Multiple Comparisons Method

A GRAPHICAL MULTIPLE COMPARISONS PROCEDURE FOR SEVERAL STANDARD DEVIATIONS

Senin J. Banga and Gregory D. Fox

June 18, 2013

ABSTRACT

A new graphical procedure for multiple comparisons of k standard deviations is provided. As a test for homogeneity of variances, the new procedure has similar type I and type II error properties as the Brown and Forsythe (1974) version of the Levene (1960) test, W_{50}. The graphical display associated with the multiple comparisons test, however, provides a useful visual tool for screening samples with different standard deviations.

Index terms: Homogeneity of variances, Levene's test, Brown-Forsythe test, Layard's test, multiple comparisons

1. Introduction

The Brown and Forsythe (1974) modification of Levene's test (1960), commonly referred to as test W_{50}, is perhaps one of the most widely used procedures for testing the homogeneity (equality) of variances. In part, test W_{50} is popular because it is robust and is asymptotically distribution free. Compared to other tests of the homogeneity of variances, test W_{50} is also easy to calculate. (For a comparison of such tests, see Conover et al. (1981).) In addition, test W_{50} is easily accessible because it is available in many statistical software packages such as SAS, Minitab, R, and JMP.

However, for some distributions, the power of test W_{50} can be very low, particularly in small samples. For example, Pan (1999) shows that for some distributions, including the normal distribution, test W_{50} may not have sufficient power to detect differences between two standard deviations, regardless of the magnitude of the differences. It is not clear from Pan's analysis whether the same limitation would apply to multi-sample designs. One might expect that this limitation would not apply to designs with more than two samples, simply because such designs are likely to include more data than two-sample designs. Test W_{50} is known to have good large-sample properties (Miller, 1968; Brown and Forsythe, 1974; Conover et al., 1981).

It has become common practice to follow a significant test W_{50} with a simultaneous pairwise comparison procedure based on a Bonferroni multiplicity correction. As pointed out by Pan (1999), however, such an approach is likely to fail or to yield misleading results because of the low power of test W_{50} in two-sample designs. Using the Bonferroni correction worsens the problem because it is conservative, particularly when the number of pairwise comparisons is large. In contrast, many effective multiple comparison procedures are available for comparing means following a one-way ANOVA. For examples, see Tukey (1953), Hochberg et al. (1982), and Stoline (1981). An analogous post-hoc analysis for comparisons among sample variances would be useful.

In this paper, we propose a graphical method for comparing the variances (or standard deviations) of multiple samples. The analysis is based on "uncertainty intervals" for variances that are similar to the uncertainty intervals described by Hochberg et al. (1982) for means. First, a multiple pairwise comparisons procedure is based on the Bonett's (2006) modified version of Layard's (1973) test for the equality of variances for two-sample designs. The multiplicity correction used in the pairwise comparisons is based on a large-sample generalization of the Tukey-Kramer method (Tukey, 1953; Kramer, 1956), proposed by Nakayama (2009). The uncertainty intervals, which we refer to as "multiple comparison intervals" or "MC intervals", are derived from the pairwise comparison procedure using the best approximate procedure described by Hochberg et al. (1982). The resulting MC test rejects the null hypothesis if, and only if, at least one pair of MC intervals does not overlap. Non-overlapping MC intervals identify the samples that have significantly different variances (or standard deviations).

We perform simulation studies to assess the small-sample properties of the MC test. For comparison, we also include test W_{50} in the simulation studies.

2. Graphical Multiple Comparisons Procedure

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 $E(Y_{il}) = \mu_i$ and variance $\mathrm{Var}(Y_{il}) = \sigma_i^2 > 0$. In addition, suppose that the samples originate from populations with a common kurtosis $\gamma = E(Y-\mu)^4/\sigma^4 < \infty$.

Also, let $\bar{Y}_i$ and S_i be the mean and the standard deviation of sample i, respectively. Let m_i be the trimmed mean of sample i with trim proportion $1/[2\sqrt{n_i - 4}]$ and let $\hat{\gamma}_{ij}$ be a pooled kurtosis estimator of samples (i, j) given as

$$\hat{\gamma}_{\mathrm{ij}} = \left(\mathrm{n_i} + \mathrm{n_j}\right) \frac{\sum_{\mathrm{l=1}}^{\mathrm{n_i}} (\mathrm{Y_{il}} - \mathrm{m_i})^4 + \sum_{\mathrm{l=1}}^{\mathrm{n_j}} \left(\mathrm{Y_{jl}} - \mathrm{m_j}\right)^4}{\left[\sum_{\mathrm{l=1}}^{\mathrm{n_i}} (\mathrm{Y_{il}} - \bar{\mathrm{Y}}_{\mathrm{i}})^2 + \sum_{\mathrm{l=1}}^{\mathrm{n_j}} \left(\mathrm{Y_{jl}} - \bar{\mathrm{Y}}_{\mathrm{j}}\right)^2\right]^2}$$

$$= \left(\mathrm{n_i} + \mathrm{n_j}\right) \frac{\sum_{\mathrm{l=1}}^{\mathrm{n_i}} (\mathrm{Y_{il}} - \mathrm{m_i})^4 + \sum_{\mathrm{l=1}}^{\mathrm{n_j}} \left(\mathrm{Y_{jl}} - \mathrm{m_j}\right)^4}{\left[(\mathrm{n_i} - 1)\mathrm{S_i^2} + (\mathrm{n_j} - 1)\mathrm{S_j^2}\right]^2}$$

Note that $\hat{\gamma}_{ij}$ is asymptotically equivalent to Layard's (1973) pooled kurtosis estimator where the sample mean $\bar{Y}_i$ has been replaced with the trimmed mean m_i. Thus, $\hat{\gamma}_{ij}$ is a consistent estimator of the unknown common kurtosis γ, so long as the population variances are equal. Bonett (2006) proposes this estimator in place of Layard's pooled kurtosis estimator to improve the small-sample performance of Layard's test in two-sample problems. Throughout, we refer to Bonett's (2006) modified version of Layard's test simply as Bonett's test.

Suppose that there are more than two independent groups or samples to compare ($k > 2$). The graphical multiple comparison procedure we propose is derived from the multiple pairwise comparisons that are based on the Bonett's test. An alternative approach is to base the pairwise comparisons on test W_{50}. In two-sample designs, however, the power performance of test W_{50} is problematic for some distributions including the normal distribution (Pan, 1999). Moreover, Banga and Fox (2013) show that confidence intervals for the ratio of variances that are based on Bonett's test are generally superior to those based on test W_{50}.

Given any pair (i, j) of samples, a two-sided Bonett's test with significance level α' rejects the null hypothesis of the equality of variances if, and only if,

$$\left|\ln(c_i S_i^2) - \ln(c_j S_j^2)\right| > z_{\alpha'/2} \sqrt{\frac{\hat{\gamma}_{ij} - k_i}{n_i - 1} + \frac{\hat{\gamma}_{ij} - k_j}{n_j - 1}}$$

where $z_{\alpha'/2}$ is the $\alpha'/2 \times 100^{\text{th}}$ upper percentile point of the standard normal distribution,

$$k_i = \frac{n_i - 3}{n_i}, k_j = \frac{n_j - 3}{n_j}, c_i = \frac{n_i}{n_i - z_{\alpha/2}}, c_j = \frac{n_j}{n_j - z_{\alpha/2}}$$

Since there are multiple pairwise comparisons, exactly $k(k-1)/2$ comparisons, a multiplicity adjustment is required. For example, if a target overall or family-wise significance level, α, is given, then one common approach, known as the Bonferroni correction, is to choose the significance level of each of the $k(k-1)/2$ pairwise comparisons, $\alpha' = 2\alpha/(k(k-1))$.The Bonferroni correction, however, is well known to yield increasingly conservative pairwise comparison procedures as the number of samples to compare increases. An alternative and better approach is proposed by Nakayama (2009) and is based on a large-sample approximation of the Tukey-Kramer method (Tukey, 1953; Kramer, 1956). Specifically, the overall multiple pairwise comparisons test is significant if, and only if, the following is true for some pair (i, j) of samples:

$$\left|\ln(c_i S_i^2) - \ln(c_j S_j^2)\right| > \frac{q_{k,\alpha}}{\sqrt{2}} \sqrt{\frac{\hat{\gamma}_{ij} - k_i}{n_i - 1} + \frac{\hat{\gamma}_{ij} - k_j}{n_j - 1}}$$

where $q_{\alpha,k}$ is the upper α point of the range of k independent and identically distributed standard normal random variables. That is, $q_{\alpha,k}$ satisfies

$$\Pr\left(\max_{1 \le i < j \le k} \left|Z_i - Z_j\right| \le q_{\alpha,k}\right) = 1 - \alpha$$

where $Z_1, \ldots, Z_k$ are independent and identically distributed standard normal random variables. Barnard (1978) provides a simple numerical algorithm based on a 16-point Gaussian quadrature for computing the distribution function of the normal range.

As suggested by Hochberg et al. (1982), a graphical multiple comparisons procedure that approximates the multiple pairwise comparison procedure described above would reject the null hypothesis if, and only if,

$$\left|\ln(c_i S_i^2) - \ln(c_j S_j^2)\right| > q_{\alpha,k}(V_i + V_j)/\sqrt{2}$$

where the V_i are selected to minimize the following:

$$\sum_{i \ne j} \sum \left(V_i + V_j - b_{ij}\right)^2$$

where

$$b_{ij} = \sqrt{\frac{\hat{\gamma}_{ij} - k_i}{n_i - 1} + \frac{\hat{\gamma}_{ij} - k_j}{n_j - 1}}$$

The solution to this problem, as illustrated in Hochberg et al. (1982), is to choose

$$V_i = \frac{(k-1)\sum_{j \ne i} b_{ij} - \sum\sum_{1 \le j < l \le k} b_{jl}}{(k-1)(k-2)}$$

It follows that a test of homogeneity of variances based on this approximate procedure rejects the null hypothesis if, and only if, at least one pair of the intervals given below do not overlap:

$$\left[S_i \sqrt{c_\mathrm{i} \exp(-q_{\alpha,k} V_i/\sqrt{2}\,)}\,, S_i \sqrt{c_i \exp(q_{\alpha,k} V_i/\sqrt{2})}\right], i = 1, \ldots, k$$

The graphical MC procedure consists of displaying these intervals on a graph to visually identify the samples with non-overlapping intervals. In addition, the p-value of the overall test of the homogeneity of variance (or standard deviation) can be determined. We provide detailed algorithms for calculating the p-value in the next section. But, first, we point out some simple facts about the MC procedure.

REMARK

1. The pooled kurtosis estimator, $\hat{\gamma}_{ij}$, based on the pair (i, j) of samples, could have been replaced with the overall pooled kurtosis estimator, based on all the k samples. Although this approach somewhat simplifies the computations, simulation results that are not shown here, indicate that using $\hat{\gamma}_{ij}$ yields better results.
2. The interval corresponding to sample i is not a confidence interval for the standard deviation of the sample parent population. Hochberg et al. (1982) refer to such an interval as an “uncertainty interval”. We refer to it as a “multiple comparison interval” or an “MC interval”. MC intervals are useful only for comparing the standard deviations or variances for multi-sample designs.
3. The MC intervals that are described in this paper can be used only to compare more than two standard deviations. When there are only two samples, comparison intervals can be constructed, but they convey the same information that is provided by the test results. It is much more informative to construct a confidence interval for the ratio of the standard deviations, such as that described by Banga and Fox (2013) and provided with Minitab's Two-Sample Variance command.

3. P-value of the Graphical Multiple Comparisons Method

Before we describe the algorithm for calculating the p-value of the graphical (MC) method, we first derive the p-value associated with Bonett’s (2006) modification of Layard’s test in two-sample designs. We then show how to apply the two-sample design results to the multiple comparisons procedure.

3.1 P-value in two-sample designs

As mentioned earlier, Bonett’s (2006) adjustment of Layard’s test in two-sample designs rejects the null hypothesis of homogeneity of variances if, and only if,

$$\left|\ln(c_1 S_1^2) - \ln(c_2 S_2^2)\right| > z_{\alpha/2} se$$

or equivalently

$$\left|\ln(c_{\alpha/2}\, S_1^2\, / S_2^2)\right| > z_{\alpha/2} se$$

where

$$se = \sqrt{\frac{\hat{\gamma}_{12} - k_1}{n_1 - 1} + \frac{\hat{\gamma}_{12} - k_2}{n_2 - 1}}$$

$$c_{\alpha/2} = \frac{c_1}{c_2} = \frac{n_1}{n_1 - z_{\alpha/2}} \frac{n_2 - z_{\alpha/2}}{n_2}$$

Bonett introduced the constant $c_{\alpha/2}$ as a small-sample adjustment to mitigate the effect of unequal tail-error probabilities in small-sample unbalanced designs. The effect of the constant, however, is negligible in large-sample unbalanced designs, and the constant has no effect in balanced designs.

It follows that, if the design is balanced, then the p-value of the two-sided test for the homogeneity of variances is simply calculated as

$$P = 2\,\Pr(Z > |Z_0\,|)$$

where

$$Z_0 = \frac{\ln(S_1^2) - \ln(S_2^2)}{se}$$

If the design is unbalanced, then $P = 2\min(\alpha_L, \alpha_U)$, where α_L is the smallest solution for α in the equation,

$\exp[\ln(c_\alpha S_1^2/S_2^2) - z_\alpha se] = 1$ (1)

and α_U is the smallest solution for α in the equation,

$\exp[\ln(c_\alpha S_1^2/S_2^2) + z_\alpha se] = 1$ (2)

Algorithms for finding α_L and α_U are given below. The mathematical details of the algorithms are deferred to the Appendix section.

Let

$$L(z, n_1, n_2, S_1, S_2) = \ln\frac{n_1}{n_2} + \ln\frac{n_2 - z}{n_1 - z} - z\,se + \ln\frac{S_1^2}{S_2^2}, z < \min(n_1, n_2)$$

Also let

$$z_m = \frac{n_1 + n_2 - \sqrt{(n_1 - n_2)(n_1 - n_2 - \frac{4}{se})}}{2}$$

The solutions α_L and α_U are calculated in the following steps:

Case 1: $n_1 < n_2$

- Calculate z_m as given in the above result and evaluate $L(z_m, n_1, n_2, S_1, S_2)$.
- If $L(z_m) \leq 0$, then find the root, z_L, of $L(z, n_1, n_2, S_1, S_2)$ in the interval, $(-\infty, z_m]$ and calculate $\alpha_L = \Pr(Z > z_L)$.

- If $L(z_m) > 0$, then the function $L(z, n_1, n_2, S_1, S_2)$ has no root. Set $\alpha_L = 0.0$.

Case 2: $n_1 > n_2$

- Calculate $L(0, n_1, n_2, S_1, S_2) = \ln S_1^2 / S_2^2$.
- If $L(0, n_1, n_2, S_1, S_2) \geq 0$ then find the root, z_o, of $L(z, n_1, n_2, S_1, S_2)$ in the interval $[0, n_2)$, otherwise find the root z_L in the interval $(-\infty, 0)$.
- Calculate $\alpha_L = \Pr(Z > z_L)$.

To calculate α_U, we simply apply the above steps using the function, $L(z, n_2, n_1, S_2, S_1)$, instead of the function, $L(z, n_1, n_2, S_1, S_2)$.

3.2 P-value of the Graphical Multiple comparisons

Assuming that there are k $(k > 2)$ samples in the design, we let P_{ij} be the p-value of the test associated with any pair (i, j) of samples. We recall that the multiple comparisons test rejects the null hypothesis of the homogeneity of variances if, and only if, at least one pair of the k comparison intervals does not overlap. It follows that the overall p-value associated with the multiple comparisons procedure is

$$P = \min\{P_{ij}, 1 \leq i < j \leq k\}$$

To calculate P_{ij} we perform the algorithm of the two-sample designs using

$$se = V_i + V_j$$

where V_i is as defined as before.

If $n_i \neq n_j$, then

$$P_{ij} = \min(\alpha_L, \alpha_U)$$

where $\alpha_L = \Pr(Q > z_L\sqrt{2})$, $\alpha_U = \Pr(Q > z_U\sqrt{2})$, z_L is the smallest root of the function, $L(z, n_i, n_j, S_i, S_j)$, z_U is the smallest root of the function, $L(z, n_j, n_i, S_j, S_i)$, and Q is a random variable as defined previously. The quantities z_L and z_U are found by applying the two-sample design algorithm described earlier to the pair (i, j) of samples.

If $n_i = n_j$ then $P_{ij} = \Pr(Q > |z_o|\sqrt{2})$, where

$$z_o = \frac{\ln S_i^2 - \ln S_j^2}{V_i + V_j}$$

4. Simulation Study and Results

Two major simulation studies are conducted to investigate the small-sample performance of the MC test as an overall test for the homogeneity of variances. All simulations were conducted using Version 8 of the Mathematica software package.

Study 1

The first study is designed to evaluate and compare the type I error properties of the MC test and test W_{50}. We compare the performance of the two tests with samples generated from various distributions in three different designs: a 3-sample design, a 4-sample design, and a 6-sample design. In each design, the sample sizes are varied from 10 to 50 increments of 10. Samples are drawn from the following parent distributions:

- the normal distribution
- symmetric light-tailed distributions, represented by the uniform distribution and a Beta distribution with parameters of (3, 3)
- symmetric heavy-tailed distributions, represented by a t-distribution with 5 degrees of freedom ($t(5)$) and the Laplace distribution
- skewed and heavy-tailed distributions, represented by the exponential distribution, a chi-square distribution with 1 degree of freedom ($\chi^2(1)$), and a chi-square distribution with 5 degrees of freedom ($\chi^2(5)$)
- a contaminated normal distribution (CN(0.9, 3)) for which 90% of observations are drawn from the standard normal distribution and the remaining 10% are drawn from a normal distribution with a mean of 0 and a standard deviation of 3.

Each simulation consists of 10,000 sampling replicates. The targeted nominal α level is 0.05. The simulation error is approximately 0.002. The simulated significance levels for each test are reported in Table 1.

Table 1 Comparison of Simulated Significance Levels ($\alpha = 0.05$)

Description	Distribution [Kurtosis]	n_i	$k = 3$		$k = 4$		$k = 6$	
			MC	W_{50}	MC	W_{50}	MC	W_{50}
Normal	Normal [3.0]	10	.038	.033	.038	.031	.036	.029
		20	.039	.038	.040	.038	.041	.033
		30	.043	.041	.044	.038	.046	.039
		40	.046	.043	.046	.041	.048	.041
		50	.046	.046	.046	.044	.052	.047
Symmetric with Light Tails	Uniform [1.8]	10	.029	.029	.025	.024	.023	.020
		20	.028	.026	.030	.026	.028	.023
		30	.037	.035	.034	.032	.034	.030
		40	.038	.037	.037	.037	.035	.033
		50	.041	.041	.036	.036	.036	.036

Description	Distribution [Kurtosis]	n_i	$k = 3$		$k = 4$		$k = 6$	
			MC	W_{50}	MC	W_{50}	MC	W_{50}
	Beta(3, 3) [2.5]	10	.031	.032	.031	.029	.031	.025
		20	.035	.031	.036	.027	.037	.026
		30	.041	.035	.037	.034	.037	.032
		40	.040	.036	.039	.035	.040	.033
		50	.044	.039	.044	.037	.044	.035
Symmetric with Heavy Tails	Laplace [6.0]	10	.056	.038	.063	.041	.071	.039
		20	.054	.044	.058	.043	.059	.041
		30	.051	.042	.053	.043	.052	.044
		40	.048	.045	.048	.045	.048	.046
		50	.045	.045	.051	.046	.049	.047
	$t(5)$ [9.0]	10	.042	.032	.044	.031	.042	.031
		20	.043	.039	.045	.038	.045	.040
		30	.039	.040	.040	.040	.041	.040
		40	.041	.042	.040	.041	.039	.038
		50	.040	.050	.039	.046	.038	.046
Skewed with Heavy Tails	$\chi^2(5)$ [5.4]	10	.040	.039	.046	.040	.048	.039
		20	.040	.043	.040	.040	.042	.039
		30	.039	.047	.042	.044	.043	.042
		40	.040	.046	.041	.044	.039	.042
		50	.037	.047	.038	.047	.040	.048
	Exponential [9.0]	10	.063	.051	.073	.049	.076	.048
		20	.051	.049	.053	.048	.057	.046
		30	.042	.048	.046	.051	.049	.049
		40	.034	.050	.038	.046	.037	.049
		50	.033	.045	.037	.047	.038	.046
	$\chi^2(1)$ [15.0]	10	.084	.048	.098	.050	.118	.050
		20	.053	.046	.060	.047	.068	.046

Description	Distribution [Kurtosis]	n_i	$k = 3$		$k = 4$		$k = 6$	
			MC	W_{50}	MC	W_{50}	MC	W_{50}
		30	.041	.041	.045	.045	.050	.047
		40	.044	.049	.046	.047	.045	.047
		50	.038	.050	.037	.049	.040	.049
Contaminated Normal	CN(0.9, 3) [8.3]	10	.020	.016	.018	.012	.016	.010
		20	.014	.015	.012	.013	.008	.007
		30	.012	.014	.010	.011	.007	.008
		40	.009	.017	.009	.014	.006	.008
		50	.009	.016	.007	.012	.006	.009

The results show that both tests perform well for most distributions. Most of the simulated significance levels are close to the target of 0.05. However, the simulated significance levels for both tests tend to be conservative (lower than 0.05) when small samples are drawn from normal and symmetric light-tailed distributions. For these distributions, the simulated significance levels for the MC test are closer to the target significance level than are those for test W_{50}.

When small samples are drawn from heavy-tailed distributions, test W_{50} tends to be conservative and the MC test tends to be liberal. The MC test is even more liberal when small samples are drawn from extremely skewed distributions. For example, when samples of size 10 are drawn from a chi-square distribution with 1 degree of freedom, the simulated significance levels for the MC test are 0.084, 0.098, and 0.118 for the 3-, 4-, and 6-sample designs, respectively.

Both tests are influenced by outliers. Significance levels for the contaminated normal distribution are extremely conservative even when sample sizes are as large as 50.

Study 2

The second study evaluates and compares the type II error properties (power) of the two procedures in a 4-sample design. We use the same samples for this study as those used for the samples of size 20 and $k = 4$ condition in Study 1. The observations are scaled by a factor of 1, 2, 3, or 4. For example, in the condition denoted as 1:1:4:4, the observations for samples 1 and 2 are the same as those used in Study 1. The observations in samples 3 and 4 are scaled by a factor of 4.

We include the condition 1:1:1:1 for comparison. Notice that the results for this condition are the same as those reported in Study 1 for samples of size 20 and $k = 4$. We choose samples of size 20 because the results of Study 1 suggest that, for both tests, samples of size 20 yield achieved significance levels that are near the target level for most distributions.

The simulated power levels in these experiments are calculated as the proportion of sample replicates that lead to rejections of the null hypothesis of homogeneity of variances.

The results are reported in Table 2.

Table 2 Comparison of Simulated Power Levels ($\alpha = 0.05$)

Description	Distribution	Standard deviation ratio							
		1:1:1:1		1:1:2:2		1:2:3:4		1:1:4:4	
		MC	W_{50}	MC	W_{50}	MC	W_{50}	MC	W_{50}
	Normal	.040	.038	.846	.853	.998	.994	1.000	1.000
Symmetric Light Tailed	Uniform	.030	.026	.985	.962	1.000	.999	1.000	1.000
	Beta(3, 3)	.036	.027	.938	.916	1.000	.999	1.000	1.000
Symmetric Heavy Tailed	Laplace	.058	.043	.597	.629	.931	.921	.996	.998
	$t(5)$	.045	.038	.657	.703	.952	.949	.997	.998
Skewed Heavy Tailed	$\chi^2(5)$	.040	.040	.625	.704	.949	.949	.996	.999
	Exponential	.053	.048	.431	.507	.804	.779	.963	.978
	$\chi^2(1)$	.060	.047	.298	.291	.602	.504	.838	.824
Contaminated	CN(0.9, 3)	.012	.013	.499	.612	.889	.917	.989	.998

The results suggest that the type II error properties (power) of the MC test and test W_{50} are similar. In general, the simulated power levels achieved with both tests are of the same order of magnitude. In only one case does the power for the two tests differ by more than 0.1.

The simulated power levels for the MC test are slightly better than those of test W_{50} when samples are drawn from symmetric distributions with light to moderated tails. On the other hand, test W_{50} appears to be slightly more powerful than the MC test when samples are drawn from distributions with heavy tails.

5. Example

In this section, we apply the graphical MC procedure and test W_{50} to a data set obtained from Ott et al. (2010), page 397. The data is described as follows:

> *A casting company has several ovens in which they heat the raw materials prior to pouring them into a wax mold. It is very important that these metals be heated to a precise temperature with very little variation. Three ovens are selected at random and their temperatures are recorded (℃) very accurately on 10 successive heats. The collected data are as follows:*

Oven 1	1670.87	1670.88	1671.51	1672.01	1669.63	1670.95	1668.70	1671.86	1669.12	1672.52
Oven 2	1669.16	1669.60	1669.76	1669.18	1671.92	1669.69	1669.45	1669.35	1671.89	1673.45
Oven 3	1673.08	1672.75	1675.14	1674.94	1671.33	1660.38	1679.94	1660.51	1668.78	1664.32

Figure 1 shows boxplots of the temperatures for each oven. The boxplots suggest that there are no outliers in the recorded temperatures and that the temperature variability for Oven 3 is different from that of Oven 1 or Oven 2.

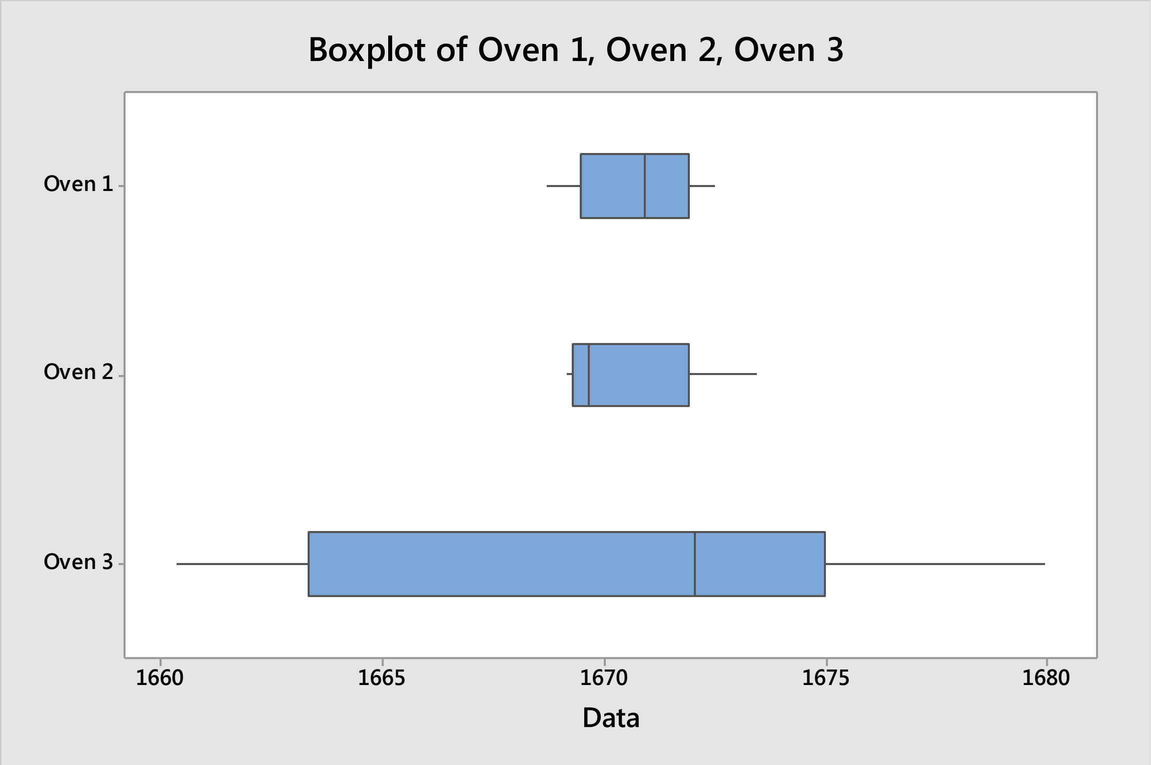

Figure 1 Boxplots of oven temperature (℃)

Figure 2 shows the MC intervals for the same data, as well as the results of the overall MC test and test W_{50}, which is referred to in the legend as Levene's Test. The significant p-values for both tests indicate that the variability in temperatures is different among the three ovens. The non-overlapping MC intervals confirm that the variability for Oven 3 is different from Oven 2 or Oven 1. The MC intervals are $(0.896, 2.378)$, $(1.072, 2.760)$, and $(4.366, 12.787)$ for Ovens 1, 2, and 3, respectively.

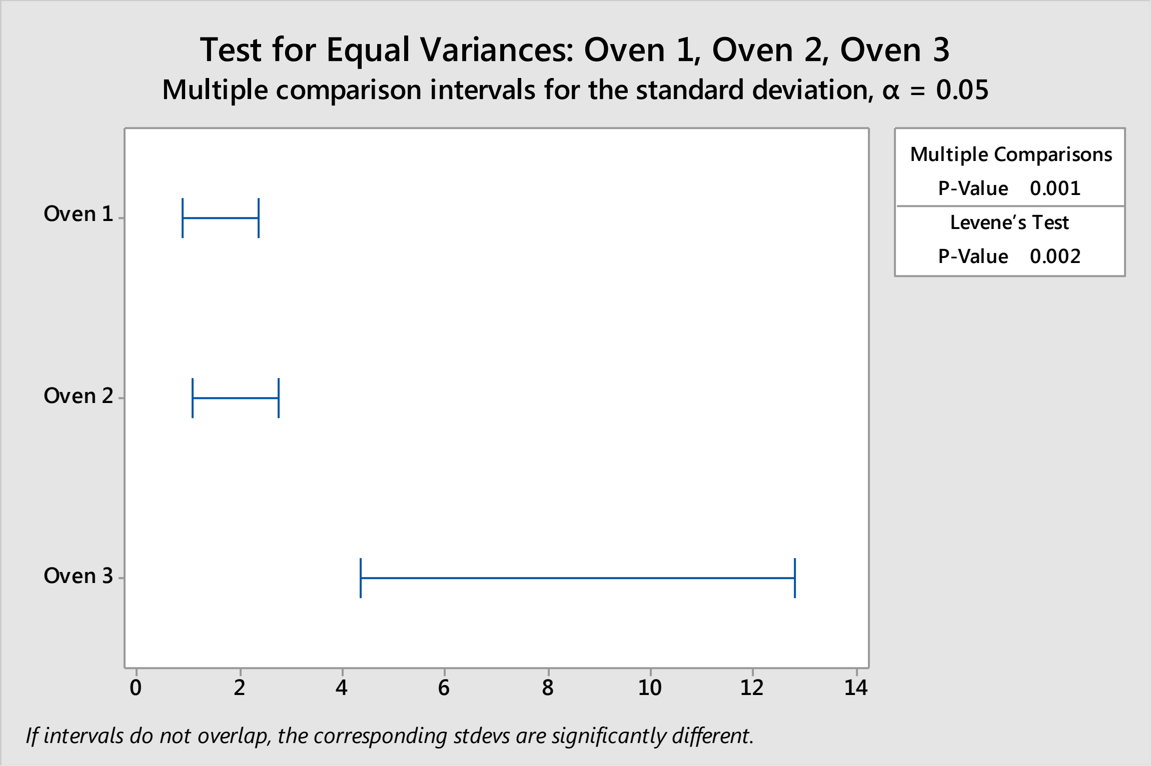

Figure 2 MC intervals and p-values for the MC test and test W_{50} (Levene's Test)

6. Conclusion

Overall, the simulation results show that, for designs with multiple small samples, the performance of the MC test is similar to that of test W_{50}. The MC test is slightly more suited to symmetric or nearly symmetric distributions with light to moderate tails, while test W_{50} might be preferred when data are drawn from highly skewed distributions and distributions with heavy tails. One clear advantage of the MC procedure is that it provides a powerful visual tool for screening samples with different standard deviations or variances when the overall test for the homogeneity of standard deviations is significant. The graphical MC procedure is available in Minitab, release 17.

7. Appendix

Bonett's (2006) adjustment of Layard's test in two-sample designs rejects the null hypothesis of homogeneity of variances if, and only if,

$$\left|\ln(c_1 S_1^2) - \ln(c_2 S_2^2)\right| > z_{\alpha/2} se$$

or equivalently

$$\left|\ln(c_{\alpha/2}\, S_1^2\, / S_2^2)\right| > z_{\alpha/2} se$$

where

$$se = \sqrt{\frac{\hat{\gamma}_{12} - k_1}{n_1 - 1} + \frac{\hat{\gamma}_{12} - k_2}{n_2 - 1}}$$

$$c_{\alpha/2} = \frac{c_1}{c_2} = \frac{n_1}{n_1 - z_{\alpha/2}} \frac{n_2 - z_{\alpha/2}}{n_2}$$

Thus, if the design is balanced, then $c_{\alpha/2} = 1$, so that the p-value of the test is simply

$$P = 2\,\Pr(Z > |Z_0\,|)$$

where

$$Z_0 = \frac{\ln(S_1^2) - \ln(S_2^2)}{se}$$

If the design is unbalanced, then $P = 2\min(\alpha_L, \alpha_U)$ where

$\alpha_L$is the smallest solution for α in the equation

$\exp[\ln(c_\alpha S_1^2/S_2^2) - z_\alpha se] = 1$ (1)

and α_U is the smallest solution α of the equation

$\exp[\ln(c_\alpha S_1^2/S_2^2) + z_\alpha se] = 1$ (2)

The approach for solving these equations for α is to first solve the equations for $z \equiv z_\alpha$ and then obtain $\alpha = \Pr(Z > z)$ where the random variable Z has the standard normal distribution. Before we describe how to solve these equations, we note that equation (1) can be re-expressed as the equation $L(z) = 0$ where

$$L(z, n_1, n_2, S_1, S_2) = \ln\frac{n_1}{n_2} + \ln\frac{n_2 - z}{n_1 - z} - z\, se + \ln\frac{S_1^2}{S_2^2}, z < \min(n_1, n_2)$$

Similarly, equation (2) is equivalent to the equation $U(z) = 0$, where

$$U(z, n_1, n_2, S_1, S_2) = \ln\frac{n_1}{n_2} + \ln\frac{n_2 - z}{n_1 - z} + z\, se + \ln\frac{S_1^2}{S_2^2}, z < \min(n_1, n_2)$$

We note that $L(z, n_2, n_1, S_2, S_1) = -U(z, n_1, n_2, S_1, S_2)$. Consequently, the roots of only one of the two functions must be found.

The algorithm for solving equation (1), or (2), is derived from the following result:

Result

Let n_1, n_2, S_1 and S_2 be given and fixed. For unbalanced designs, the function, $L(z, n_1, n_2, S_1, S_2)$, has, at most, two roots.

4. *If $n_1 < n_2$ then $L(z, n_1, n_2, S_1, S_2)$ is convex: It satisfies $L(-\infty, n_1, n_2, S_1, S_2) = L(n_1, n_1, n_2, S_1, S_2) = +\infty$, and reaches its minimum at*

$$z_m = \frac{n_1 + n_2 - \sqrt{(n_1 - n_2)(n_1 - n_2 - \frac{4}{se})}}{2}$$

   Thus, if $L(z_m, n_1, n_2, S_1, S_2) \leq 0$, then there are two roots: One in the interval $(-\infty, z_m]$ and the other in the interval $[z_m, n_1)$. On the other hand, if $L(z_m, n_1, n_2, S_1, S_2) > 0$, then the function $L(z, n_1, n_2, S_1, S_2)$ has no root.

5. If $n_1 > n_2$, then $L(z, n_1, n_2, S_1, S_2)$ decreases monotonically from $+\infty$ to $-\infty$ and therefore has a unique root. If $L(0, n_1, n_2, S_1, S_2) = ln\, S_1^2/S_2^2 \geq 0$, then the root is in the interval $[0, n_2)$; otherwise it lies in the interval $(-\infty, 0)$.

Proof

In the following, we let $L(z) \equiv L(z, n_1, n_2, S_1, S_2)$.

First, we want to prove that, if $n_1 < n_2$ then $L(z)$ is convex and reaches its minimum at

$$z_m = \frac{n_1 + n_2 - \sqrt{(n_1 - n_2)(n_1 - n_2 - \frac{4}{se})}}{2}$$

As previously defined

$$L(z) = \ln\frac{n_1}{n_2} + \ln\frac{n_2 - z}{n_1 - z} - z\, se + \ln\frac{S_1^2}{S_2^2}, z < \min(n_1, n_2)$$

Then, we have $\lim_{z \to -\infty} L(z) = +\infty$ and

$$\lim_{z \to \min(n_1, n_2)} L(z) = \begin{cases} +\infty \text{ if } n_1 < n_2 \\ -\infty \text{ if } n_2 < n_1 \end{cases}$$

Also, note that the derivative of $L(z)$ satisfies

$$-\frac{(n_1 - z)(n_2 - z)}{se} L'(z) = z^2 - (n_1 + n_2)z + n_1 n_2 + \frac{n_1 - n_2}{se}$$

Let

$$Q(z) = -\frac{(n_1 - z)(n_2 - z)}{se} L'(z)$$

If $n_1 < n_2$, then quadratic $Q(z)$ has two roots given as

$$z_1 = \frac{n_1 + n_2 - \sqrt{(n_1 - n_2)(n_1 - n_2 - \frac{4}{se})}}{2}$$

and

$$z_2 = \frac{n_1 + n_2 + \sqrt{(n_1 - n_2)(n_1 - n_2 - \frac{4}{se})}}{2}$$

Since $Q(n_1) = \frac{n_1 - n_2}{se} < 0$, we have $z_1 < n_1 = \min(n_1, n_2) < z_2$ so that $Q(z) > 0$ for z in $(-\infty, z_1)$ and so that $Q(z) < 0$ for z in (z_1, n_1). It follows that $L'(z) < 0$ for z in $(-\infty, z_1)$ and that $L'(z) > 0$ for z in (z_1, n_1). Thus $L(z)$ is convex on the domain $(-\infty, \min(n_1, n_2))$ and reaches its minimum value at $z_1 \equiv z_m$.

If $n_1 > n_2$, then there are two cases: the case where $n_1 - n_2 > 4/se$ and the case where $0 < n_1 - n_2 < 4/se$. In the first case, z_1 and z_2 are the roots of $Q(z)$ such that $n_2 = \min(n_1, n_2) < z_1 < z_2$. (This is because $n_2 - \frac{z_1 + z_2}{2} = \frac{n_2 - n_1}{2} < 0$). Thus $Q(z) > 0$ for z in the domain $(-\infty, \min(n_1, n_2))$. In the second case, $Q(z)$ has no roots so that $Q(z) > 0$ on the domain.

It follows that if $n_1 > n_2$, then $L'(z) < 0$ so that $L(z)$ decreases monotonically from $+\infty$ to $-\infty$.

8. References

Banga, S. J. and Fox, G. D. (2013). On Bonett's Robust Confidence Interval for a Ratio of Standard Deviations. In press.

Barnard, J. (1978). Probability Integral of the Normal Range. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 27, 197–198.

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

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.

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

Kramer, C. Y. (1956). Extension of Multiple Range Tests to Group Means with Unequal Numbers of Replications. *Biometrics*, 12, 307–310.

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.

Miller, R. G. (1968). Jackknifing Variances. *Annals of Mathematical Statistics*, 39, 567–582.

Nakayama, M. K. (2009). Asymptotically Valid Single-Stage Multiple-Comparison Procedures. *Journal of Statistical Planning and Inference*, 139, 1348–1356.

Ott, R. L. and Longnecker, M. (2010). *An introduction to Statistical Methods and Data Analysis, sixth edition*, Brooks/Cole, Cengage Learning.

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

Stoline, M. R. (1981). The Status of Multiple of Comparisons: Simultaneous Estimation of All Pairwise Comparisons in One-Way ANOVA Designs. *The American Statistician*, 35, 134–141.

Tukey, J. W. (1953). *The Problem of Multiple Comparisons*. Mimeographed monograph.

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