Two-Sample Testing for Equality of Variances

David Allingham and J. C. W. Rayner

School of Mathematical and Physical Sciences,
The University of Newcastle, NSW 2308, Australia

David.Allingham@newcastle.edu.au and John.Rayner@newcastle.edu.au

Abstract

To test for equality of variances given two independent random samples from univariate normal populations, popular choices would be the two-sample F test and Levene's test. The latter is a nonparametric test while the former is parametric: it is the likelihood ratio test, and also a Wald test. Another Wald test of interest is based on the difference in the sample variances. We give a nonparametric analogue of this test and call it the R test. The R, F and Levene tests are compared in an indicative empirical study.

For moderate sample sizes when assuming normality the R test is nearly as powerful as the F test and nearly as robust as Levene's test. It is also an appropriate test for testing equality of variances without the assumption of normality, and so it can be strongly recommended.

Keywords: Bartlett's test; Levene's test; Wald tests.

1. Introduction

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

The corresponding dispersion problem is of interest to confirm the validity of, for example, the pooled t-test, and for its own sake. As an example, testing for reduced variability is of interest in confirming natural selection. In exploratory data analysis it is sensible to test whether one population is more variable than another. If it is, the cause may be that one population is bi-modal relative to the other; the consequences of this in both the scenario and the model can then be explored in depth.

Here, a new test for equality of variances based on what might be called a nonparametric version of a very natural Wald test is introduced. In an indicative empirical study we show that, in moderately-sized samples, the new test is nearly as powerful as the F test when normality may be assumed, and is nearly as robust as Levene's test when normality is in doubt. See [3, p.519], who say 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 usefulness of the new test is limited to moderate sample sizes of at least 25 each, a reasonable expectation in a serious study aiming at reasonable power which could not be hoped for with samples of size 10 or so.

We are aware of more expansive comparative studies such as those of [1] and [2]. Our goal here is not to emulate these studies but to merely show that the new test is competitive and interesting. Reflecting our limited study, we restrict attention to samples of equal size from both populations and a 5% level of significance.

In Section 2 the new test is introduced. In Section 3 we investigate test size. It is shown that when normality may be assumed the asymptotic χ^2 critical values may be used for moderate sample sizes, achieving test sizes 'close' to nominal. We then show that when sampling from t distributions with various 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 is challenged somewhat for small degrees of freedom, but its performance is only slightly inferior to the Levene test.

In Section 4 it is shown that when normality holds the new test is not as powerful as the Levene test for small sample sizes, but overtakes it for moderate sample sizes of about 25. The new test is always inferior to the optimal F test, but has power that approaches that of the F test, its power being at least 95% of that of the F test throughout most of the parameter space for sample sizes of at least 80. This, in conjunction with the fact that the new test is valid when normality doesn't hold, is a strong reason for preferring the new test for moderate sample sizes.

2. Competitor Tests for the Two-Sample Dispersion Problem

We assume independent random samples of sizes m and n from normal populations, $N(\mu_i, \sigma_i^2)$ for $i = 1, 2$. We wish to test H: $\sigma_1^2 = \sigma_2^2$ against the alternative K: $\sigma_1^2 \neq \sigma_2^2$. If S_i^2, $i = 1, 2$ are the unbiased sample variances, then the so-called F test is equivalent to the likelihood ratio test and is based on the quotient of the sample variances, $S_2^2 / S_1^2 = F$, say. It is well-known, and will be confirmed yet again in Section 3, that the null distribution of F, namely $F_{m-1,\, n-1}$, is sensitive to departures from normality. If the cumulative distribution function of this distribution is $F_{m-1,\, n-1}(x)$, and if c_p is such that $F_{m-1,\, n-1}(c_p) = p$, then the F test rejects H at the 100α% level when $F \leq c_{\alpha/2}$ and when $F \geq c_{1-\alpha/2}$.

Common practice when normality is in doubt is to use a nonparametric test such as the Levene test or the Mood test. In the two-sample case, Levene's test is just the pooled t-test applied to the sample residuals. There are different versions of Levene's test using different definitions of residual. The two most common versions use the group means, $| X_{ij} - \overline{X}_{i.} |$, and the group medians, $| X_{ij} - \tilde{X}_{i.} |$, in obvious notation. The latter is called the Brown-Forsythe test. The distribution of the test statistics, say L and B, that are the squares of the pooled t-test statistics using mean- and median-based residuals, respectively, is approximately $F_{1,\, m+n-2}$. Again it is well-known that the tests based on L and B 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 this paper presents an indicative, rather than exhaustive, study, we will henceforth make comparisons only with the Levene test.

We now construct a new test that we will call the R test. For univariate parameters θ, a Wald test statistic for H: $\theta = \theta_0$ against the alternative K: $\theta \neq \theta_0$ 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 $\mathrm{var}(\hat{\theta})$. This test statistic has an asymptotic χ_1^2 distribution. As well as being 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: $\theta \neq 1$.

A Wald test for testing H: $\theta = \sigma_2^2 - \sigma_1^2 = 0$ against K: $\theta \neq 0$ is derived in [4]. The test statistic is

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

say. Being a Wald test, the asymptotic distribution of W is χ_1^2, while its exact distribution is not obvious. However, W is a one-to-one function of F, and so the two tests are equivalent. Since the exact distribution of F is known, the F test is the more convenient test.

The variances var(S_j^2) used in W 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 in [5]. For a random sample $X_1, \ldots, X_n$ and population and sample central moments μ_r and $m_r = \sum_{j=1}^{n} (X_j - \overline{X})^r / n$, $r = 2, 3, \ldots$, [5] gives

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

Applying [5, 10.5], μ_2^2 may be estimated to $\mathrm{O}(n^{-1})$ 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 var(m_2) may be estimated to order $\mathrm{O}(n^{-2})$ by $(m_4 - m_2^2)/n$. A robust alternative to W is thus

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

say, in which m_{i4}, i=1, 2, are the fourth central sample moments for the ith sample. We call the test based on R the R test. In large samples the denominator in R will approximate var($S_1^2 - S_2^2$) and R will have asymptotic distribution χ_1^2.

We emphasise that the R test is a Wald test 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. It can be expected to have good properties in large samples no matter what distribution is sampled.

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

3. Test Size Under Normality and Non-normality

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

Two empirical assessments of test size, defined as the probability of rejecting the null hypothesis when it is true, will now be undertaken. The test statistics are scale invariant, and so it is sufficient under the null hypothesis to take both population variances to be one. As this is an indicative study, we take $m = n$ and the significance level to be 5%.

In the first assessment we assume normality. In Figure 1, the extent of the error caused by using the asymptotic critical point 3.841... in the R test is shown, using the proportion of rejections in K = 100,000 random samples. For $m = n$ = 10 and 30 these proportions are approximately 20% and 8%. Most would hopefully agree that the former is not acceptably 'close' to 5%, whilst the latter is.

For various n, we estimated the 5% critical points for each test by generating K = 100,000 pairs of random samples of size n, calculating the test statistics, ordering them and identifying the 0.95Kth percentile. The estimated critical points of R approach the χ_1^2 5% critical point 3.841.... These estimated critical points will be used in the subsequent power study later to give tests with test size exactly 5%.

Even if the R test has good power, the test is of little value unless it is robust in the sense that, even when the sampled distributions are not normal, the p-values are reasonably accurate. Thus in the second assessment we estimate the proportion of rejections when the null hypothesis is true and both the populations sampled are non-normal. We consider different kurtoses via t distributions with various degrees of freedom. If the degrees of freedom are large, say 50 or more, the sampled distribution will be sufficiently normal that the proportion of rejections should be close to the nominal.

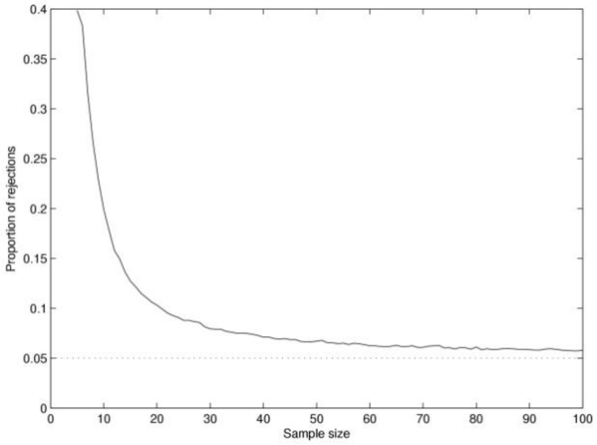

Figure 1: Proportion of rejections of the R test using the 5% critical point 3.841... for sample sizes up to 100.

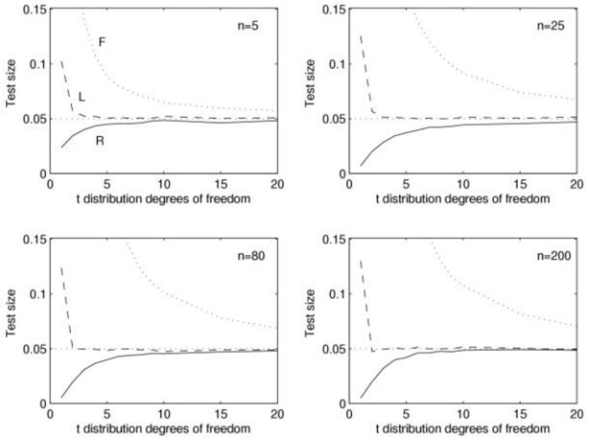

Figure 2: Test sizes for the F (dots), L (dashes) and R (solid line) tests for t distributions with varying degrees of freedom.

In Figure 2 we show the proportion of rejections for the Levene, F and R tests when sampling from t_ν distributions, for $\nu = 1, ..., 50$, with sample sizes of $m = n$ = 5, 25, 80 and 200. Interestingly, the achieved test size is closer to the nominal 5% value for smaller samples, in all cases.

It is apparent that the F test performs increasingly poorly as the degrees of freedom diminish. It is also interesting to note that in this scenario the F test is always liberal (exact size greater than 5%) while the R test is always conservative (exact size less than 5%). 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. Moreover, while the level of the R test is almost always reasonable, for very small ν the level is not as close to the exact level as perhaps would be preferable.

4. Power Under Normality

For the F, Levene and R tests we estimated the power as the proportion of rejections from K = 100,000 pairs of random samples of size n, where 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, estimated critical values that give virtually exact 5% level tests were used. It is apparent that for sample sizes of about 20 the Levene test is superior to the R test; that between approximately 20 and 30 the R test takes over from the Levene test; and that thereafter the R test is always more powerful than the L test. These results are shown in Figure 3.

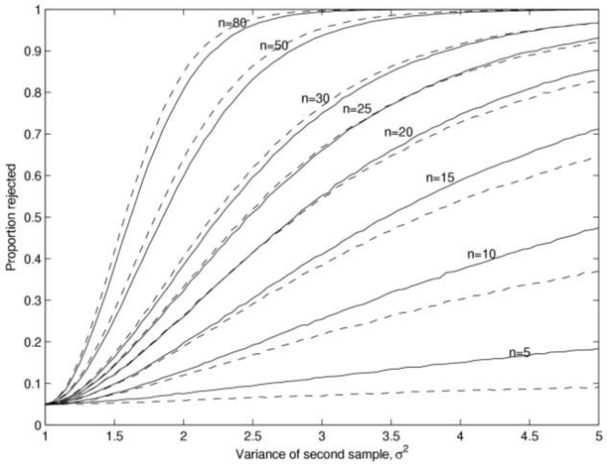

Figure 3: Power of the 5% level L test (solid line) and R test (dashed line) for various sample sizes.

When normality holds, both the Levene and R tests are always less powerful that the F test. This is explored in Figure 4, which 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. The figure shows a contour plot of the regions in which the ratio of the power of the stated test to the F test is either less than 95%, between 95% and 99.99%, or greater than 99.99%. The corresponding regions are far smaller for the Levene test than the R test. Moreover, it appears that for approximately $m = n > 80$, the power of the R test is always at least 95% of the power of the F test.

5. Recommendations

The R test is a nonparametric Wald test, so that when sampling from any non-normal distribution it can be expected to be at least as powerful as any competitor test in sufficiently large samples.

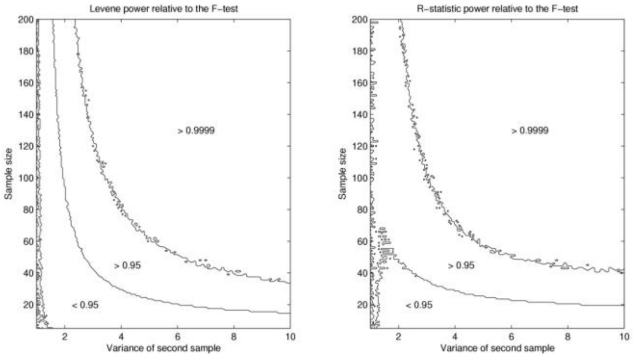

Figure 4: Contour plots of the power of the L test (left) and R test (right) relative to the F test power, showing regions in which the power ratios are less than 95%, between 95% and 99.99%, and greater than 99.99%.

If normality can be assumed then the F test is both the likelihood ratio test and a Wald test, and is the appropriate test to apply. However, if normality is doubtful then the well-known non-robustness of the F test means that tests such as the Levene test are more appropriate for small-to-moderate sample sizes. For sample sizes of at least 30, though, the R test is more powerful than the Levene test, and may be implemented using the asymptotic χ_1^2 distribution to obtain critical values and p-values.

If normality cannot be assumed, then the F test is no longer an optimal test, whereas the R test is. For moderate sample sizes of at least 30 in each sample, the R test has test size very close to the nominal and is more powerful than both the F and Levene tests. It should then be the test of choice.

References

[1] BOOS, Dennis D. And BROWNIE, Cavell. (1989). Bootstrap methods for testing homogeneity of variances. *Technometrics*, 31, 1, 69-82.

[2] CONOVER, W.J., JOHNSON, Mark E. and JOHNSON, Myrle M. (1981). A comparative study of tests of homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, 23, 4, 351- 361.

[3] MOORE, D.S. and McCABE, G.P. (2006). *Introduction to the Practice of Statistics*. New York: W.H. Freeman.

[4] RAYNER, J.C.W. (1997). The Asymptotically Optimal Tests. *J.R.S.S., Series D (The Statistician)*, 46(3), 337-346.

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