A Nonparametric Two-Sample Wald Test of Equality of Variances

David Allingham*[1] and J. C. W. Rayner[2]

Abstract

We develop a test for equality of variances given two independent random samples of observations. The test can be expected to perform well when both sample sizes are at least moderate and the sample variances are asymptotically equivalent to the maximum likelihood estimators of the population variances.

The test is motivated by and is here assessed for the case when both populations sampled are assumed to be normal. Popular choices of test would be the two sample F test if normality can be assumed and Levene’s test if this assumption is dubious. Another competitor is the Wald test for the difference in the population variances. We give a nonparametric analogue of this test and call it the R test.

In an indicative empirical study when both populations are normal, we find that for moderate sample sizes the R test is nearly as robust as Levene’s test and nearly as powerful as the F test.

Key Words: Bartlett’s test; Levene’s test; Wald tests.

1. Introduction: Testing Equality of Variances for Two Independent Samples

In the two-sample problem we are given two independent random samples X_{11}, ..., X_{1n_1} and X_{21}, ..., X_{2n_2}. The location problem attracts most attention. Assuming that the samples are from normal populations, the pooled t-test is used to test equality of means assuming equal variances and Welch’s test can be used when equality of variances is suspect but normality is not. When normality is in doubt the Wilcoxon test is often used.

The corresponding dispersion problem is of interest to confirm the validity of, for example, the pooled t-test, and for its own sake. For example, testing for reduced variability is of interest in confirming natural selection (see, for example, [1, Section 5.5]) and if some processes are in control. In exploratory data analysis it is sensible to assess if one population is more variable than another. If it is, the cause may be that one population is bi-modal and the other is not; the consequences of this in both the scenario and the model can then be explored in depth.

The study here introduces a new test of equality of variances. The asymptotic null distribution of the test statistic is χ_1^2, but, depending on the populations sampled from, this may or may not be satisfactory for small to moderate sample sizes.

To assess this, and other aspects of the proposed test, an indicative empirical study is undertaken. We give comparisons when both populations sampled are normal, when both samples have the same sample size, and for 5% level tests. First

*Author to whom correspondence should be addressed.

[1] Centre for Computer-Assisted Research Computation and its Applications, School of Mathematical and Physical Sciences, University of Newcastle, NSW 2308, Australia
e-mail: David.Allingham@newcastle.edu.au

[2] 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
e-mail: John.Rayner@newcastle.edu.au

we derive a finite sample correction to the critical values based on the asymptotic null distribution when sampling from normal populations. Different corrections would be needed when sampling from different distributions. Next we show that in moderate samples the new test is nearly as powerful as the F test when normality may be assumed, and finally, that it is nearly as robust as Levene's test when normality is in doubt. Moore and McCabe [2, p.519] claim that the "F test and other procedures for inference about variances are so lacking in robustness as to be of little use in practice." The new test gives a counterexample to that proposition.

We acknowledge that the new test is most effective for at least moderate sample sizes. In the normal case each random sample should have at least 25 observations, which is what we would expect of a serious study aiming at reasonable power that cannot be hoped for with samples of size 10 or so. See Section 4.

We are aware of more expansive comparative studies such as [3 and 4]. Our goal here is not to emulate these studies but to show that the new test is competitive in terms of test size, robustness and power.

In Section 2 the new test is introduced. Sections 3, 4 and 5 give the results of an empirical investigation when the populations sampled are assumed to be normal. In Section 3 we investigate test size, showing that the asymptotic χ^2 critical values should only be used for moderate to large sample sizes. For smaller sample sizes a finite sample correction to the asymptotic 5% χ^2 critical value is given. This results actual test sizes between 4.6% and 5.3%.

In Section 4 we show that when normality holds the new test is not as powerful as the Levene test for smaller sample sizes, but overtakes it for moderate samples of about 25. The new test is always inferior to the optimal F test. However the R test has power that approaches that of the F test, being at least 95% that of the F test throughout most of the parameter space in samples of at least 80.

In Section 5 we show that if we sample from t distributions with varying degrees of freedom, the F test is highly non-robust for small degrees of freedom, as is well-known for fat-tailed distributions. The new test performs far better than the F test, and although it is challenged somewhat for small degrees of freedom, its performance is only slightly inferior to the Levene test.

That the new test gives a good compromise between power and robustness, and is valid when normality doesn't hold, are strong reasons for preferring the new test for sample sizes that are at least moderate and normality is dubious.

2. Competitor Tests for Equality of Variance

Initially we assume that we have two independent random samples X_{i1}, ..., X_{in_i} from normal populations, $N(\mu_i,\ \sigma_i^2)$ for i = 1 and 2. We wish to test H: $\sigma_1^2 = \sigma_2^2$ against the alternative K: $\sigma_1^2 \neq \sigma_2^2$, with the population means being unknown nuisance parameters. If $S_i^2 = \sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\bullet})^2/(n_i - 1)$, in which $\overline{X}_{i\bullet} = \sum_{j=1}^{n_i} X_{ij}/n_i$, i = 1, 2, then the S_i^2 are the unbiased sample variances, and the so-called F test is based on their quotient, $S_2^2/S_1^2 = F$, say. It is well-known, and will be confirmed yet again in Section 5, that the null distribution of F, F_{n_1-1,n_2-1}, is sensitive to departures from normality. If $F_{n_1-1,n_2-1}(x)$ is the cumulative distribution function of the F_{n_1-1,n_2-1}

distribution, and if c_p is such that $F_{n_1-1,n_2-1}(c_p) = p$, then the F test rejects H at the $100\alpha\%$ level when $F \le c_{\alpha/2}$ and when $F \ge c_{(1-\alpha/2)}$.

Common practice when normality is in doubt is to use Levene's test or a nonparametric test such as Mood's test. Levene's test is based on the ANOVA F test applied to the residuals. There are different versions of Levene's test using different definitions of residual. The two most common versions use residuals based on the group means, $\left|X_{ij} - \bar{X}_{i\bullet}\right|$, and the group medians, $\left|X_{ij} - \tilde{X}_{i\bullet}\right|$, in which $\tilde{X}_{i\bullet}$ is the median of the ith sample. The latter is called the Brown-Forsythe test. Again it is well-known that these tests are robust in that when the population variances are equal but the populations themselves are not normal, they achieve levels close to nominal. However this happens at the expense of some power. As the empirical study in this paper is intended to be indicative rather than exhaustive, we will henceforth make comparisons only with the Levene test, based on a test statistic we denote by L.

We now construct a new test that we call the R test. For univariate parameters θ, a Wald test statistic of H: $\theta = \theta_0$ against the alternative K: $\theta \ne \theta_0$ is based on $\hat{\theta}$, the maximum likelihood estimator of θ, usually via the test statistic $(\hat{\theta} - \theta_0)^2 / \mathrm{est\,var}(\hat{\theta})$, where $\mathrm{est\,var}(\hat{\theta})$ is a consistent estimate of $\mathrm{var}(\hat{\theta})$. Under the null hypothesis this test statistic has an asymptotic χ_1^2 distribution. As well as being equivalent to the likelihood ratio test, the F test is also a Wald test for testing H: $\theta = \sigma_2^2 / \sigma_1^2 = 1$ against K: $\theta \ne 1$.

Rayner [5] derived the Wald test for testing H: $\theta = \sigma_2^2 - \sigma_1^2 = 0$ against K: $\theta \ne 0$. The test statistic is

$$\frac{(S_1^2 - S_2^2)^2}{\dfrac{2S_1^4}{n_1+1} + \dfrac{2S_2^4}{n_2+1}} = W, \text{ say.}$$

Being a Wald test, the asymptotic distribution of W is χ_1^2, while its exact distribution is not immediately obvious. However W is a 1-1 function of F, so the two tests are equivalent. Since the exact distribution of F is available, the F test is the obvious test to use.

In W, the variances var(S_j^2) are estimated optimally using the Rao-Blackwell theorem. This depends very strongly on the assumption of normality. If normality is in doubt then we can estimate var($S_1^2 - S_2^2$) using results given, for example, in [6]. For a random sample $Y_1, \ldots, Y_n$ with population and sample central moments μ_r and $m_r = \sum_{j=1}^{n} (Y_j - \bar{Y})^r / n$, r = 2, 3, ... , [6] gives that

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

Applying [6, 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. We thus propose a robust alternative to W, given by

$$\frac{(S_1^2 - S_2^2)^2}{\frac{m_{14} - S_1^4}{n_1} + \frac{m_{24} - S_2^4}{n_2}} = R \text{ say,}$$

in which m_{i4}, are the fourth central sample moments for the ith sample, i = 1, 2. We call the test based on R the R test. As the sample sizes increase, the distributions of the sample variances approach normality, the denominator in R will approximate var($S_1^2 - S_2^2$), and R will have asymptotic distribution χ_1^2. Thus, if c_α is the point for which the χ_1^2 distribution has weight α in the right hand tail, then the R test rejects H at approximately the 100α% level when $R \geq c_\alpha$.

We emphasise that although the motivation for the derivation of R is under the assumption of sampling from normal populations, it is a valid test statistic for testing equality of variances no matter what the populations sampled.

If the sample variances are equal to or asymptotically equivalent to the maximum likelihood estimators of the population variances, as is the case when sampling from normal populations, then the R test is a Wald test for equality of variances in the sense described above. Since it doesn't depend on any distributional assumptions about the data, it can be thought of as a nonparametric Wald test. As such it can be expected to have good properties in large samples.

We note that all the above test statistics are invariant under transformations $a(X_{ij} - b_i)$, for constants a, b_1 and b_2 and for j = 1, ..., n_i and i = 1, 2.

The next three sections report an empirical study when the distributions sampled are assumed to be normal. As this is an indicative study, we fix the samples sizes to be equal, $n_1 = n_2 = n$, say, and the significance level to be 5% throughout.

3. Test Size Under Normality

Under the null hypothesis, the distribution of F is known exactly, the distribution of L is known approximately, and, as mentioned above, the distribution of R is known asymptotically. In analysing data these distributions are used to determine p-values and critical values. We now investigate their use in determining test size, the probability of rejecting the null hypothesis when it is true.

Two empirical assessments of test size will now be undertaken. Since the test statistics are scale invariant, it is sufficient under the null hypothesis to take both population variances to be one.

In the first assessment we assume normality. For various values of the common sample size n, we estimate the 5% critical points for each test by generating 100,000 pairs of random samples of size n, calculating the test statistics, ordering them and hence identifying the 0.95th percentile. The estimated critical points of R approach the χ_1^2 5% critical point 3.841. These estimated critical points will subsequently be used in the power study to give tests with test size of exactly 5%.

To see the extent of the error caused by using the asymptotic critical point 3.841, Figure 1 gives the proportion of rejections in 100,000 pairs of random samples for sample sizes up to 100. For n = 10 the proportion of rejections is nearly 20% and although for n = 40 this has dropped to nearly 7%, most users would hope for observed test sizes closer to 5%.

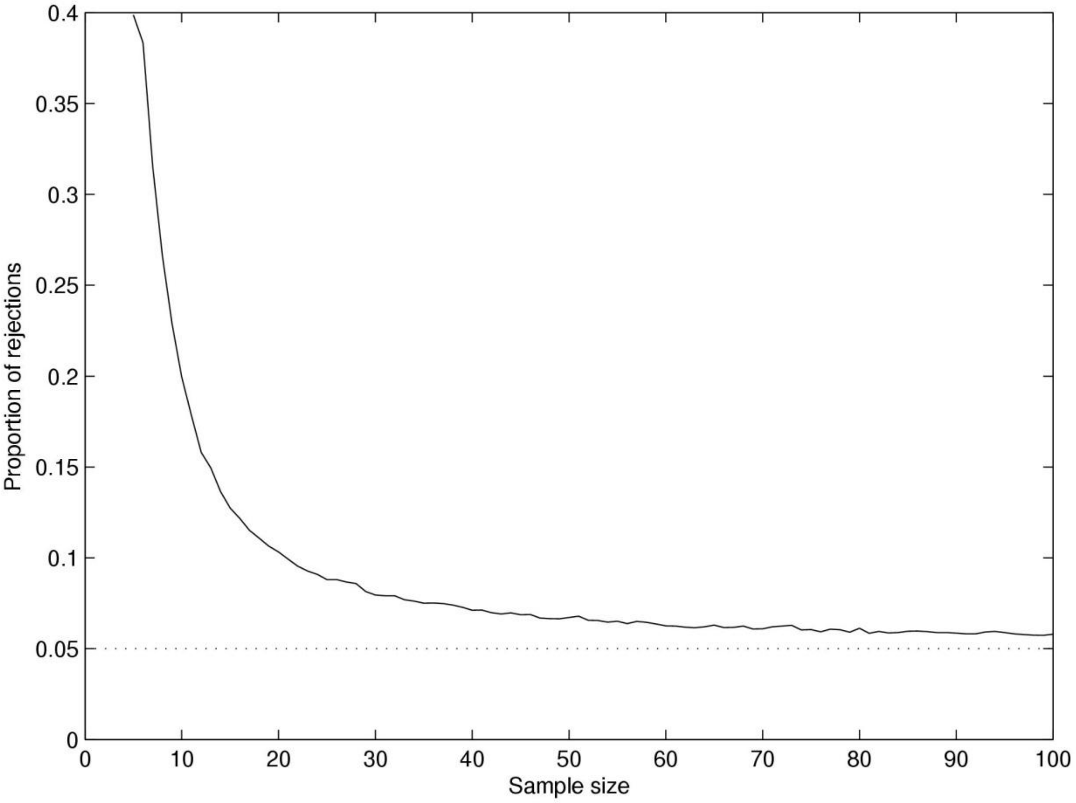

Figure 1. Proportion of rejections of the R test using the χ_1^2 5% critical point 3.841 for sample sizes up to 100

To improve the application of the test we plotted the estimated 5% critical points against n and used standard curve fitting techniques to find that the exact critical points were well approximated between n = 10 and 100 by $c(n, 0.05) = 3.84146(1.339 - 4.953/\sqrt{n} + 24.171/n)$. For larger n it is sufficient to use the asymptotic 5% value 3.84146, the error being at most 0.8%. We checked the exact probabilities of rejection under the null hypothesis when applying the test with critical value $c(n, 0.05)$, and all were between 4.6% and 5.3%.

For levels other than 5%, and when sample sizes are unequal, further empirical work needs to be done to find critical values. However, as this study was intended to be indicative, we leave extensive tabulation of critical values for another time.

4. Power Under Normality

For the F, Levene and R tests we estimate the power as the proportion of rejections from 100,000 pairs of random samples of size n when the first sample is from a N(0, 1) population and the second is from a N(0, σ^2) population with $\sigma^2 \geq 1$. To compare like with like, we use estimated critical values that give exact 5% level tests. It is apparent that for sample sizes less than about 20 the Levene test is more powerful than the R test and that between approximately 20 and 30 the R test takes over from the Levene test; thereafter the R test is always more powerful than the Levene test. This is shown in Figure 2.

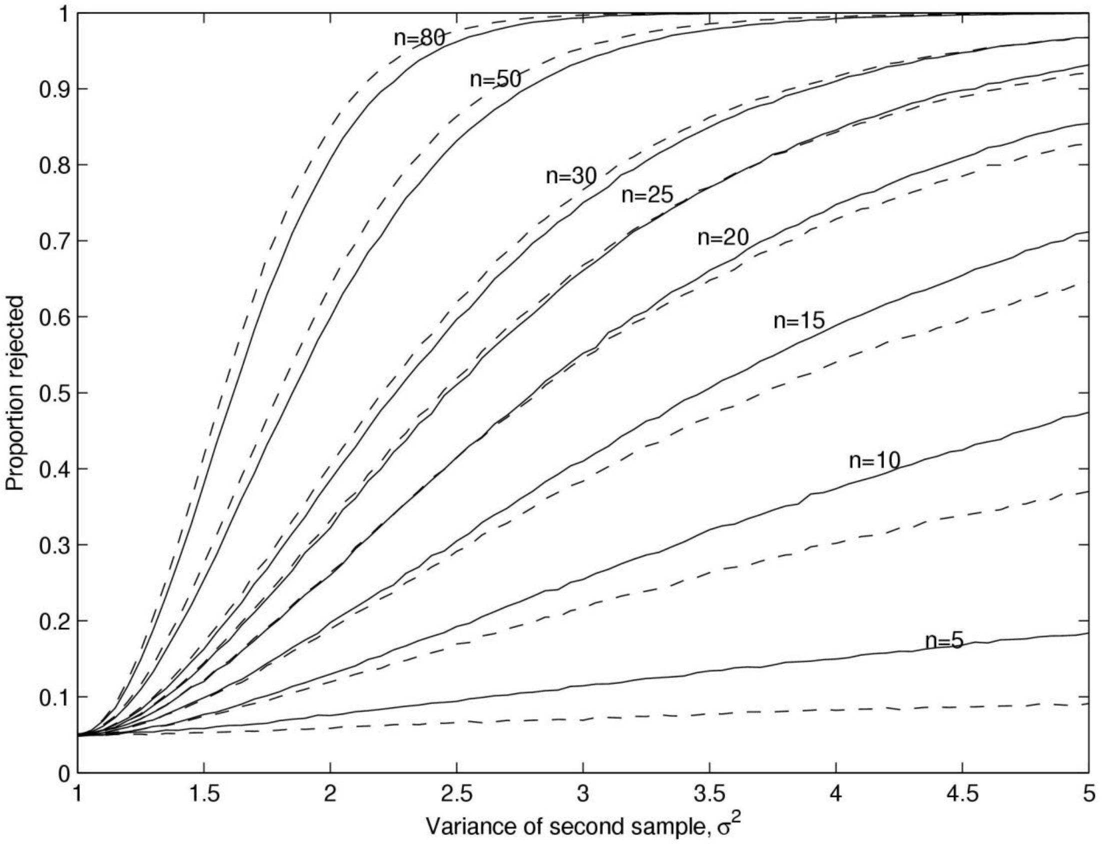

Figure 2. Power of the 5% level L test (solid line) and R test (dashed line) for various sample sizes

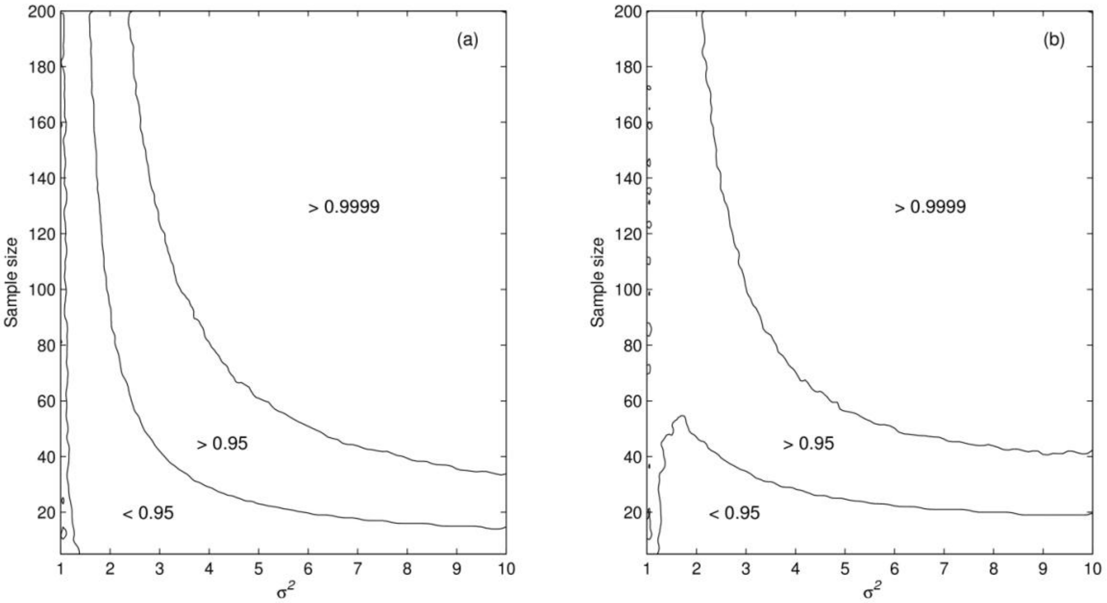

Figure 3. Contour plots of the L test (left panel) and R test (right panel) relative to the F test showing regions in which the power ratios are less than 95%, between 95% and 99.99%, and greater than 99.99%.

Both the Levene and R tests are always less powerful that the F test. This is explored in Figure 3 that compares the Levene test to the F test in the left hand panel and the R test to the F test in the right hand panel. What is given is a contour plot of the regions in which the ratios of the power of the stated test to the F test is less than 95%, between 95% and 99.99%, and greater than 99.99%. Generally, for any given *n* and σ^2, it is clear that the power of the Levene test is at most that of the R test. For example it appears that for $n_1 = n_2 = 80$ approximately the power of the R test is always at least 95% that of the F test, whereas there is a considerable region where the power of the Levene test is less than 95% that of the F test.

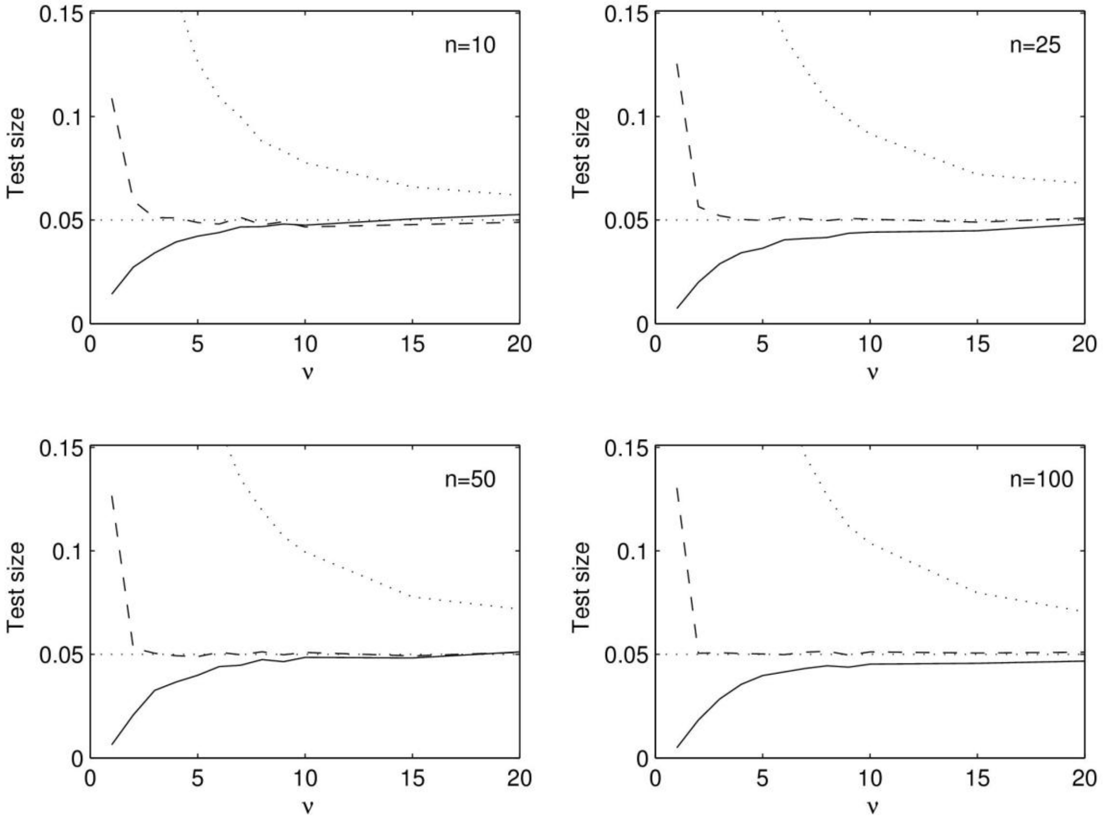

Figure 4. Test sizes for the F (dots), L (dashes) and R (solid line) tests for t distributions with varying degrees of freedom, ν, from 1 to 20

5. Robustness

Even if the R test has good power, the test is of little value unless it is robust in the sense that when the distributions from which we are sampling are not from the nominated population (here, the normal distribution) the p-values are reasonably accurate. It is thus of interest to estimate the proportion of rejections when the null hypothesis is true and both the populations from which we sample are not normal. We could have looked at variable skewness through gamma distributions with varying shape parameter, but instead we consider variable kurtosis via t distributions with varying degrees of freedom. If the degrees of freedom, ν say, are 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 here are $c(n, 0.05)$ given in Section 3.

In Figure 4 we plot the proportion of rejections for the Levene, F and R tests when sampling from t_ν distributions for $\nu = 1, ..., 20$. We show curves for each test with common sample sizes n = 10, 25, 50 and 100. The critical values used are the $c(n, 0.05)$ from Section 3.

It is apparent that the F test performs increasingly poorly as the degrees of freedom reduce and the tails of the distribution become fatter. Interestingly, in this scenario, the F test is always liberal (exact test size more than the significance level) while the R test is almost always conservative (exact test size less than the significance level). In general the latter is to be preferred.

The Levene test generally has exact level closer to the nominal level than the R test except for small degrees of freedom. However the level of the R test 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.

5. Conclusion

First, we reflect on testing for equality of variances when it is assumed that the populations sampled are normal. The F test is both the likelihood ratio test and a Wald test, and is the appropriate test to apply. When normality does not hold, the F test is no longer an asymptotically optimal test and its well-known non-robustness means that tests such as the Levene are more appropriate for small to moderate sample sizes. However for sample sizes of about 25 or more the R test is more powerful than the Levene and with the small sample corrected critical values it holds its nominal significance level well. For these sample sizes it can be preferred to the Levene test.

Second, consider testing for equality of variances when both samples are drawn from the same population. If that population is nominated then the R test may be applied after determining critical values or using p-values calculated by Monte Carlo methods. When the sample variances are asymptotically equivalent to the maximum likelihood estimators of the population variances (as, for example, is the case when sampling from normal populations but not Poisson populations), the R test is a nonparametric Wald test and hence will have good power in sufficiently large samples. If the population is not specified the R test can be confidently applied when the sample sizes are large, using the asymptotic χ_1^2 null distribution to calculate p-values or critical values.

References

[1] B.F.J. Manley, *The Statistics of Natural Selection*, Chapman and Hall, London, 1987.

[2] D.S. Moore and G.P. McCabe, *Introduction to the Practice of Statistics* (5th ed.), W.H. Freeman, New York, 2006.

[3] W.J. Conover, Mark E Johnson and Myrle M. Johnson, “A comparative study of tests of homogeneity of variances, with applications to the outer continental shelf bidding data,” *Technometrics*, vol. 23, no. 4, pp. 351- 361, 1981.

[4] Dennis D. Boos and Cavell Brownie, “Bootstrap methods for testing homogeneity of variances,” *Technometrics*, vol. 31, no. 1, pp. 69-82, 1989.

[5] J.C.W. Rayner, “The Asymptotically Optimal Tests*,” Journal of the Royal Statistical Society: Series D,* vol. 46, no. 3, 337-346, 1997.

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