Bonett's Method

ON BONETT'S ROBUST CONFIDENCE INTERVAL FOR A RATIO OF STANDARD DEVIATIONS

Senin J. Banga and Gregory D. Fox
June 18, 2013

ABSTRACT

We propose an alternative procedure to correct a subtle misstep in Bonett's (2006) confidence interval (CI) for the ratio of two standard deviations. The pooled kurtosis estimator for Layard's (1973) test statistic, on which Bonett's interval is based, is consistent only when the population variances are equal. We derive an alternative estimator that is consistent when the population variances are equal and when they are unequal and use the new estimator to calculate the correct CI. Simulation studies reveal that the new CI is, in general, more accurate and more precise than the CI based on the Levene/Browne-Forsythe test W_{50} and Pan's (1999) test L_{50}. Consistent with Pan, we observe that CIs based on test W_{50} display a loss of precision with small samples, often resulting in intervals that have infinite width. CIs that are based on test L_{50} perform well with symmetric and nearly symmetric distributions, but perform poorly when the populations are skewed.

Index terms: homogeneity of variances, Levene's test, Brown-Forsythe test, Layard's test, confidence interval (CI) for the ratio of variances

1.Introduction

It is widely known that the classical F test, and associated confidence intervals (CIs) are extremely sensitive to departures from normality—so sensitive, in fact, that the classical F test is not appropriate for most practical applications. For this reason, many have proposed more robust alternatives. Among these, the test known as "Test W50" is often preferred because it has very good type I error properties, yet is simple to calculate, and is simple to interpret. (For comparative analyses, see Conover et al. (1981), Balakrishnan and Ma (1990), and Lim and Loh (1996).) Test W_{50} is based on a procedure that was originally proposed by Levene (1960) and later enhanced by Brown and Forsythe (1974). Test W_{50} has been widely adopted and is available

in most well-known statistical software packages, such as Minitab Statistical Software, SAS, R, and JMP.

The type II error properties of test W_{50} are somewhat less remarkable than its type I error properties. Pan (1999) shows that, for some distributions, including the normal distribution, the power of test W_{50} in two-sample problems has an upper bound that is possibly far below 1. And this upper bound is not affected by the magnitude of the difference between the population variances. This deficiency naturally extends to CIs that are based on test W_{50}. Pan shows that there is a non-negligible probability that a CI for the ratio of the population variances that is based on test W_{50} will be infinite $(0, +\infty)$, and thus uninformative. Pan's observation is consistent with the results of our own simulations, which we report later in this paper.

Pan proposes an alternative procedure, called L_{50}, to correct the limitations of the W_{50} procedure. Based on simulation results, Pan concludes that test L_{50} is more powerful than test W_{50}, yet is equally robust and shares its desired asymptotic properties. The samples for Pan's simulations, however, were drawn from symmetric or mildly skewed distributions with heavy to light tails. The potential impact of skewness on the performance of the L_{50} test in small samples was not specifically discussed.

Pan also argues that the L_{50} procedure is as powerful as other, notably robust procedures such as the modified Fligner-Killeen rank test and the Hall-Padmanabhan adaptive test. Practically, however, the modified Fligner-Killeen rank test and the Hall-Padmanabhan adaptive test are somewhat less useful than tests L_{50} and W_{50} because they are computationally laborious and intensive.

Recently, Bonett (2006) proposed an alternative CI procedure that is based on the two-sample version of Layard's (1973) test of the homogeneity of variances. Bonett includes several adjustments to improve the small-sample performance of Layard's procedure. For example, Bonett proposes a pooled kurtosis estimator that is asymptotically equivalent to Layard's, but which displays less small-sample bias.

Unfortunately, neither Layard's original pooled kurtosis estimator, nor Bonett's proposed replacement are consistent when population variances are not equal. Thus, the intervals that Bonett (2006) proposes are not proper CIs, but are better described as acceptance intervals for the test of the equality of variances. Thus, subtracting the simulated coverage probabilities reported in Bonett (2006) from unity yields the type I error rates for the test of the equality of variances. Comparing these type I error rates to those of Layard's original test confirms that Bonett's adjustments successfully enhance the small-sample performance of Layard's test. The CI for the ratio of the variances proposed by Bonett, however, must be revisited.

Note also that Bonett compares the proposed intervals to CIs based on Shoemaker's (2003) approximate F test. However, the CI for the variance ratio associated with Shoemaker's test—as briefly described on page 106 of Shoemaker's article—is also based on Layard's pooled kurtosis estimator. Therefore, the CIs calculated in section 7 of Shoemaker's paper are also best described as acceptance intervals for the test of the equality of variances. Despite these errors, one can conclude from Bonett's simulation results that his adjustment improved the small-

sample performance of Layard's test of the equality of variances and that the resulting test for the equality of variances performs better than does Shoemaker's test.

In the present paper, we correct the misstep in Bonett (2006) by extending the two-sample form of Layard's test to test null hypotheses about the ratio of the variances or standard deviations. To accomplish this, we propose a pooled kurtosis estimator that is consistent for any given hypothesized ratio. We then invert the test statistic to obtain the CI for the ratio. Finally, we conduct simulation studies to assess the robustness properties of the new CI in small-sample designs. Moreover, we compare the small-sample performance of the new CI to the performance of CIs associated with the classical F test, test W_{50}, and test L_{50}.

2. Layard Test and Some Extension

Let $Y_{i1}, \ldots, Y_{in_i}, \ldots, Y_{k1}, \ldots, Y_{kn_k}$ be k independent samples, each sample being independent and identically distributed with mean $E(Y_{ij}) = \mu_i$ and variance $\mathrm{Var}(Y_{ij}) = \sigma_i^2 > 0$. In addition, suppose that the samples originate from populations with a common kurtosis $\gamma = E(Y-\mu)^4/\sigma^4 < \infty$. We note that Layard uses the kurtosis excess $\gamma_e = \gamma - 3$.

Let $\bar{Y}_i$ and S_i be the mean and standard deviation of sample i, respectively. Also, let $\tau^2 = 2 + (1 - 1/\bar{n})\gamma_e = 2 + (1 - 1/\bar{n})(\gamma - 3)$, where $\bar{n} = \sum n_i/k$. As indicated in Layard (1973), for large samples, $\tau^2 \cong \mathrm{Var}((n_i - 1)^{1/2} \ln S_i^2)$.

To test the null hypothesis of equality of variances, Layard performs an orthogonal transformation on the vector whose components $Z_i = (n_i - 1)^{1/2} \ln S_i^2 \,/\tau$ are asymptotically distributed as the standard normal distribution under the null hypothesis. Then he uses the distance preservation property of orthogonal transformations to show that the test statistic S' (given below) is asymptotically distributed as a chi-square distribution with $k - 1$ degrees of freedom under the null hypothesis of equality of variances:

$$S' = \sum_{i=1}^{k} (n_i - 1)\left(\ln S_i^2 - \frac{\sum_{i=1}^{k}(n_i - 1)\ln S_i^2}{\sum_{i=1}^{k}(n_i - 1)}\right)^2 /\tau^2$$

In general, $Z_i = (n_i - 1)^{\frac{1}{2}}(\ln S_i^2 - \ln \sigma_i^2)/\tau$ is asymptotically distributed as the standard normal distribution. Therefore, one can apply Layard's techniques to derive the more generalized test statistic T_k':

$$T_k' = \sum_{i=1}^{k} (n_i - 1)\frac{\left(\ln S_i^2 - \ln \sigma_i^2\right)^2}{\tau^2} - \left(\sum_{i=1}^{k} (n_i - 1)\frac{\ln S_i^2 - \ln \sigma_i^2}{\tau\sqrt{\sum_{i=1}^{k}(n_i - 1)}}\right)^2$$

T_k' is asymptotically distributed as a chi-square distribution with $k - 1$ degrees of freedom under both the null hypothesis and the alternative hypothesis.

One can express T_k' in a form that is more similar to that of S'. Expressing the squared term as a double sum and performing some algebra yields the following:

$$T_k' = \sum_{i=1}^{k}(n_i - 1)\left(\ln S_i^2 - \ln \sigma_i^2 - \frac{\sum_{i=1}^{k}(n_i - 1)\,(\ln S_i^2 - \ln \sigma_i^2)}{\sum_{i=1}^{k}(n_i - 1)}\right)^2 / \tau^2$$

If all the variances are equal, then $T_k' = S'$. Therefore, $S'$and T_k' are the same test statistic when testing the null hypothesis of equality of variances. However, T_k' can also be used more generally to test any hypotheses that are expressed as functions of the variances. For example, one can use T_k' to test any null hypothesis in the form $H_0: \sigma_i = \sigma_{0i}$ for any given $\sigma_{0i} > 0, i = 1, \ldots, k$.

Because $\tau^2 = 2 + (1 - 1/\bar{n})(\gamma - 3)$ is unknown, a test based on S' or T_k' requires an estimator for the common kurtosis of the populations, γ. For example, to test the null hypothesis of homogeneity of variances, Layard proposes the following pooled estimator of the common kurtosis:

$$\hat{\gamma} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^4}{\left[\sum_{i}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2\right]^2}\sum_{i=1}^{k} n_i$$

Layard points out, however, that $\hat{\gamma}$ is not necessarily a consistent estimator of the common kurtosis when the variances are not equal.

In the special case of two-sample designs, one may assess the magnitude of the difference between the standard deviations by testing the null hypothesis $H_0: \sigma_1/\sigma_2 = \rho_0$ for some given hypothesized ratio $\rho_0 > 0$. However, one may assess this difference more directly by calculating the CI for the ratio of the standard deviations.

If $\rho_0 = 1$, then the null hypothesis is equivalent to the hypothesis of homogeneity of variance. Therefore, one can base the test on $T_2' = S'$, after one substitutes the two-sample version of Layard's kurtosis estimator for γ in the expression of $\tau^2 = 2 + (1 - 1/\bar{n})(\gamma - 3)$ to obtain $\hat{\tau}^2$.

However, if $\rho_0 \neq 1$, then the test must be based on T_2' rather than S'. In addition, if $\rho_0 \neq 1$, then Layard's pooled kurtosis estimator is not necessarily consistent, and thus cannot be used to estimate the common kurtosis of the populations. Therefore, an alternative pooled kurtosis estimator—one that is consistent for any hypothesized ratio $\rho_0 > 0$—is required.

We next derive such an estimator. Because it is a function of ρ_0, we denote the estimator as $\hat{\gamma}_P(\rho_0)$. We also define the test statistic $T_2 = \tau^2 T_2'/\hat{\tau}^2$, where $\hat{\tau}^2 = 2 + (1 - 1/\bar{n})(\hat{\gamma}_P(\rho_0) - 3)$. By Slutzky's theorem, T_2 is asymptotically distributed as a chi-square distribution with 1 degree of freedom. Finally, we invert T_2 to obtain CIs for $\rho = \sigma_1/\sigma_2$.

3. CI for the Ratio of the Standard Deviations

The previous section details the need for an alternative kurtosis estimator when testing null hypotheses that are stated in terms of the ratio of the variances or standard deviations. The following result provides that estimator.

RESULT 1

For any given $\rho = \sigma_1/\sigma_2 > 0$, a consistent pooled kurtosis estimator of the common population kurtosis in the two-sample model may be given as

$$\hat{\gamma}_P(\rho) = (n_1 + n_2)\frac{\sum_{j=1}^{n_1}\left(Y_{1j} - \bar{Y}_1\right)^4 + \rho^4 \sum_{j=1}^{n_2}\left(Y_{2j} - \bar{Y}_2\right)^4}{[(n_1 - 1)S_1^2 + \rho^2(n_2 - 1)S_2^2]^2}$$

The proof for this result can be found in Appendix A.

As expected, $\hat{\gamma}_P(1)$ is identical to Layard's pooled kurtosis estimator, $\hat{\gamma}$, since $\sigma_1/\sigma_2 = 1$ implies that the standard deviations (or variances) are equal.

The statistic T_2', which is the two-sample version of the general statistic T_k', is given as

$$T_2' = \frac{(\ln \hat{\rho}^2 - \ln \rho^2)^2}{\left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1}\right)\tau^2}$$

where $\hat{\rho} = S_1/S_2$, $\rho = \sigma_1/\sigma_2$, and $\tau^2 = 2 + (1 - 1/\bar{n})\hat{\gamma}_e = 2 + (1 - 1/\bar{n})(\gamma - 3)$.

As indicated in Layard (1973), in large samples, $\tau^2 \cong \mathrm{Var}\left((n_i - 1)^{1/2} \ln S_i^2\right)$. Bonett (2006) uses an alternative approximation, which is also adopted in Shoemaker (2003), $\mathrm{Var}\left((n_i - 1)^{1/2} \ln S_i^2\right) \cong \gamma - (n_i - 3)/n_i$. In large samples, these approximations are equivalent. However, Shoemaker reports that the latter version is advantageous when using his test of the equality of variances with small samples. Using this adjustment, the statistic T_2' can be modified as

$$T_2' = \frac{(\ln \hat{\rho}^2 - \ln \rho^2)^2}{\frac{\gamma - g_1}{n_1 - 1} + \frac{\gamma - g_2}{n_2 - 1}}$$

where $g_i = (n_i - 3)/n_i$.

It follows then, that the test statistic $T_2 = \tau^2 T_2'/\hat{\tau}^2$ for testing the null hypothesis $H_0: \rho = \rho_0$ is given as

$$T_2 = \frac{(\ln \hat{\rho}^2 - \ln \rho_0^2)^2}{\frac{\hat{\gamma}_P(\rho_0) - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P(\rho_0) - g_2}{n_2 - 1}}$$

In this expression of T_2, the square root of the denominator can be viewed as a large-sample estimate of the standard error for the pooled kurtosis.

Moreover, in the expression of $\hat{\gamma}_P(1) \equiv \hat{\gamma}$, Bonett (2006) uses the trimmed sample means with the trim proportion $1/[2(n_i - 4)^{1/2}]$. Accordingly, we make the same adjustment to the pooled kurtosis estimator:

$$\hat{\gamma}_P(\rho) = (n_1 + n_2)\frac{\sum_{j=1}^{n_1}(Y_{1j} - m_1)^4 + \rho^4 \sum_{j=1}^{n_2}(Y_{2j} - m_2)^4}{[(n_1 - 1)S_1^2 + \rho^2(n_2 - 1)S_2^2]^2}$$

where m_i is the trimmed mean for sample i, with the trim proportion $1/[2(n_i - 4)^{1/2}]$. This version of the pooled kurtosis estimator and the earlier version are asymptotically equivalent since the trimmed mean m_i is a consistent estimator of the population mean μ_i. This alternative version, however, may improve the small-sample performance of the test based on T_2.

The test statistic T_2 may now be inverted to derive an approximate CI for the ratio of the variances or standard deviations. But, first, we briefly describe the misstep in the derivation of the Bonett (2006) CIs for the ratio of the standard deviations.

3.1 Bonett's Intervals

Rather than inverting T_2 to obtain the CI, Bonett (2006) inverts the following statistic

$$T = \frac{(\ln \hat{\rho}^2 - \ln \rho^2)^2}{\frac{\hat{\gamma}_P(1) - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P(1) - g_2}{n_2 - 1}}$$

Consequently, the resulting interval is simply the acceptance region for the test of the equality of variances. This is because the pooled kurtosis estimator $\hat{\rho}_P(1)$ is consistent only when the variances are equal, or equivalently when the hypothesized ratio is 1. The resulting interval is reported in Bonett (2006) as

$$\exp\left[\ln(c\, S_1^2/S_2^2) \pm z_{\alpha/2} se\right]$$

where

$$se^2 = \frac{\hat{\gamma}(1) - g_1}{n_1 - 1} + \frac{\hat{\gamma}(1) - g_2}{n_2 - 1}$$

The constant c is included as a small-sample adjustment to mitigate the effect of unequal tail error probabilities in unbalanced designs. This constant is given by

$$c = \frac{n_1}{n_1 - z_{\alpha/2}} \frac{n_2 - z_{\alpha/2}}{n_2}$$

The constant vanishes when the designs are balanced and its effect becomes negligible with increasing sample sizes.

Table 1 illustrates the consequences of misinterpreting the above intervals as CIs. These results are based on a small simulation study in which we compute simulated coverage probabilities based on Bonett's (2006) intervals. For the equal variance cases (left column), we draw two independent samples from the standard normal distribution. For the unequal variance cases (right column), we scale the observations of the second sample by a constant factor of 4. The

estimated coverage probabilities are based on 100,000 replicates. The targeted nominal coverage is 0.95.

Table 1 Effect of Unequal Population Variances on Bonett's (2006) CIs ($\alpha = 0.05$)

	Simulated Coverage Probabilities	
n_1, n_2	Equal Variances	Unequal Variances
10, 10	0.963	0.972
50, 50	0.952	0.991
100, 100	0.952	0.994

If the intervals were based on a consistent pooled kurtosis estimator, then one would expect the coverage probabilities in the two cases to be identical. However, notice that the intervals are consistently more conservative when the variances are unequal. Furthermore, the coverage probabilities approach 1 as the sample sizes increase. Note that similar results are obtained with Shoemaker's (2003) approximate CIs.

3.2 Calculations for the CI

Consider the problem of testing the null hypothesis H_0: $\rho = \rho_0$ against the alternative hypothesis H_A: $\rho \neq \rho_0$, where $\rho = \sigma_1/\sigma_2$ and $\rho_0 > 0$, based on the test statistic T_2 given earlier. Under the null hypothesis, the test statistic

$$T_2 = \frac{(\ln \hat{\rho}^2 - \ln \rho_0^2)^2}{\frac{\hat{\gamma}_P(\rho_0) - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P(\rho_0) - g_2}{n_2 - 1}}$$

is asymptotically distributed as a chi-square distribution with 1 degree of freedom. Thus, the test rejects the null hypothesis at the α level of significance if and only if

$$(\ln \hat{\rho}^2 - \ln \rho_0^2)^2 > z_{\alpha/2}^2 \left(\frac{\hat{\gamma}_P(\rho_0) - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P(\rho_0) - g_2}{n_2 - 1} \right)$$

where z_α denotes the $\alpha \times 100$th upper percentile point of the standard normal distribution. Note that the $\alpha \times 100$th upper percentile point of the chi-square distribution with 1 degree of freedom, $\chi_{1,\alpha}^2$, satisfies the following condition: $\chi_{1,\alpha}^2 = z_{\alpha/2}^2$.

Bonett's (2006) simulation results show that the small-sample adjustment to reduce the effect of unequal tail error probabilities in unbalanced designs worked well. Thus, we make a similar adjustment for the test based on T_2. When this adjustment is made, the test rejects the null hypothesis if an only if

$$(\ln \rho_0^2 - \ln(c\hat{\rho}^2))^2 > z_{\alpha/2}^2 \left(\frac{\hat{\gamma}_P(\rho_0) - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P(\rho_0) - g_2}{n_2 - 1} \right)$$

where c is Bonett's adjustment constant given as

$$c = \frac{n_1}{n_1 - z_{\alpha/2}} \frac{n_2 - z_{\alpha/2}}{n_2}$$

Equivalently, an approximate $(1-\alpha)100$ percent confidence set for $\rho = \sigma_1/\sigma_2$ based on T_2 is given by

$$\left\{\rho \in (0,\infty): (\ln \rho^2 - \ln(c\hat{\rho}^2))^2 - z_{\alpha/2}^2 \left(\frac{\hat{\gamma}_P(\rho) - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P(\rho) - g_2}{n_2 - 1}\right) \leq 0\right\}$$

Note that c has no effect in balanced designs and has only a negligible effect in large-sample unbalanced designs.

The next result provides an alternative expression of the confidence set in a form that is convenient for describing its nature. In this expression, the pooled kurtosis estimator is rewritten in terms of the individual sample kurtoses given as

$$\hat{\gamma}_i = n_i \frac{\sum_{j=1}^{n_i} (Y_{ij} - m_i)^4}{\left[(n_i - 1)S_i^2\right]^2}, i = 1,2$$

RESULT 2

An approximate $(1-\alpha)100$ percent confidence set for $\rho = \sigma_1/\sigma_2$ based on T_2 may be expressed as

$$\hat{\rho}\sqrt{c}\,\{r \in (0,\infty): H(r^2) \leq 0\}$$

or equivalently, the confidence set for $\rho^2 = \sigma_1^2/\sigma_2^2$ may be expressed as

$$c\hat{\rho}^2\,\{r \in (0,\infty): H(r) \leq 0\}$$

where

$$H(x) = (\ln x)^2 - z_{\alpha/2}^2 se^2(cx), x > 0$$

$$se^2(x) = A\frac{\hat{\gamma}_1 K^2/n_1 + \hat{\gamma}_2 x^2/n_2}{(K+x)^2} - B$$

$$A = \frac{(n_1+n_2)(n_1+n_2-2)}{(n_1-1)(n_2-1)}, B = \frac{g_1}{n_1-1} + \frac{g_2}{n_2-1}, K = \frac{n_1-1}{n_2-1}$$

For the proof of this result, see Appendix B.

It is easily verified that the function $H(x)$ is continuous on the positive real line, with $H(0) = H(+\infty) = +\infty$ and $H(1) < 0$. Therefore, by the intermediate values theorem, the function $H(x)$ admits at least one root in the interval $(0,1)$ and at least one root in the interval $(0,+\infty)$.

The next result describes the confidence set as an interval or union of disjoint intervals.

RESULT 3

If the function $H(x)$ has exactly two roots, x_L and x_U, then $0 < x_L < 1 < x_U$ and the confidence set for $\rho^2 = \sigma_1^2/\sigma_2^2$ is the interval given by

$$[c\hat{\rho}^2 x_L, c\hat{\rho}^2 x_U]$$

It follows, then, that the CI for $\rho = \sigma_1/\sigma_2$ is the interval,

$$\left[\hat{\rho}\sqrt{cx_L}, \hat{\rho}\sqrt{cx_U}\right]$$

On the other hand, if the function $H(x)$ has more than two roots, then the confidence set for $\rho^2 = \sigma_1^2/\sigma_2^2$ is the union of non-overlapping intervals. The endpoints of each interval are the consecutive roots where the function opens upward.

For the proof of this result, see Appendix C.

REMARK

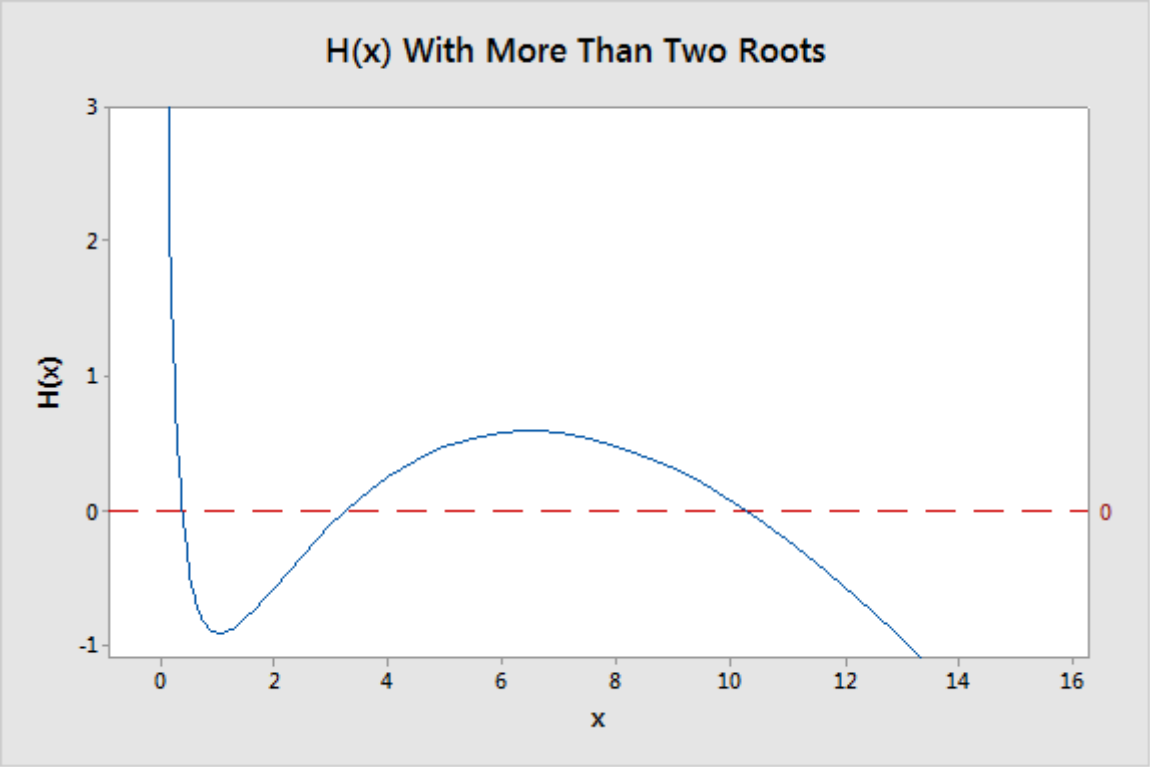
Although it is mathematically possible for the function $H(x)$ to admit more than two roots, we have observed that this occurs only with extremely unusual and practically meaningless designs where one or both samples are either too small or severely unbalanced. We conjecture that $H(x)$ has either two or four roots.

The following example is based on data that were fabricated to force the function $H(x)$ to have more than two roots. The data are summarized as follows: $n_1 = 169,\ n_2 = 7,\ S_1 = 301.855,$ $S_2 = 4606.170,\ \hat{\gamma}_1 = 1.877,\ \hat{\gamma}_2 = 6.761,\ c = 0.728,\ A = 30.381,\ B = 0.101,$ and $K = 28.000.$

For $\alpha = 0.05$, the function $H(x)$ is given as

$$H(x) = (\ln x)^2 - 1.960^2\left(30.381\frac{1.877 \times 28^2/169 + 6.761 \times (.728x)^2/7}{(28.000 + 0.728x)^2} - 0.101\right)$$

The function $H(x)$ in this case has four roots. The graph of the function is displayed below. Note that the fourth root is not visible on the graph because it is too large. However, we know that the fourth root exists because $H(+\infty) = +\infty$.



The four roots are numerically computed as $x_1 = 0.389$, $x_2 = 3.282$, $x_3 = 10.194$, and $x_4 = 39685.0$. The estimated ratio of the standard deviations is $\hat{\rho} = S_1/S_2 = 0.066$. The confidence set for $\rho^2 = \sigma_1^2/\sigma_2^2$ may be expressed as

$$[c\,\hat{\rho}^2 x_1, c\,\hat{\rho}^2 x_2] \cup [c\,\hat{\rho}^2 x_3, c\,\hat{\rho}^2 x_4] = [0.001, 0.010] \cup [0.032, 124.072]$$

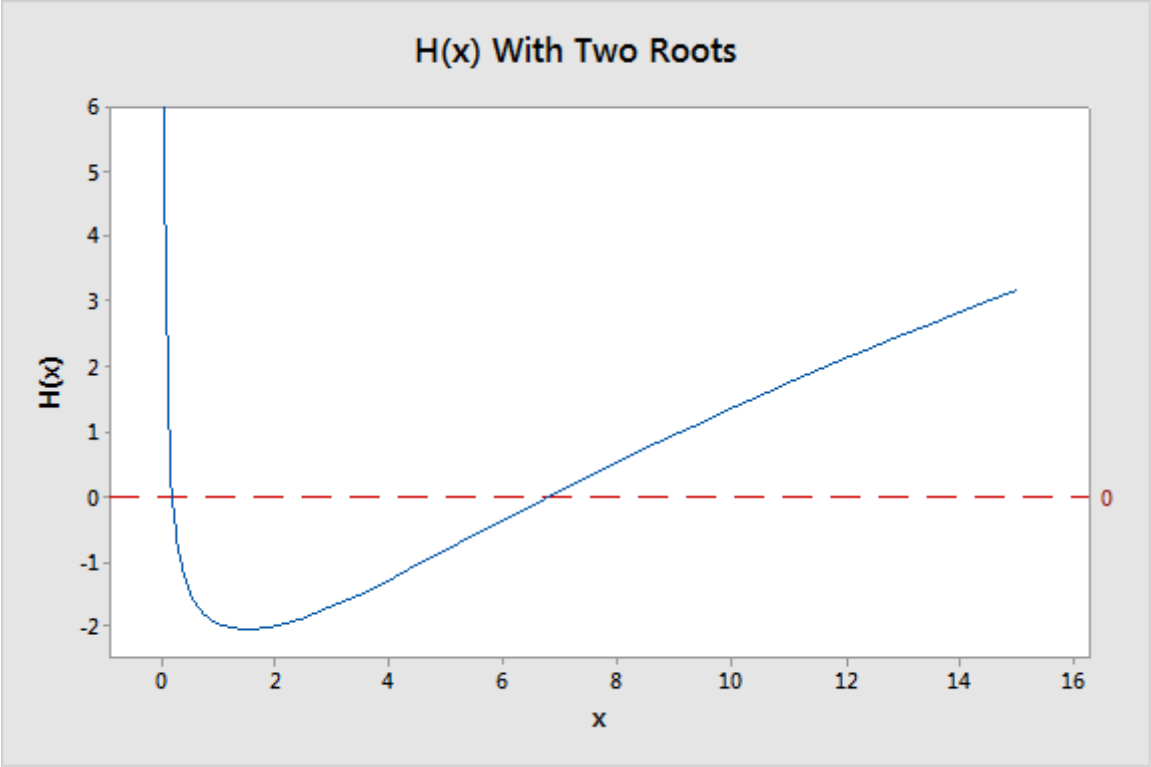
The confidence set for the ratio of the standard deviations, ρ, is obtained by taking the square root of the endpoints of the intervals.

When the samples are not too small ($n_i \geq 10$) and the disparity between their sizes is not great, the function $H(x)$ typically admits two roots. One root is below unity, and the other root is above unity as described in Result 2. Here is an example that is based on randomly generated data. The data can be summarized as follows: $n_1 = 10$, $n_2 = 12$, $S_1 = 1.150$, $S_2 = 1.043$, $\hat{\gamma}_1 = 2.704$, $\hat{\gamma}_2 = 3.671$, $c = 1.041$, $A = 4.444$, $B = 0.146$, and $K = 0.818$.

For $\alpha = 0.05$, the function $H(x)$ is given in this case as

$$H(x) = (\ln x)^2 - 1.960^2 \left(4.444 \frac{2.704 \times 0.818^2/10 + 3.671 \times (1.041x)^2/12}{(0.818 + 1.041x)^2} - 0.146 \right)$$

The function $H(x)$ has two roots as shown below:



The two roots are numerically computed as $x_1 = 0.200$ and $x_2 = 6.824$. The estimated ratio of the standard deviations is $\hat{\rho} = S_1/S_2 = 1.102$. The confidence set for $\rho^2 = \sigma_1^2/\sigma_2^2$ is the interval given as

$$[c\,\hat{\rho}^2 x_1, c\,\hat{\rho}^2 x_2] = [0.253, 8.634]$$

The CI for the ratio of the standard deviations, ρ, is obtained by taking the square root of the endpoints of the above interval.

We now describe two algorithms for finding the confidence limits.

The first algorithm consists of using a numerical root finder procedure to find the roots of the function $H(x)$. The root that corresponds to a lower confidence limit for the ratio of the variances is confined in the interval $(0, 1)$. If we denote this root by x_L, then, by Result 3, the lower confidence limit for the ratio of the variances is calculated as $c\hat{\rho}^2 x_L$, and the lower confidence limit for the ratio of the standard deviation is obtained as $\hat{\rho}\sqrt{cx_L}$. Similarly, the upper confidence limit for the ratio of the variances is $c\hat{\rho}^2 x_U$, and the upper confidence limit for the ratio of the standard deviations is $\hat{\rho}\sqrt{cx_U}$, where $x_U > 1$ is the other root of $H(x)$. A simple approach for finding the upper confidence limit is to use the fact that the lower limit for $1/\rho^2$ is the upper limit for ρ^2. First, the roles of the first sample and the second sample are interchanged in the expression of the function $H(x)$ as if one were computing the confidence limit for the ratio $1/\rho^2 = \sigma_2^2/\sigma_1^2$. Second, the algorithm for finding the lower bound is applied to the new function $H(x)$. Finally, the resulting limit is inverted to obtain the desired upper confidence limit.

An alternative approach consists of recursively calculating the lower confidence limit for the ratio of the variances using the recurrence relation given by

$$\rho_0^2 = 1$$

$$\rho_{i+1}^2 = \exp\left[\ln(c\,\hat{\rho}^2) - z_{\alpha/2}\sqrt{\frac{\hat{\gamma}_P(\rho_i) - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P(\rho_i) - g_2}{n_2 - 1}}\right], i = 0,1,2,\ldots$$

The lower confidence limit for the ratio of the variances is ρ_{j+1}^2, such that $\left|\rho_{j+1}^2 - \rho_j^2\right| < \varepsilon$, where $j > 0$ and ε is chosen to be small (for example $\varepsilon = 10^{-6}$). To find the upper confidence limit, we simply replace $-z_{\alpha/2}$ with $+z_{\alpha/2}$ in the above.

Evidently, the two algorithms for computing the confidence limits are equivalent because the recursive procedure is essentially an iterative procedure for solving the equation $H(\rho^2/(c\hat{\rho}^2)) = 0$ for ρ^2. The recursive algorithm is easier to implement, and therefore is a useful alternative when a root finder procedure is not available.

4. Simulation Studies and Results

In this paper, we derived a procedure to extend Layard's test for equality of two variances to test the ratio of variances. We call this procedure the Extended Layard's Test for the Ratio, or ELTR. In this section, we investigate the small-sample properties of CIs based on the ELTR procedure. We follow the general approach taken by Bonett (2006).

We compare CIs that are based on the ELTR procedure to CIs that are based on test L_{50} (Pan, 1999) and test W_{50} (the Levene/Brown-Forsythe test). For Study 1, we also include, for comparison, CIs that are based on the classical F test. It is well-known that, when the data are normally distributed, the classical F test is optimal. Note that the calculations for CIs based on tests $W_{50}$and L_{50} are given in Pan (1999). The calculations for CIs based on the F test can be found in many introductory statistics text books. They are also given in Bonett (2006).

We conducted three simulation studies, each with 100,000 sampling replicates. Each replicate consists of two independent samples that are small to moderate in size. Each sample was drawn from a parent distribution with known properties including symmetry, asymmetry, heavy tails, and light tails. The standard error associated with each simulation is approximately 0.0009, 0.0007, and 0.0003 for nominal confidence levels of 90%, 95%, and 99%, respectively.

To evaluate the performance of each procedure, we report the achieved coverage probability and the mean width of the simulated intervals for the ratio of the variances. Some of the intervals associated with test W_{50} had infinite width (a possibility exposed by Pan (1999)). In such cases, we report both the mean width of the finite intervals and the percentage of intervals with infinite width. All simulations were conducted using Version 8 of the Mathematica software package.

Study 1: Comparison of Coverage Probabilities for Normal Data

In the first study, we generate random samples of various sizes from the normal distribution. The results are presented in Table 2.

Table 2 Comparison of Coverage Probabilities and Average Interval Widths

			Procedure			
$1-\alpha$	n_1, n_2	Measure	F	ELTR	L_{50}	W_{50}
0.90	10, 10	Coverage	0.898	0.918	0.913	0.921
		Width	*3.72*	*5.06*	*4.72*	*8.03 (0.01%)*
	30, 10	Coverage	0.900	0.909	0.897	0.911
		Width	*2.42*	*3.01*	*3.58*	*3.17*
	25, 25	Coverage	0.902	0.907	0.914	0.916
		Width	*1.61*	*1.73*	*1.85*	*1.938*
	50, 50	Coverage	0.900	0.901	0.906	0.907
		Width	*1.03*	*1.06*	*1.13*	*1.15*
0.95	10, 10	Coverage	0.949	0.963	0.958	0.964
		Width	*4.90*	*7.72*	*6.52*	*497.24 (0.20%)*
	30, 10	Coverage	0.950	0.957	0.945	0.959
		Width	*2.98*	*4.91*	*4.67*	*4.07*
	25, 25	Coverage	0.951	0.955	0.958	0.961
		Width	*1.99*	*2.24*	*2.31*	*2.49*
	50, 50	Coverage	0.951	0.952	0.953	0.954
		Width	*1.25*	*1.31*	*1.38*	*1.41*

			Procedure			
$1-\alpha$	n_1, n_2	Measure	F	ELTR	L_{50}	W_{50}
0.99	10, 10	Coverage	0.989	0.993	0.992	0.994
		Width	*8.29*	*17.76*	*12.52*	*$> 10^4$ (8.8%)*
	30, 10	Coverage	0.990	0.992	0.986	0.994
		Width	*4.26*	*15.76*	*8.26*	*6.77*
	25, 25	Coverage	0.990	0.992	0.992	0.993
		Width	*2.86*	*3.66*	*3.43*	*4.03*
	50, 50	Coverage	0.990	0.991	0.991	0.991
		Width	*1.71*	*1.89*	*1.92*	*2.02*

The shaded rows display the achieved coverage probabilities (Coverage) for each procedure at each confidence level ($1-\alpha$) and each combination of sample sizes (n_1, n_2). The mean of the interval widths (Width) is displayed below each coverage probability. If any intervals for a condition were infinite, then we report both the mean for the finite intervals and the percentage of intervals that were infinite.

As expected, the results show that the CIs associated with the F procedure are the most accurate and the most precise. The coverage probabilities achieved with the F procedure are closer to the target coverage than are those associated with the other procedures. And the average widths of the intervals associated with the F procedure are smaller than those associated with the other procedures. The table also reveals, however, that CIs that are constructed using the ELTR and L_{50} procedures are almost as accurate and precise as those based on the F procedure.

The intervals based on test W_{50} are also fairly accurate. However, W_{50} intervals can be very wide and can even have infinite width, depending on the size of the samples. Note that, when both samples have only 10 observations, at least 0.01% of the intervals produced by the W_{50} procedure are infinitely wide. And the percentage of infinite intervals increases when the target coverage increases. Under most conditions, the mean widths of the ELTR and L_{50} intervals are smaller than the mean widths of the W_{50} intervals.

Study 2: Comparison of Coverage Probabilities for Nonnormal Data

The second study is designed to evaluate and compare the performance of the ELTR, L_{50}, and W_{50} procedures when parent distributions are not normal. We also include a contaminated normal distribution in order to assess the impact of outliers on the performance of the procedures. We denote this contaminated distribution as CN(0.1, 3) to indicate that, while 90%

of observations are drawn from the standard normal distribution, the remaining 10% are drawn from a normal population with a mean of 0 and a standard deviation of 3. The results are presented in Table 3.

Table 3 Comparison of Coverage Probabilities and Average Interval Widths in some Nonnormal Models Nominal Confidence Level is $1 - \alpha = 0.95$

Distribution $[\gamma]$ n_1, n_2	**ELTR**	L_{50}	W_{50}	**Distribution** $[\gamma]$ n_1, n_2	**ELTR**	L_{50}	W_{50}
Uniform				$\chi^2(5)$			
[1.8]				[5.4]			
10, 10	0.971	0.971	0.966	10, 10	0.956	0.938	0.956
	5.27	*4.87*	*42.08 (0.1%)*		*11.61*	*8.78*	*$> 10^4$ (2.6%)*
10, 30	0.964	0.961	0.957	10, 30	0.959	0.923	0.956
	2.51	*2.4*	*2.89*		*6.25*	*4.14*	*190.645 (0.3%)*
25, 25	0.967	0.972	0.968	25, 25	0.956	0.944	0.954
	1.43	*1.79*	*1.88*		*3.66*	*2.92*	*3.26*
50, 50	0.959	0.962	0.959	50, 50	0.959	0.946	0.952
	0.83	*1.06*	*1.08*		*2.07*	*1.7*	*1.77*
Beta (3, 3)				**Exponential**			
[2.5]				[9]			
10, 10	0.968	0.966	0.966	10, 10	0.947	0.916	0.950
	6.26	*5.59*	*254.62 (0.1%)*		*20.99*	*14.47*	*$> 10^4$ (9.1%)*
10, 30	0.960	0.954	0.960	10, 30	0.954	0.896	0.953
	3.14	*2.76*	*3.71*		*10.46*	*6.19*	*$> 10^4$ (4.1%)*
25, 25	0.959	0.966	0.965	25, 25	0.956	0.931	0.951
	1.81	*2.06*	*2.18*		*6.09*	*4.13*	*5.48 (0.008%)*
50, 50	0.957	0.959	0.958	50, 50	0.962	0.942	0.952
	1.06	*1.23*	*1.26*		*3.18*	*2.24*	*2.38*

Distribution [γ] n_1, n_2	ELTR	L_{50}	W_{50}	Distribution [γ] n_1, n_2	ELTR	L_{50}	W_{50}
Laplace				$\chi^2(1)$			
[6]				[15]			
10, 10	0.946	0.935	0.961	10, 10	0.928	0.889	0.947
	13.47	*10.45*	*> 10^4 (3.0%)*		*55.09*	*37.4*	*> 10^5 (25.1%)*
10, 30	0.947	0.919	0.957	10, 30	0.943	0.882	0.956
	6.78	*4.82*	*> 10^4 (0.4%)*		*18.71*	*11.14*	*> 10^6 (25.7%)*
25, 25	0.945	0.940	0.952	25, 25	0.952	0.925	0.954
	4.00	*3.372*	*3.86*		*10.97*	*6.84*	*> 10^4 (0.4%)*
50, 50	0.952	0.949	0.955	50, 50	0.958	0.936	0.951
	2.19	*1.91*	*1.99*		*5.08*	*3.31*	*3.75 (0.001%)*
$t(5)$				**Lognormal**			
[9]				[113.9]			
10, 10	0.957	0.946	0.965	10, 10	0.923	0.876	0.955
	11.07	*8.81*	*> 10^3 (2.0%)*		*59.22*	*46.15*	*> 10^5 (23.0%)*
10, 30	0.957	0.930	0.959	10, 30	0.949	0.866	0.958
	6.06	*4.24*	*> 10^3 (0.7%)*		*29.13*	*17.67*	*> 10^6 (31.6%)*
25, 25	0.954	0.948	0.960	25, 25	0.947	0.917	0.965
	3.54	*2.93*	*4.86 (0.01%)*		*16.21*	*8.73*	*> 10^4 (2.4%)*
50, 50	0.954	0.947	0.954	50, 50	0.955	0.928	0.960
	2.10	*1.71*	*1.77 (0.003%)*		*8.62*	*4.11*	*164.38 (0.2%)*

Distribution [γ] n_1, n_2	ELTR	L_{50}	W_{50}	Distribution [γ] n_1, n_2	ELTR	L_{50}	W_{50}
Half Normal				CN(0.1, 3)			
[3.9]				[8.3]			
10, 10	0.956	0.942	0.954	10, 10	0.977	0.965	0.979
	10.41	*7.89*	*> 10^4 (1.5%)*		*12.64*	*9.52*	*> 10^4 (4.9%)*
10, 30	0.959	0.930	0.954	10, 30	0.981	0.952	0.979
	5.18	*3.64*	*13.00 (0.02%)*		*7.82*	*4.71*	*944.68 (1.1%)*
25, 25	0.959	0.952	0.959	25, 25	0.982	0.972	0.981
	3.01	*2.62*	*2.88*		*4.63*	*3.22*	*3.71*
50, 50	0.960	0.951	0.954	50, 50	0.983	0.972	0.978
	1.69	*1.54*	*1.59*		*2.64*	*1.83*	*1.91*

The shaded rows display the achieved coverage probabilities for each procedure, parent distribution, and combination of sample sizes. The mean of the interval widths is displayed below each coverage probability. If any intervals for a condition were infinite, then we report both the mean for the finite intervals and the percentage of intervals that were infinite. The kurtosis (γ) of each parent distribution is displayed in brackets.

For symmetric, light-tailed distributions, the results indicate that all three methods yield similarly conservative coverage probabilities. However, the ELTR and L_{50} intervals are more precise for small samples than are the W_{50} intervals. For example, when samples are drawn from a Beta distribution with parameters of (3, 3), the achieved coverage probabilities for the ELTR and L_{50} intervals are at least as accurate as those of the W_{50} intervals, but the W_{50} intervals are consistently wider.

The ELTR and W_{50}intervals are also a bit conservative for symmetric, heavy-tailed distributions, while the L_{50} intervals are liberal. The L_{50} intervals are even more liberal when designs are unbalanced. For example, when samples of sizes 10 and 30 are drawn from the Laplace distribution, the achieved coverage probability for the L_{50} intervals is 0.919. And when the same sized samples are drawn from a t-distribution with 5 degrees of freedom, the achieved coverage probability for the L_{50} intervals is 0.930.

The L_{50} intervals are also quite liberal when small samples are drawn from highly skewed, heavy-tailed distributions. For example, when samples are drawn from a lognormal distribution, the achieved coverage can be as low as 0.866. For these distributions, the W_{50} method is the least liberal of the three methods. However, too many of the W_{50} intervals have infinite width. For example, when samples are drawn from the chi-square distribution with 1 degree of freedom

($\chi^2(1)$), more than 25% of the W_{50} intervals can have infinite width. The ELTR intervals are somewhat less accurate, but considerably narrower and thus more informative than the W_{50} intervals.

Finally, we note that all three procedures are adversely affected by outliers. The L_{50} method is the least affected, which might be expected because the L_{50} method was derived to reduce the effect of outliers on test W_{50} (Pan, 1999). When samples are drawn from the contaminated normal distribution, CN(0.1, 3), the minimum of the achieved coverage probabilities for the ELTR and W_{50} procedures is 0.977. Additional simulation results (not shown) indicate that these intervals improve only slowly with increasing sample sizes.

Study 3: Sensitivity to the Equal-Kurtosis Assumption

Our final study investigates the sensitivity of the ELTR procedure to the assumption of equal kurtosis under which it is derived. We examine the performance of the ELTR procedure when the kurtoses of the parent populations are not equal, that is when $\gamma_1 \neq \gamma_2$. We also include the L_{50} and W_{50} procedures, because they are derived under the assumption that the populations are similar. This similarity assumption is undermined when the kurtoses of the parent populations are not equal. The results are presented in Table 4.

Table 4 Sensitivity of the ELTR Procedure to the Equal-Kurtosis Assumption Nominal Confidence Level is $1 - \alpha = 0.95$

Dist. 1, Dist. 2 $[\gamma_1, \gamma_2]$ n_1, n_2	**ELTR**	L_{50}	W_{50}	**Dist. 1, Dist. 2** $[\gamma_1, \gamma_2]$ n_1, n_2	**ELTR**	L_{50}	W_{50}
Beta (3, 3), Normal [2.5, 3]				**Normal, CN (0.9, 3)** [3, 8.3]			
10, 10	0.964	0.961	0.964	10, 10	0.955	0.948	0.951
	0.27	*0.23*	*204.50 (0.20%)*		*6.88*	*5.16*	*> 10^4 (4.89%)*
30, 10	0.946	0.939	0.946	30, 10	0.941	0.910	0.942
	0.16	*0.17*	*0.15*		*5.26*	*3.77*	*3.20*
10, 30	0.966	0.956	0.967	10, 30	0.961	0.950	0.958
	0.14	*0.11*	*0.17*		*4.26*	*2.40*	*630.42 (1.10%)*
50, 50	0.951	0.950	0.949	50, 50	0.936	0.910	0.907
	0.04	*0.05*	*0.05*		*1.27*	*1.11*	*1.19*

Dist. 1, Dist. 2 $[\gamma_1, \gamma_2]$ n_1, n_2	ELTR	L_{50}	W_{50}	Dist. 1, Dist. 2 $[\gamma_1, \gamma_2]$ n_1, n_2	ELTR	L_{50}	W_{50}
Normal, Laplace [3, 6]				**Half Normal, $\chi^2(5)$** [3.9, 5.4]			
10, 10	0.941	0.935	0.947	10, 10	0.956	0.940	0.954
	6.67	*5.17*	*$> 10^6$ (2.90%)*		*0.42*	*0.32*	*304.41 (2.60%)*
30, 10	0.912	0.888	0.914	30, 10	0.954	0.918	0.949
	5.06	*3.85*	*3.21*		*0.33*	*0.22*	*0.20*
10, 30	0.963	0.943	0.955	10, 30	0.962	0.934	0.958
	3.33	*2.25*	*$> 10^3$ (0.40%)*		*0.23*	*0.15*	*3.28 (0.30%)*
50, 50	0.935	0.894	0.889	50, 50	0.955	0.941	0.945
	0.98	*1.04*	*1.12*		*0.07*	*0.06*	*0.07*
Normal, Half Normal [3, 3.9]				**$\chi^2(5)$, Exponential** [5.4, 9]			
10, 10	0.956	0.948	0.957	10, 10	0.938	0.914	0.940
	28.16	*20.65*	*$> 10^4$ (1.50%)*		*211.17*	*137.88*	*$> 10^6$ (9.10%)*
30, 10	0.946	0.924	0.947	30, 10	0.928	0.875	0.929
	20.59	*14.83*	*12.78*		*194.70*	*93.02*	*83.02*
10, 30	0.961	0.946	0.962	10, 30	0.968	0.930	0.954
	14.06	*9.37*	*49.11 (0.02%)*		*102.35*	*55.29*	*$> 10^5$ (3.90%)*
50, 50	0.953	0.950	0.952	50, 50	0.950	0.920	0.923
	4.32	*4.16*	*4.33*		*29.64*	*23.37*	*25.54*

The shaded rows display the achieved coverage probabilities for each procedure, combination of parent distributions (Dist. 1, Dist. 2), and combination of sample sizes. The mean of the interval widths is displayed below each coverage probability. If any intervals for a condition are infinite, then we report both the mean for the finite intervals and the percentage of intervals that were infinite. The kurtosis of each parent distribution (γ_1, γ_2) is displayed in brackets.

In general, the performance of the ELTR procedure does not appear to be adversely affected by unequal kurtoses when the samples are large enough. However, when designs are unbalanced and the smaller sample is obtained from the heavier-tailed distribution, the achieved coverage probabilities are liberal. The achieved coverage probabilities are better when the larger sample is drawn from the heavier-tailed distribution.

When sample sizes are large enough, the L_{50} and W_{50} intervals also appear to be generally robust to the dissimilarity of distributions that results from unequal kurtoses. However, note that when samples are drawn from a normal distribution and a Laplace distribution, or from a normal distribution and a contaminated normal distribution, the coverage probabilities for the L_{50} and W_{50} intervals are not stable, even for samples as large as 50.

The L_{50} intervals are generally more liberal than the ELTR and W_{50} intervals. In three cases, the achieved coverage probabilities for the L_{50} intervals are less than 0.90. In contrast, only one of the achieved coverage probabilities for the W_{50} intervals is less than 0.90. The lowest achieved coverage probability for the ELTR intervals is 0.912.

The previous study (Table 3) shows that all three procedures produce intervals that are noticeably more conservative when both samples are drawn from the contaminated normal distribution, CN(0.1, 3). The present study shows that all three procedures perform notably better when only one sample is drawn from CN(0.1, 3). However, note that the performance of the L_{50} and W_{50} intervals appears to degrade considerably when the sample size increases to 50.

5. Example

In this section, we apply all four procedures—F, ELTR, L_{50} and W_{50}—to a data set obtained from Pan (1999). Ott (1993, page 352) describes the data as follows:

A chemist at an iron mine suspects that the variance in the amount (weight, in ounces) of iron oxide per pound of ore tends to increase as the mean amount of iron oxide per pound increases. To test this theory, ten 1-pound specimens of iron ore are selected from each of two locations, one, location 1, containing a much higher mean content of iron oxide than the other, location 2. The amounts of iron oxide contained in the ore specimen are shown below:

Location 1	8.1	7.4	9.3	7.5	7.1	8.7	9.1	7.9	8.4	8.8
Location 2	3.9	4.4	4.7	3.6	4.1	3.9	4.6	3.5	4.0	4.2

The 95% CIs for $\sigma_2/\sigma_1 = 1/\rho$ calculated using the four different methods are given in the following table:

Procedure	95% CI
F	(0.262, 1.055)
ELTR	(0.277, 0.924)

Procedure	95% CI
L_{50} (Pan)	(0.295, 0.938)
W_{50} (Levene/Brown-Forsythe)	(0.237, 0.908)

6. Conclusion

Our simulations show that, in general, the CIs based on the ELTR procedure are as accurate as CIs derived from tests L_{50} and W_{50}. However, the ELTR intervals and the L_{50} intervals are more precise than the W_{50} intervals for most distributions. The W_{50} intervals tend to be more accurate than the ELTR intervals and the L_{50} intervals when small samples are drawn from severely skewed and heavy-tailed distributions. However, this advantage is usually offset by a remarkable loss of precision. The resulting W_{50} intervals are typically too wide and are likely to have infinite width.

As designed, the L_{50} intervals improve the precision of the W_{50} intervals. For skewed populations, however, the L_{50} intervals are so short that they yield excessively liberal coverage probabilities. In contrast, the ELTR intervals are more stable in general. The ELTR intervals are usually not too long or too short, so the coverage probabilities are usually not too conservative or too liberal. Therefore, the ELTR procedure appears to be the best procedure for most practical purposes.

The ELTR intervals are a bit more laborious to compute than the intervals based on test L_{50} or test W_{50}. In general, however, the increased precision (compared to the W_{50} intervals) and the increased accuracy (compared to the L_{50} intervals) more than outweigh the extra computational effort. The ELTR procedure has been implemented as part of the Two-Sample Variance analysis in Release 17 of Minitab Statistical Software, where it is referred to as Bonett's procedure.

For future research, one may consider investigating the small-sample properties of Layard's test in multi-sample designs when Layard's pooled kurtosis estimator is replaced with Bonett's more robust version, given as

$$\hat{\gamma}_A = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-m_i\right)^4}{\left[\sum_{i}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-m_i\right)^2\right]^2}\sum_{i=1}^{k} n_i$$

where m_i is the trimmed mean for sample i, with the trim proportion $1/[2(n_i-4)^{1/2}]$, and $i = 1, \ldots, k$.

In addition, it may be beneficial to use Shoemaker's approximation of the asymptotic variance of the log-transformed sample variance.

Finally, we note that the intervals proposed by Bonett (2006), while not suitable as CIs, are nonetheless remarkably accurate and precise for most distributions when interpreted as acceptance regions for the test of the equality of two variances. These acceptance regions are well suited to serve as the basis for a graphical procedure for comparing multiple variances.

Hochberg, Weiss, and Hart (1982) proposed a similar procedure for testing the equality of means. Such a procedure has been implemented as part of the Test for Equal Variances analysis in Release 17 of Minitab Statistical Software, where it is referred to as the Multiple Comparisons procedure.

7. Appendices

Appendix A: Proof of Result 1

Using the notations and assumptions of Section 2, let $X_j = \rho Y_{2j}$ for a given $\rho = \sigma_1/\sigma_2$. Then

$$\mathrm{Var}(X_j) = \rho^2 \mathrm{Var}(Y_{2j}) = \rho^2\sigma_2^2 = \sigma_1^2 = \mathrm{Var}(Y_{1j})$$

and

$$E\left(X_j - \mu_{X_j}\right)^4/\sigma_{X_j}^4 = \rho^4\, E(Y_{2j} - \mu_2\,)/(\rho^4\,\sigma_2^4\,) = E(Y_{2j} - \mu_2\,)/\sigma_2^4 = \gamma$$

Since $E(Y_{1j} - \mu_1)^4/\sigma_1^4 = \gamma$ by assumption, it follows that the parent populations of the two samples Y_{1j} and $X_j = \rho Y_{2j}$have the same variance σ_1^2 and the same kurtosis γ. By Layard (1973), a consistent pooled kurtosis estimator of γ based on the two samples Y_{1j} and X_j is given as

$$\hat{\gamma}\,' = (n_1 + n_2)\frac{\sum_{j=1}^{n_1}(Y_{1j} - \bar{Y}_1)^4 + \sum_{j=1}^{n_2}(X_j - \bar{X})^4}{[(n_1-1)S_1^2 + (n_2-1)S_X^2]^2} = (n_1 + n_2)\frac{\sum_{j=1}^{n_1}(Y_{1j} - \bar{Y}_1)^4 + \rho^4\sum_{j=1}^{n_2}(Y_{2j} - \bar{Y}_2)^4}{[(n_1-1)S_1^2 + \rho^2(n_2-1)S_2^2]^2}$$
$$= \hat{\gamma}_P(\rho)$$

as required.

Appendix B: Proof of Result 2

We have already established that an approximate $(1-\alpha)100$ percent confidence set for $\rho = \sigma_1/\sigma_2$ based on T_2 is given by

$$\left\{\rho \in (0,\infty)\colon (\ln\rho^2 - \ln(c\hat{\rho}^2))^2 - z_{\alpha/2}^2\left(\frac{\hat{\gamma}_P(\rho) - k_1}{n_1 - 1} + \frac{\hat{\gamma}_P(\rho) - k_2}{n_2 - 1}\right) \le 0\right\}$$

The pooled kurtosis estimator can be expressed in terms of the kurtosis estimators for the individual samples, which is given by

$$\hat{\gamma}_i = n_i\frac{\sum_{j=1}^{n_i}(Y_{ij} - m_i)^4}{\left[(n_i-1)S_i^2\right]^2}, i = 1{,}2$$

More specifically, if we let $t = \rho/\hat{\rho}$, then

$$\hat{\gamma}_P(\rho) = (n_1 + n_2)\frac{\sum_{j=1}^{n_1}(Y_{1j} - m_1)^4 + \rho^4\sum_{j=1}^{n_2}(Y_{2j} - m_2)^4}{[(n_1-1)S_1^2 + \rho^2(n_2-1)S_2^2]^2} = (n_1 + n_2)\frac{\hat{\gamma}_1\,K^2/n_1 + \hat{\gamma}_2\,t^4/n_2}{(K + t^2)^2}$$

where $K = (n_1 - 1)/(n_2 - 1)$.

Consequently, the squared standard error term can be expressed as

$$\frac{\hat{\gamma}_P(\rho) - k_1}{n_1 - 1} + \frac{\hat{\gamma}_P(\rho) - k_2}{n_2 - 1} = A\frac{\hat{\gamma}_1 K^2/n_1 + \hat{\gamma}_2 t^4/n_2}{(K + t^2)^2} - B$$

where

$$A = \frac{(n_1 + n_2)(n_1 + n_2 - 2)}{(n_1 - 1)(n_2 - 1)}, B = \frac{k_1}{n_1 - 1} + \frac{k_2}{n_2 - 1}$$

Thus, if we let $r^2 = \rho^2/(c\hat{\rho}^2)$, then it is easily seen that

$$(\ln \rho^2 - \ln(c\hat{\rho}^2))^2 - z_{\alpha/2}^2 \left(\frac{\hat{\gamma}_P(\rho) - k_1}{n_1 - 1} + \frac{\hat{\gamma}_P(\rho) - k_2}{n_2 - 1}\right)$$
$$= (\ln r^2)^2 - z_{\alpha/2}^2 \left(A\frac{\hat{\gamma}_1 K^2/n_1 + \hat{\gamma}_2 c^2 r^4/n_2}{(K + c\, r^2)^2} - B\right)$$

It follows that an approximate $(1 - \alpha)100$ percent confidence set for $\rho = \sigma_1/\sigma_2$ based on T_2 may be given as

$$\hat{\rho}\sqrt{c}\,\{r \in (0, \infty) : H(r^2) \leq 0\}$$

or equivalently, the confidence set for $\rho^2 = \sigma_1^2/\sigma_2^2$ may be expressed as

$$c\hat{\rho}^2\,\{r \in (0, \infty) : H(r) \leq 0\}$$

where

$$H(x) = (\ln x)^2 - z_{\alpha/2}^2 se^2(cx), x > 0$$

and

$$se^2(x) = A\frac{\hat{\gamma}_1 K^2/n_1 + \hat{\gamma}_2 x^2/n_2}{(K + x)^2} - B$$

Appendix C: Proof of Result 3

It is easily verified that $H(x)$ is continuous on the positive real line, with $H(0) = H(+\infty) = +\infty$ and $H(1) < 0$. By the intermediate value theorem, the function $H(x)$ admits at least one root in the interval $(0, 1)$ and at least one root in the interval $(0, +\infty)$. Thus, if the function $H(x)$ has exactly two roots, then one root is below 1 and the other is above 1. Since this function opens upward, the inequality $H(r) \leq 0$ is satisfied if r lies between the roots. These roots define the endpoints of the CI for $\rho^2/(c\hat{\rho}^2)$. Thus, if we let $x_L < 1 < x_U$ be the two roots, then, by Result 2, the lower confidence limit for the ratio of the variances, ρ^2, is calculated as $c\hat{\rho}^2 x_L$, and the lower confidence limit for the ratio of the standard deviation is obtained as $\hat{\rho}\sqrt{cx_L}$. Similarly, the upper confidence limit for the ratio of the variances is $c\hat{\rho}^2 x_U$ and the upper confidence limit for the ratio of the standard deviations is $\hat{\rho}\sqrt{cx_U}$.

On the other hand, if the function $H(x)$ has more than two roots, then the inequality $H(r) \leq 0$ is satisfied if r lies between consecutive roots where the function opens upward. Thus, the confidence set is a union of non-overlapping intervals.

8. References

Balakrishnan, N. and Ma, C. W. (1990). A Comparative Study of Various Tests for the Equality of Two Population Variances. *Journal of Statistical Computation and Simulation*, 35, 41–89.

Bonett D. G. (2006). Robust Confidence Interval for a Ratio of Standard Deviations. *Applied Psychological Measurements*, 30, 432–439.

Boos, D. D. and Brownie, C. (1989). Bootstrap Methods for Testing Homogeneity of Variances. *Technometrics*, 31, 69–82.

Brown, M. B., and Forsythe A. B. (1974). Robust Tests for the Equality of Variances. *Journal of the American Statistical Association*, 69, 364–367.

Conover, W. J., Johnson, M. E. and Johnson, M. M. (1981). A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data. *Technometrics*, 23, 351–361.

Fligner, M. A. and Killeen, T. J. (1976). Distribution-Free Two-Sample Tests for Scale. *Journal of the American Statistical Association*, 71, 210–213.

Hall, P. and Padmanabhan, A. R. (1997). Adaptive Inference for the Two-Sample Scale Problem. *Technometrics*, 39, 412–422.

Hochberg, Y., Weiss, G., and Hart S., (1982). On Graphical Procedures for Multiple Comparisons. *Journal of the American Statistical Association*, 77, 767–772.

Layard, M. W. J. (1973). Robust Large-Sample Tests for Homogeneity of Variances. *Journal of the American Statistical Association*, 68, 195–198.

Levene, H. (1960). "Robust Tests for Equality of Variances," in I. Olkin, ed., *Contributions to Probability and Statistics*, Palo Alto, CA: Stanford University Press, 278–292.

Lim, T.-S. and Loh, W.-Y. (1996). A Comparison of Tests of Equality of Variances. *Computational Statistics and Data Analysis*, 22, 287–301.

Ott, L. (1993). *An Introduction to Statistical Methods and Data Analysis*, Belmont, CA: Duxbury Press.

Pan, G. (1999). On a Levene Type Test for Equality of Two Variances. *Journal of Statistical Computation and Simulation*, 63, 59–71.

Shoemaker, L. H. (2003). Fixing the F Test for Equal Variances. *The American Statistician*, 57, 105–114.

Wolfram, S. (1999). *The Mathematica Book*, 4th ed. Wolfram Media/Cambridge University Press.