For a significance level of 5% we compare four populations (panels (a), (b) and (c)) and eight populations (panels (d), (e) and (f)), and sample sizes of $N = 10$ (panels (a) and (d)), 30 (panels (b) and (e)), and 50 (panels (c) and (f)).

For $N > 50$, or for levels other than 5%, or for unequal sample sizes, critical values can be estimated using the MATLAB or R code available at the URL given in the conclusion.

To assess the effect of non-normality, instead of sampling from the standard normal distribution, we sampled from t_ν distributions with varying degrees of freedom, ν. If ν is large, the distribution sampled will be close enough to normal that we could expect the proportion of rejections to be close to the nominal. The critical values used in this assessment are both $c(m, N, 0.05)$ and the estimated 'exact' critical points.

In Figure 3 we plot the proportion of rejections - actual test size - for the tests based on L, T_{OC} and, with both critical values, T_{AR}. Sampling is from t_ν distributions for $\nu = 1, ..., 30$. We show curves for each test with common sample sizes $N = 10$, 30 and 50.

It is apparent that the test based on T_{OC} performs increasingly poorly as the degrees of freedom reduce and the tails of the distribution become fatter. Although not shown in the figure, this parallels the results for the Bartlett test.

The Levene test generally has exact level closer to the nominal level than the tests based on T_{AR} except for small degrees of freedom. However the level of the test based on T_{AR} is almost always reasonable, and while for very small ν the level is not as close to the exact level as perhaps we may prefer, the same is the case for the Levene test.

In general, a Wald test of H: $\theta = 0$ against K: $\theta \neq 0$ is based on a quadratic form $\hat{\theta}^T\hat{\Sigma}^{-1}\hat{\theta}$, in which $\hat{\theta}$ is asymptotically equivalent to the maximum likelihood estimator of θ and $\hat{\Sigma}$ is the asymptotic covariance matrix of $\hat{\theta}$. Thus, the tests based on T_{OC} and T_{AR} are both Wald tests and will have the weak optimality these tests enjoy. In particular they are asymptotically equivalent to the likelihood ratio test, and thus will have power approaching the power of the Bartlett test in large samples. As the tests based on T_{AR}, using either $c(m, N, 0.05)$ or estimated 'exact' critical values, have actual test sizes close to the nominal 5% test size, are asymptotically optimal and are robust (at least to fat-tailed distributions), they can be recommended.

6 Example: *National Institute of Standards and Technology Data*

We analysed data from the National Institute of Standards and Technology, involving ten groups of ten observations (NIST/SEMATECH, 2006). The collection of all unadjusted observations is not consistent with normality (Shapiro-Wilk p-value less than 1%) but if the group means are subtracted from each observation the collection of all centred observations is consistent with normality (Shapiro-Wilk p-value 0.35). We conclude the data are consistent with the assumption that they are $N(\mu_j, \sigma_j^2)$ distributed.

Without loss of generality all observations are multiplied by 1000, giving the following standard deviations for the ten groups:

4.3461, 5.2164, 3.9777, 3.8528, 7.5785, 9.8860, 7.8775, 3.6271, 4.1379 and 5.3292.

We find that B and T_{AR} take the values of 20.786 and 26.535, respectively, with corresponding Monte Carlo p-values, based on 100,000 simulations, of 0.013 and 0.445. The χ_9^2 p-value for B is 0.014, consistent with the Monte Carlo p-value. The sample sizes N are too small to use the asymptotic distribution to obtain a p-value for T_{AR}.

From the Bartlett test it appears that there is evidence, at the 5% level, that the variances are not consistent. The Bartlett test is much more critical of the null hypothesis than the test based on T_{AR}. This is consistent with the powers shown in Figure 2, where for small sample sizes the Bartlett test appears to be more powerful.

We doubled and trebled N by duplicating the data within each population and found that the T_{AR} test was then equally critical of the data.

If the data are not consistent with normality then the Bartlett test cannot be used. However, the test based on T_{AR} is valid and can be implemented by finding a Monte Carlo p-value. This merely involves generating sets of standard normal values for samples of size $n_1, ..., n_m$ and calculating T_{AR} for each data set. The proportion of exceedances of the observed T_{AR} value is the Monte Carlo p-value. MATLAB code to implement this procedure is available at the web site given in the conclusion.

7 Conclusion

We have given a compact derivation for the test based on T_{MR}, introduced in Mather and Rayner (2002). A new test, based on T_{OC}, is also derived. In fact T_{MR} and T_{OC} are identical. A test due to Rippon and Rayner (2010), based on T_{MP}, is T_{OC} minus a 'correction factor'. The test based on T_{MP} approaches its asymptotic distribution quite slowly. Thus unless resampling methods are used to calculate p-values, the T_{OC} test is to be preferred to T_{MP}.

Although the test new test based on T_{OC} cannot be claimed to be superior to that based on B in terms of size and power, it is competitive. The Bartlett test adheres to its nominal significance level more closely, no doubt in part because the test statistic has been adjusted; were the new test to be so adjusted it is very probable that both tests will perform similarly. For larger sample sizes the powers are similar.

A robustness study shows that when sampling is from fat-tailed t distributions instead of the normal, the tests based on B and T_{OC} do not have actual test size close to nominal. However tests based on L and T_{AR} are acceptable in this regard. As the test based on T_{AR} is a Wald test it can be expected to have good power in large samples.

Even when not sampling from normal distributions, test statistics of the same form as T_{AR}, quadratic forms with vector asymptotically equivalent to the maximum likelihood estimator of the parameter of interest and matrix the asymptotic covariance matrix of $\hat{\theta}$, are Wald tests and can be expected to have excellent properties in large samples.

To apply the test based on T_{AR} three options are available. For $m = 4$ and $m = 8$, $\alpha = 5\%$ and equal sample sizes N between 10 and 50 from each population, the Bartlett approximate critical values $c(m, N, 0.05)$ given in Section 5, may be used. For other configurations of $(n_1, ..., n_m)$ and other α, critical values may be estimated using the MATLAB or R code available at http://hdl.handle.net/1959.13/922372. Code to calculate the Monte Carlo p-value for a data set is available at the same URL.

References

Allingham, David and Rayner, J. C. W. (2011). A nonparametric two-sample Wald test of equality of variances. Submitted.

Mather, Katrina J. and Rayner, J.C.W. (2002). Testing equality of corresponding variances from multiple covariance matrices. In *Advances in Statistics, Combinatorics and Related Areas*. Chandra Gulati, Yan-Xia Lin, Satya Mishra and John Rayner (eds). Singapore: World Scientific Press, 157-166.

NIST/SEMATECH (2006). Data used for chi-square test for the standard deviation, http://www.itl.nist.gov/div898/handbook/eda/section3/eda3581.htm. Accessed November 2011.

Rayner, J.C.W. (1997). The asymptotically optimal tests. *J.R.S.S., Series D (The Statistician)*, 46(3), 337-346.

Rippon, Paul and Rayner, J.C.W. (2010). Generalised score and Wald tests. *Advances in Decision Sciences*, Article ID 292013, 8 pages. doi:10.1155/2010/292013.

Stuart, A. and Ord, J.K. (1994). *Kendall's Advanced Theory Of Statistics, Vol.1: Distribution theory* (6th ed.), London: Hodder Arnold.

Appendix: Derivation of T_{MR}

First, C is augmented by the row vector (0, … , 0, 1) to form the $m \times m$ matrix

$$C^* = \begin{pmatrix} & C & \\ 0 & \ldots & 0 \quad 1 \end{pmatrix}.$$

The inverse C^{*-1} is readily shown to satisfy $(C^{*-1})_{ij} = 1$ for $j \geq i$ and 0 otherwise. To see this define $\theta^* = (\theta^T, \theta_m^*)^T = C^*\phi$ so that $\theta_1 = \phi_1 - \phi_2, \ldots, \theta_{m-1} = \phi_{m-1} - \phi_m$, $\theta_m^* = \phi_m$, whence $\phi_m = \theta_m^*$, $\phi_{m-1} = \theta_{m-1} + \phi_m = \theta_{m-1} + \theta_m^*$, $\phi_{m-2} = \theta_{m-2} + \theta_{m-1} + \theta_m^*, \ldots, \phi_1 = \theta_1 + \theta_2 + \ldots + \theta_m^*$. Thus $\phi = C^{*-1}\theta^*$ with C^{*-1} as stated.

To calculate $T_{\mathrm{MR}} = \hat{\theta}^T \hat{\mathrm{cov}}^{-1}(\hat{\theta})\hat{\theta}$ $\hat{\mathrm{cov}}(\hat{\theta})$ is first required. Now as above $\hat{\mathrm{cov}}(\hat{\phi}) = D$ so that $\hat{\mathrm{cov}}(\hat{\theta}) = CDC^T$. Similarly $\hat{\mathrm{cov}}(\hat{\theta}^*) = C^*DC^{*T}$, so that $\hat{\mathrm{cov}}^{-1}(\hat{\theta}^*) = C^{*T-1}D^{-1}C^{*-1}$ is known. We require $\hat{\mathrm{cov}}^{-1}(\hat{\theta})$, the leading $(m-1) \times (m-1)$ block of $\hat{\mathrm{cov}}^{-1}(\hat{\theta}^*)$. This may be found using Lemma 3.1 with $A = \hat{\mathrm{cov}}(\hat{\theta})$ and, recalling that $\lambda_j = 1/d_j$ for all j, (clearly all $d_j > 0$), we have

$$\begin{pmatrix} E & F \\ G & H \end{pmatrix} = C^{*T-1}D^{-1}C^{*-1} = \begin{pmatrix} \lambda_1 & \lambda_1 & \cdots & \lambda_1 \\ \lambda_1 & \lambda_1 + \lambda_2 & \cdots & \lambda_1 + \lambda_2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1 & \lambda_1 + \lambda_2 & \cdots & \lambda_1 + \ldots + \lambda_m \end{pmatrix}.$$

Here E is $(m-1) \times (m-1)$, $G = (\lambda_1, \lambda_1 + \lambda_2, \ldots, \lambda_1 + \lambda_2 + \ldots + \lambda_{m-1})$, $F = G^T$ and H is the scalar $\lambda_1 + \lambda_2 + \ldots + \lambda_m$. The Wald test statistic is $T_{\mathrm{MR}} = \hat{\theta}^T \hat{\mathrm{cov}}^{-1}(\hat{\theta})\hat{\theta} = \hat{\theta}^T A^{-1}\hat{\theta} = W_1 - W_2$ in which $W_1 = \hat{\theta}^T E\hat{\theta}$ and $W_2 = \hat{\theta}^T FH^{-1}G\hat{\theta}$. Now

$$W_1 = \lambda_1\hat{\theta}_1^2 + (\lambda_1 + \lambda_2)\hat{\theta}_2^2 + \ldots + (\lambda_1 + \ldots + \lambda_{m-1})\hat{\theta}_{m-1}^2 +$$

$$2\{\lambda_1\hat{\theta}_1\hat{\theta}_2 + \ldots + \lambda_1\hat{\theta}_1\hat{\theta}_{m-1} + (\lambda_1 + \lambda_2)\hat{\theta}_2\hat{\theta}_3 + \ldots + (\lambda_1 + \lambda_2)\hat{\theta}_2\hat{\theta}_{m-1} + \ldots + (\lambda_1 + \ldots + \lambda_{m-2})\hat{\theta}_{m-2}\hat{\theta}_{m-1}\}$$

$$= \lambda_1\{\hat{\theta}_1^2 + \ldots + \hat{\theta}_{m-1}^2 + 2\hat{\theta}_1\hat{\theta}_2 + \ldots + 2\hat{\theta}_{m-2}\hat{\theta}_{m-1}\}^2$$

$$+ \lambda_2\{\hat{\theta}_2^2 + \ldots + \hat{\theta}_{m-1}^2 + 2\hat{\theta}_2\hat{\theta}_3 + \ldots + 2\hat{\theta}_{m-2}\hat{\theta}_{m-1}\}^2 + \ldots + \lambda_{m-1}\hat{\theta}_{m-1}^2$$

$$= \lambda_1(\hat{\theta}_1 + \hat{\theta}_2 + \ldots + \hat{\theta}_{m-1})^2 + \lambda_2(\hat{\theta}_2 + \hat{\theta}_3 + \ldots + \hat{\theta}_{m-1})^2 + \ldots + \lambda_{m-1}\hat{\theta}_{m-1}^2$$

and

$$G\hat{\theta} = \lambda_1\hat{\theta}_1 + (\lambda_1+\lambda_2)\hat{\theta}_2 + \ldots + (\lambda_1+\ldots+\lambda_{m-1})\hat{\theta}_{m-1} =$$
$$\lambda_1\left(\hat{\theta}_1+\hat{\theta}_2+\ldots+\hat{\theta}_{m-1}\right) + \lambda_2\left(\hat{\theta}_2+\ldots+\hat{\theta}_{m-1}\right) + \ldots + \lambda_{m-1}\hat{\theta}_{m-1}.$$

Thus, writing $\lambda. = \lambda_1 + \lambda_2 + \ldots + \lambda_m$, we have

$$\lambda.\, T_{\mathrm{MR}} = \lambda.\, \{\lambda_1\left(\hat{\theta}_1+\hat{\theta}_2+\ldots+\hat{\theta}_{m-1}\right)^2 + \lambda_2\left(\hat{\theta}_2+\hat{\theta}_3+\ldots+\hat{\theta}_{m-1}\right)^2 + \ldots + \lambda_{m-1}\hat{\theta}^2_{m-1}\} -$$
$$\left\{\lambda_1\left(\hat{\theta}_1+\hat{\theta}_2+\ldots+\hat{\theta}_{m-1}\right) + \lambda_2\left(\hat{\theta}_2+\ldots+\hat{\theta}_{m-1}\right) + \ldots + \lambda_{m-1}\hat{\theta}_{m-1}\right\}^2 =$$
$$= \lambda_1(\lambda. - \lambda_1)\left(\hat{\theta}_1+\hat{\theta}_2+\ldots+\hat{\theta}_{m-1}\right)^2 + \lambda_2(\lambda. - \lambda_2)\left(\hat{\theta}_2+\ldots+\hat{\theta}_{m-1}\right)^2 + \ldots +$$
$$\lambda_{m-1}(\lambda. - \lambda_{m-1})\hat{\theta}^2_{m-1} - 2\lambda_1\lambda_1\left(\hat{\theta}_1+\hat{\theta}_2+\ldots+\hat{\theta}_{m-1}\right)\left(\hat{\theta}_2+\hat{\theta}_3+\ldots+\hat{\theta}_{m-1}\right) - \ldots -$$
$$2\lambda_{m-2}\lambda_{m-1}\,\hat{\theta}_{m-2}\hat{\theta}_{m-1}$$
$$= \lambda_1(\lambda_2+\ldots+\lambda_m)\left(\hat{\theta}_1+\hat{\theta}_2+\ldots+\hat{\theta}_{m-1}\right)^2 + \lambda_2(\lambda_1+\lambda_3+\ldots+\lambda_m)\left(\hat{\theta}_2+\ldots+\hat{\theta}_{m-1}\right)^2 + \ldots +$$
$$\lambda_{m-1}(\lambda_1+\ldots+\lambda_{m-2}+\lambda_m)\hat{\theta}^2_{m-1} - 2\lambda_1\lambda_2\left(\hat{\theta}_1+\hat{\theta}_2+\ldots+\hat{\theta}_{m-1}\right)\left(\hat{\theta}_2+\hat{\theta}_3+\ldots+\hat{\theta}_{m-1}\right) - \ldots$$
$$- 2\lambda_{m-1}\lambda_{m-2}\,\hat{\theta}_{m-2}\hat{\theta}_{m-1} =$$
$$= \lambda_1\lambda_2\left\{\left(\hat{\theta}_1+\hat{\theta}_2+\ldots+\hat{\theta}_{m-1}\right) - \left(\hat{\theta}_2+\hat{\theta}_3+\ldots+\hat{\theta}_{m-1}\right)\right\}^2 + \ldots + \lambda_{m-2}\lambda_{m-1}\left\{\left(\hat{\theta}_{m-2}+\hat{\theta}_{m-1}\right) - \hat{\theta}_{m-1}\right\}^2$$
$$+ \lambda_1\lambda_3\left\{\left(\hat{\theta}_1+\hat{\theta}_2+\ldots+\hat{\theta}_{m-1}\right) - \left(\hat{\theta}_3+\hat{\theta}_4+\ldots+\hat{\theta}_{m-1}\right)\right\}^2 + \ldots +$$
$$\lambda_{m-3}\lambda_{m-1}\left\{\left(\hat{\theta}_{m-3}+\hat{\theta}_{m-2}+\hat{\theta}_{m-1}\right) - \hat{\theta}_{m-1}\right\}^2 + \ldots + \lambda_1\lambda_m\left(\hat{\theta}_1+\hat{\theta}_2+\ldots+\theta_{m-1}\right)^2 + \ldots +$$
$$\lambda_2\lambda_m\left(\hat{\theta}_2+\hat{\theta}_3+\ldots+\hat{\theta}_{m-1}\right)^2 + \ldots + \lambda_{m-1}\lambda_m\hat{\theta}^2_{m-1} =$$
$$= \sum_{\substack{i<j \\ i,j=1}}^{m} \lambda_i\lambda_j\left(\hat{\phi}_i - \hat{\phi}_j\right)^2.$$

In particular for $m = 1$ $T_{\mathrm{MR}} = \lambda_1\lambda_2\left(\hat{\phi}_1 - \hat{\phi}_2\right)^2 / (\lambda_1+\lambda_2)$, which agrees with (1.1) after substituting for the λ_j. Similarly for $m = 2$

$$T_{\mathrm{MR}} = \left\{\lambda_1\lambda_2\left(\hat{\phi}_1-\hat{\phi}_2\right)^2 + \lambda_2\lambda_3\left(\hat{\phi}_2-\hat{\phi}_3\right)^2 + \lambda_3\lambda_1\left(\hat{\phi}_3-\hat{\phi}_1\right)^2\right\} / (\lambda_1+\lambda_2+\lambda_3).$$

Routinely substituting for $\lambda_1, \ldots, \lambda_m$ and $\lambda.$ gives (2.1).