

Tests for Standard Deviations (Two or More Samples)

Overview

The Minitab Assistant includes two analyses to compare independent samples to determine whether their variability significantly differs. The 2-Sample Standard Deviation test compares the standard deviations of 2 samples, and the Standard Deviations test compares the standard deviations of more than 2 samples. In this paper, we refer to k -sample designs with $k = 2$ as 2-sample designs and k -sample designs with $k > 2$ as multiple-sample designs. Generally, these two types of designs are studied separately (see Appendix A).

Because the standard deviation is the square root of the variance, a hypothesis test that compares standard deviations is equivalent to a hypothesis test that compares variances. Many statistical methods have been developed to compare the variances from two or more populations. Among these tests, the Levene/Brown-Forsythe test is one of the most robust and most commonly used. However, the power performance of Levene/Brown-Forsythe test is less satisfactory than its Type I error properties in 2-sample designs. Pan (1999) shows that for some populations, including the normal population, the power of the test in 2-sample designs has an upper bound that may be far below 1 regardless of the magnitude of the difference between the standard deviations. In other words, for these types of data, the test is more likely to conclude that there is no difference between the standard deviations regardless of how big the difference is. For these reasons, the Assistant uses a new test, the Bonett test, for the 2-Sample Standard Deviation test. For the standard deviations test with multiple-sample designs, the Assistant uses a multiple comparison (MC) procedure.

The Bonett (2006) test, a modified version of Layard's (1978) test of equality of two variances, enhances the test's performance with small samples. Banga and Fox (2013A) derive the confidence intervals associated with Bonett's test and show that they are as accurate as the confidence intervals associated with the Levene/Brown-Forsythe test and are more precise for most distributions. Additionally, Banga and Fox (2013A) determined that the Bonett test is as robust as Levene/Brown-Forsythe test and is more powerful for most distributions.

The multiple comparison (MC) procedure includes an overall test of the homogeneity, or equality, of the standard deviations (or variances) for multiple samples, which is based on the comparison intervals for each pair of standard deviations. The comparison intervals are derived so that the MC test is significant if, and only if, at least one pair of the comparison intervals do not overlap. Banga and Fox (2013B) show that the MC test has Type I and Type II error properties that are similar to the Levene/Brown-Forsythe test for most distributions. One important advantage of the MC test is the graphical display of the comparison intervals, which provides an effective visual tool for identifying the samples with different standard deviations. When there are only two samples in the design, the MC test is equivalent to the Bonett test.

In this paper, we evaluate the validity of the Bonett test and the MC test for different data distributions and sample sizes. In addition, we investigate the power and sample size analysis used for the Bonett test, which is based on a large-sample approximation method. Based on these factors, we developed the following checks that the Assistant automatically performs on your data and displays in the Report Card:

- Unusual data
- Normality
- Validity of test
- Sample size (2-Sample Standard Deviation test only)

Tests for standard deviations methods

Validity of Bonett's test and MC test

In their comparative study of tests for equal variances, Conover, et al. (1981) found that the Levene/Brown-Forsythe test was among the best performing tests, based on its Type I and Type II error rates. Since that time, other methods have been proposed for testing for equal variances in 2-sample and multiple-sample designs (Pan, 1999; Shoemaker, 2003; Bonett, 2006). For example, Pan shows that despite its robustness and simplicity of interpretation, the Levene/Brown-Forsythe test does not have sufficient power to detect important differences between 2 standard deviations when the samples originate from some populations, including the normal population. Because of this critical limitation, the Assistant uses the Bonett test for the 2-Sample Standard Deviation test (see Appendix A or Banga and Fox, 2013A). For the standard deviations test with more than 2 samples, the Assistant uses an MC procedure with comparison intervals that provides a graphical display to identify samples with different standard deviations when the MC test is significant (see Appendix A and Banga and Fox, 2013B).

Objective

First, we wanted to evaluate the performance of the Bonett test when comparing two population standard deviations. Second, we want to evaluate the performance of the MC test when comparing the standard deviations among more than two populations. Specifically, we wanted to evaluate the validity of these tests when they are performed on samples of various sizes from different types of distributions.

Method

The statistical methods used for the Bonett test and the MC test are defined in Appendix A. To evaluate the validity of the tests, we needed to examine whether their Type I error rates remained close to the target level of significance (α) under different conditions. To do this, we performed a set of simulations to evaluate the validity of the Bonett test when comparing the standard deviations from 2 independent samples and other sets of simulations to evaluate the validity of the MC test when comparing the standard deviations from multiple (k) independent samples, when $k > 2$.

We generated 10,000 pairs or multiple (k) random samples of various sizes from several distributions, using both balanced and unbalanced designs. Then we performed a two-sided Bonett test to compare the standard deviations of the 2 samples or performed a MC test to compare the standard deviations of the k samples in each experiment, using a target significance level of $\alpha = 0.05$. We counted the number of times out of 10,000 replicates that the test rejected the null hypothesis (when in fact the true standard deviations were equal) and

compared this proportion, known as the simulated significance level, to the target significance level. If the test performs well, the simulated significance level, which represents the actual Type I error rate, should be very close to the target significance level. For more details on the specific methods used for the 2-sample and k-sample simulations, see Appendix B.

Results

For 2-sample comparisons, the simulated Type I error rates of the Bonett test were close to the target level of significance when the samples were moderate or large in size, regardless of the distribution and regardless of whether the design was balanced or unbalanced. However, when small samples were drawn from extremely skewed populations, the Bonett test was generally conservative, and had Type I error rates that were slightly lower than the target level of significance (that is, the target Type I error rate).

For multiple-sample comparisons, the Type I error rates of the MC test were close to the target level of significance when the samples were moderate or large in size, regardless of the distribution and regardless of whether the design was balanced or unbalanced. For small and extremely skewed samples, however, the test was generally less conservative, and had Type I error rates that were higher than the target level of significance when the number of samples in the design is large.

The results of our studies were consistent with those of Banga and Fox (2013A) and (2013B). We concluded that the Bonett test and the MC test perform well when the size of the smallest sample is at least 20. Therefore, we use this minimum sample size requirement in the Validity of test check in the Assistant Report Card (see the Data check section).

Comparison intervals

When a test to compare two or more standard deviations is statistically significant, indicating that at least one of the standard deviations is different from the others, the next step in the analysis is to determine which samples are statistically different. An intuitive way to make this comparison is to graph the confidence intervals associated with each sample and identify the samples whose intervals do not overlap. However, the conclusions drawn from the graph may not match the test results because the individual confidence intervals are not designed for comparisons.

Objective

We wanted to develop a method to calculate individual comparison intervals that can be used as both an overall test of the homogeneity of variances and as a method to identify samples with different variances when the overall test is significant. A critical requirement for the MC procedure is that the overall test is significant if, and only if, at least one pair of the comparison intervals do not overlap, which indicates that the standard deviations of at least two samples are different.

Method

The MC procedure that we use to compare multiple standard deviations is derived from multiple pairwise comparisons. Each pair of samples is compared using the Bonett's (2006) test of equality of two population standard deviations. The pairwise comparisons use a multiplicity correction based on a large-sample approximation shown in Nayakama (2009). The large-sample approximation is preferred over the commonly-used Bonferroni correction because the Bonferroni correction becomes increasingly conservative as the number of samples increases. Finally, the comparison intervals result from the pairwise comparisons based on the Hochberg et al. (1982) best approximate procedure. For details, see Appendix A.

Results

The MC procedure satisfies the requirement that the overall test of the equality of standard deviations is significant if, and only if, at least two comparison intervals do not overlap. If the overall test is not significant, then all the comparison intervals must overlap.

The Assistant displays the comparison intervals in the Standard Deviations Comparison Chart in the Summary Report. Next to this graph, the Assistant displays the p-value of the MC test, which is the overall test for the homogeneity of the standard deviations. When the standard deviations test is statistically significant, any comparison interval that does not overlap with at least one other interval is marked in red. If the standard deviations test is not statistically significant, then none of the intervals are marked in red.

Performance of theoretical power (2-sample designs only)

The theoretical power functions of the Bonett and MC tests are needed for planning sample sizes. For 2-sample designs, an approximate theoretical power function of the test can be derived using large-sample theory methods. Because this function results from large-sample approximation methods, we need to evaluate its properties when the test is conducted using small samples generated from normal and nonnormal distributions. When comparing the standard deviations of more than two groups, however, the theoretical power function of the MC test is not easily obtained.

Objective

We wanted to determine whether we could use the theoretical power function based on the large-sample approximation to evaluate the power and sample size requirements for the 2-Sample Standard Deviation test in the Assistant. To do this, we needed to evaluate whether the approximated theoretical power function accurately reflects the actual power achieved by the Bonett test when it is performed on data from several types of distributions, including normal and nonnormal distributions.

Method

The approximated theoretical power function of the Bonett test for 2-sample designs is derived in Appendix C.

We performed simulations to estimate the actual power levels (which we refer to as simulated power levels) using the Bonett test. First, we generated pairs of random samples of various sizes from several distributions, including normal and nonnormal distributions. For each distribution, we performed the Bonett test on each of 10,000 pairs of sample replicates. For each pair of sample sizes, we calculated the simulated power of the test to detect a given difference as the fraction of the 10,000 pairs of samples for which the test is significant. For comparison, we also calculated the corresponding power level using the approximated theoretical power function of the test. If the approximation works well, the theoretical and simulated power levels should be close. For more details, see Appendix D.

Results

Our simulations showed that for most distributions the theoretical and simulated power functions of the Bonett test are nearly equal for small sample sizes and are closer when the minimum sample size reaches 20. For symmetric and nearly symmetric distributions with light to moderate tails the theoretical power levels are slightly higher than the simulated (actual) power levels. However, for skewed distributions and heavy-tailed distributions they are smaller than the simulated (actual) power levels. For more details, see Appendix D.

Overall, our results show that the theoretical power function provides a good basis for planning sample sizes.

Data checks

Unusual data

Unusual data are extremely large or small data values, also known as outliers. Unusual data can have a strong influence on the results of the analysis and can affect the chances of finding statistically significant results, especially when the sample is small. Unusual data can indicate problems with data collection, or may be due to unusual behavior of the process you are studying. Therefore, these data points are often worth investigating and should be corrected when possible. The simulation studies show that when the data contain outliers, the Bonett test and the MC test are conservative (see Appendix B). The actual levels of significance of the tests are markedly smaller than the targeted level, particularly when the analysis is performed with small samples.

Objective

We wanted to develop a method to check for data values that are very large or very small relative to the overall sample and that may affect the results of the analysis.



Method

We developed a method to check for unusual data based on the method described by Hoaglin, Iglewicz, and Tukey (1986) that is used to identify outliers in boxplots.

Results

The Assistant identifies a data point as unusual if it is more than 1.5 times the interquartile range beyond the lower or upper quartile of the distribution. The lower and upper quartiles are the 25th and 75th percentiles of the data. The interquartile range is the difference between the two quartiles. This method works well even when there are multiple outliers because it makes it possible to detect each specific outlier.

When checking for unusual data, the Assistant displays the following status indicators in the Report Card:

Status	Condition
	There are no unusual data points.
	At least one data point is unusual and may have a strong influence on the results.

Normality

Unlike most tests of equality of variances, which are derived under the normality assumption, the Bonett test and the MC test for equality of standard deviations do not make an assumption about the specific distribution of the data.

Objective

Although the Bonett test and the MC test are based on large-sample approximation methods, we wanted to confirm that they perform well for normal and nonnormal data in small samples. We also wanted to inform the user about how the normality of the data relates to the results of the standard deviations tests.

Method


To evaluate the validity of the tests under different conditions, we performed simulations to examine the Type I error rate of the Bonett test and the MC test with normal and nonnormal data of various sample sizes. For more details, see the Tests for standard deviations methods section and Appendix B.

Results


Our simulations showed that the distribution of the data does not have a major effect on the Type I error properties of the Bonett test or the MC test for sufficiently large samples (minimum sample size ≥ 20). The tests produce Type I error rates that are consistently close to the target error rate for both normal and nonnormal data.

Based on these results concerning the Type I error rate, the Assistant displays the information about normality in the Report Card.

For 2-sample designs, the Assistant displays the following indicator:

Status	Condition
	This analysis uses the Bonett Test. With sufficiently large samples, the test performs well for both normal and nonnormal data.

For multiple-sample designs, the Assistant displays the following indicator:

Status	Condition
	This analysis uses a Multiple Comparison Test. With sufficiently large samples, the test performs well for both normal and nonnormal data.

Validity of test

In the Tests for standard deviations methods section, we showed that for both 2-sample and multiple (k) comparisons, the Bonett test and the MC test produce Type I error rates close to the target error rate for normal as well as nonnormal data in balanced and unbalanced designs when the samples are moderate or large in size. However, when the samples are small, the Bonett and the MC tests don't generally perform well.

Objective



We wanted to apply a rule to evaluate the validity of the standard deviation test results for 2 samples and for multiple (k) samples, based on the user's data.

Method

To evaluate the validity of the tests under different conditions, we performed simulations to examine the Type I error rate of the Bonett test and the MC test with various distributions of data, numbers of samples, and sample sizes, as described previously in the Tests for standard deviations methods section. For more details, see Appendix B.

Results

The Bonnet test and the MC test perform well when the size of the smallest sample is at least 20. Therefore, the Assistant displays the following status indicators in the Report Card to evaluate the validity of the standard deviations tests.

Status	Condition
	The sample sizes are at least 20, so the p-value should be accurate.
	Some of the sample sizes are less than 20, so the p-value may not be accurate. Consider increasing the sample sizes to at least 20.

Sample size (for 2-Sample Standard Deviations test only)

Typically, a statistical hypothesis test is performed to gather evidence to reject the null hypothesis of "no difference". If the sample is too small, the power of the test may not be adequate to detect a difference that actually exists, which results in a Type II error. It is therefore crucial to ensure that the sample sizes are sufficiently large to detect practically important differences with high probability.

Objective

If the data does not provide sufficient evidence to reject the null hypothesis, we wanted to determine whether the sample sizes are large enough for the test to detect practical differences of interest with high probability. Although the objective of sample size planning is to ensure that sample sizes are large enough to detect important differences with high probability, they should not be so large that meaningless differences become statistically significant with high probability.




Method



The power and sample size analysis for the 2-Sample Standard Deviations test is based upon an approximation of the power function of the Bonett test, which usually provides good estimates of the actual power function of the test (see the simulation results summarized in Performance of theoretical power function in the Method section).

Results

When the data does not provide enough evidence against the null hypothesis, the Assistant uses the approximate power function of the Bonett test to calculate the practical differences that can be detected with an 80% and a 90% probability for the given sample size. In addition, if the user provides a particular practical difference of interest, the Assistant uses the power function of the normal approximation test to calculate sample sizes that yield an 80% and a 90% chance of detection of the difference.

To help interpret the results, the Assistant Report Card for the 2-Sample Standard Deviations Test displays the following status indicators when checking for power and sample size:

Status	Condition
	The test finds a difference between the standard deviations, so power is not an issue. OR Power is sufficient. The test did not find a difference between the standard deviations, but the sample is large enough to provide at least a 90% chance of detecting the given difference.
	Power may be sufficient. The test did not find a difference between the standard deviations, but the sample is large enough to provide an 80% to 90% chance of detecting the given difference. The sample size required to achieve 90% power is reported.
	Power might not be sufficient. The test did not find a difference between the standard deviations, and the sample is large enough to provide a 60% to 80% chance of detecting the given difference. The sample sizes required to achieve 80% power and 90% power are reported.

Status	Condition
	<p>Power is not sufficient. The test did not find a difference between the standard deviations, and the sample is not large enough to provide at least a 60% chance of detecting the given difference. The sample sizes required to achieve 80% power and 90% power are reported.</p>
	<p>The test did not find a difference between the standard deviations. You did not specify a practical difference to detect; therefore, the report indicates the differences that you could detect with 80% and 90% chance, based on your sample size and alpha.</p>

References

- Arnold, S.F. (1990). *Mathematical statistics*. Englewood Cliffs, NJ: Prentice-Hall, Inc.
- Banga, S.J. and Fox, G.D. (2013A). On Bonnett's Robust Confidence Interval for a Ratio of Standard Deviations. *White paper, Minitab Inc.*
- Banga, S.J. and Fox, G.D. (2013B) A graphical multiple comparison procedure for several standard deviations. *White paper, Minitab Inc.*
- Bonett, D.G. (2006). Robust confidence interval for a ratio of standard deviations. *Applied Psychological Measurements*, 30, 432-439.
- Brown, M.B., & 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., & 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.
- Gastwirth, J. L. (1982). Statistical properties of a measure of tax assessment uniformity. *Journal of Statistical Planning and Inference*, 6, 1-12.
- 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.), *Probability and statistics* (278-292). Stanford University Press, Palo Alto, California.
- Nakayama, M.K. (2009). Asymptotically valid single-stage multiple-comparison procedures. *Journal of Statistical Planning and Inference*, 139, 1348-1356.
- 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 (2), 105-114.

Appendix A: Method for The Bonett test and the Multiple comparison test

The underlying assumptions for making inferences about the standard deviations or variances using the Bonett method (2-sample designs) or the multiple comparison (MC) procedure (multiple-sample designs) can be described as follow. Let $X_{11}, \dots, X_{1n_1}, \dots, X_{k1}, \dots, X_{kn_k}$ be k ($k \geq 2$) independent random samples, with each sample drawn from a distribution with an unknown mean μ_i and variance σ_i^2 , respectively, for $i = 1, \dots, k$. Let's assume that the parent distributions of the samples have a common finite kurtosis $\gamma = E(Y - \mu)^4 / \sigma^4 < \infty$. While this assumption is crucial for the theoretical derivations, it is not critical for most practical applications where the samples are sufficiently large (Banga and Fox, 2013A).

Method A1: Bonett test of equality of two variances

The Bonett test only applies to 2-sample designs where two variances or standard deviations are compared. The test is a modified version of Layard (1978) test of equality of variances in two-sample designs. A two-sided Bonett's test of equality of two variances with significance level α rejects the null hypothesis of equality if, and only if,

$$|\ln(c S_1^2/S_2^2)| > z_{\alpha/2} \sqrt{\frac{\hat{\gamma}_P - g_1}{n_1 - 1} + \frac{\hat{\gamma}_P - g_2}{n_2 - 1}}$$

where:

S_i is the sample standard deviation of sample i

$$g_i = (n_i - 3)/n_i, i = 1,2$$

$z_{\alpha/2}$ refers to the upper $\alpha/2$ percentile of the standard normal distribution

$\hat{\gamma}_P$ is the pooled kurtosis estimator given as:

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

In the expression of the pooled kurtosis estimator, m_i is the trimmed mean for sample i , with the trim proportion, $1/[2(n_i - 4)^{1/2}]$.

In the above, the constant c is included as a small sample adjustment to reduce the effect of unequal tail error probabilities in unbalanced designs. This constant is given as $c = c_1/c_2$, where

$$c_i = \frac{n_i}{n_i - z_{\alpha/2}}, i = 1,2$$

If the design is balanced, that is if $n_1 = n_2$, then the p-value of the test is obtained as

$$P = 2 \Pr(Z > z)$$

where Z is a random variable distributed as the standard normal distribution and z the observed value of the following statistics based on the data at hand. The statistic is

$$Z = \frac{\ln(C S_1^2/S_2^2)}{se}$$

where

$$se = \sqrt{\frac{\hat{y}_P - g_1}{n_1 - 1} + \frac{\hat{y}_P - g_2}{n_2 - 1}}$$

On the other hand, if the design is unbalanced then the p-value of the test is obtained as

$$P = 2\min(\alpha_L, \alpha_U)$$

where $\alpha_L = \Pr(Z > z_L)$ and $\alpha_U = \Pr(Z > z_U)$. The variable z_L is the smallest root of the function

$$L(z, S_1, S_2, n_1, n_2) = \ln \frac{n_1}{n_2} + \ln \frac{n_2 - z}{n_1 - z} - z se + \ln \frac{S_1^2}{S_2^2} - \ln \rho_0^2, z < \min(n_1, n_2)$$

and z_U is the smallest root of the function $L(z, S_2, S_1, n_2, n_1)$.

Method A2: Multiple comparison test and comparison intervals

Suppose that there are k ($k \geq 2$) independent groups or samples. Our objective was to construct a system of k intervals for the population standard deviations such that the test of equality of the standard deviations is significant if, and only if, at least two of the k intervals do not overlap. These intervals are referred to as comparison intervals. This method of comparison is similar to the procedures for multiple comparisons of the means in one-way ANOVA models, which were initially developed by Tukey-Kramer and later generalized by Hochberg, et al. (1982).

Comparing two standard deviations

For 2-sample designs, the confidence intervals of the ratio of standard deviations associated with the Bonett test can be calculated directly to assess the size of difference between the standard deviations (Banga and Fox, 2013A). In fact, we use this approach for Stat > Basic Statistics > 2 Variances in release 17 of Minitab. In the Assistant, however, we wanted to provide comparison intervals that are easier to interpret than the confidence interval of the ratio of standard deviations. To do this, we used the Bonett procedure described in Method A1 to determine the comparison intervals for two samples.

When there are two samples, the Bonett test of equality of variances is significant if, and only if, the following acceptance interval associated with the Bonett test of equality of variances does not contain 0.

$$\ln(c_1 S_1^2) - \ln(c_2 S_2^2) \pm z_{\alpha/2} \sqrt{\frac{\hat{y}_P - g_1}{n_1 - 1} + \frac{\hat{y}_P - g_2}{n_2 - 1}}$$

where the pool kurtosis estimate $\hat{\gamma}_P$, and $g_i, i = 1,2$ are as previously given.

From this interval, we deduce the following two comparison intervals such that the test of equality of variances or standard deviation is significant if, and only if, they don't overlap. These two intervals are

$$\left[S_i \sqrt{C_i \exp(-z_{\alpha/2} V_i)}, S_i \sqrt{C_i \exp(z_{\alpha/2} V_i)} \right], i = 1,2$$

where

$$V_i = \frac{\sqrt{\frac{\hat{\gamma}_P - g_i}{n_i - 1}}}{\sqrt{\frac{\hat{\gamma}_P - g_i}{n_i - 1} + \frac{\hat{\gamma}_P - g_j}{n_j - 1}}} \sqrt{\frac{\hat{\gamma}_P - g_i}{n_i - 1} + \frac{\hat{\gamma}_P - g_j}{n_j - 1}}, i = 1,2; j = 1,2; i \neq j$$

Using these intervals as a testing procedure of equality of the standard deviation is equivalent to the Bonett test of equality of standard deviations. Specifically, the intervals don't overlap if, and only if, the Bonett test of equality of standard deviation is significant. Note, however, that these intervals are not confidence intervals of standard deviations, but are only appropriate for multiple comparisons of standard deviations. Hochberg et al. refer to similar intervals for comparing means as uncertainty intervals for the same reason. We refer to these intervals as comparison intervals.

Because the comparison intervals procedure is equivalent to the Bonett test of equality of standard deviation, the p-value associated with the comparison intervals is identical to the p-value of the Bonett test of equality of two standard deviations described earlier.

Comparing multiple standard deviations

When there are more than two groups or samples, the k comparison intervals are deduced from $k(k - 1)/2$ pairwise simultaneous tests of equality of standard deviations with family wise significance level α . More specifically, let X_{i1}, \dots, X_{in_i} and X_{j1}, \dots, X_{jn_j} be the sample data for any pair (i, j) of samples. Similar to the 2-sample case, the test of equality of the standard deviations for the particular pair (i, j) of samples is significant at some α' level if, and only if, the interval

$$\ln(c_i S_i^2) - \ln(c_j S_j^2) \pm z_{\alpha'/2} \sqrt{\frac{\hat{\gamma}_{ij} - g_i}{n_i - 1} + \frac{\hat{\gamma}_{ij} - g_j}{n_j - 1}}$$

does not contain 0. In the above $\hat{\gamma}_{ij}$ is the pooled kurtosis estimator based on the pair (i, j) of samples and is given as

$$\hat{\gamma}_{ij} = (n_i + n_j) \frac{\sum_{l=1}^{n_i} (X_{il} - m_i)^4 + \sum_{l=1}^{n_j} (X_{jl} - m_j)^4}{[(n_i - 1)S_i^2 + (n_j - 1)S_j^2]^2}$$

In addition, as previously defined, m_i is the trimmed mean for sample i , with the trim proportion, $1/[2(n_i - 4)^{1/2}]$ and

$$g_i = \frac{n_i - 3}{n_i}, g_j = \frac{n_j - 3}{n_j}, c_i = \frac{n_i}{n_i - z_{\alpha'/2}}, c_j = \frac{n_j}{n_j - z_{\alpha'/2}}$$

Because there are $k(k - 1)/2$ simultaneous pairwise tests, the level α' must be chosen so that the actual family wise error rate is close to the target level of significance α . One possible adjustment is based on Bonferroni's approximation. However, Bonferroni corrections are well known to be increasingly conservative as the number of samples in the design increases. A better approach is based on a normal approximation given by Nakayama (2008). Using this approach we merely replace $z_{\alpha'/2}$ with $q_{\alpha,k}/\sqrt{2}$, where $q_{\alpha,k}$ is the upper α point of the range of k independent and identically distributed standard normal random variables; that is

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

where Z_1, \dots, Z_k are independent and identically distributed standard normal random variables.

Furthermore, using a method similar to Hochberg et al. (1982), the procedure that best approximates the pairwise procedure described above, rejects the null hypothesis of the equality of standard deviations if, and only if, for some pair (i, j) of samples

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

where V_i is chosen to minimize the quantity

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

with

$$b_{ij} = \sqrt{\frac{\hat{y}_{ij} - g_i}{n_i - 1} + \frac{\hat{y}_{ij} - g_j}{n_j - 1}}$$

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

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

It follows that the test based on the approximate procedure is significant if, and only if, at least one pair of the following k intervals don't overlap.

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

To calculate the overall p-value associated with the MC test, we let P_{ij} be the p-value associated with any pair (i, j) of samples. It follows then that the overall p-value associated with the multiple comparison test is

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

To calculate P_{ij} we perform the algorithm of the 2-sample design given in Method A1 using

$$se = V_i + V_j$$

where V_i is as given above.

More specifically, if $n_i \neq n_j$

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

where $\alpha_L = \Pr(Q > z_L\sqrt{2})$, $\alpha_U = \Pr(Q > z_U\sqrt{2})$, the variable z_L is the smallest root of the function $L(z, S_i, S_j, n_i, n_j)$, the variable z_U is the smallest root of the function $L(z, S_j, S_i, n_j, n_i)$ and Q is a random variable which has the range distribution as previously defined.

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

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

Appendix B: Validity of Bonett test and the Multiple comparison test

Simulation B1: Validity of Bonett Test (2-sample models, balanced and unbalanced designs)

We generated pairs of random samples that are small to moderate in size from distributions with different properties. The distributions included:

- Standard normal distribution ($N(0,1)$)
- Symmetric and light-tailed distributions, including the uniform distribution ($U(0,1)$) and the Beta distribution with both parameters set to 3 ($B(3,3)$)
- Symmetric and heavy-tailed distributions, including t distributions with 5 and 10 degrees of freedom ($t(5), t(10)$), and the Laplace distribution with location 0 and scale 1 (Lpl)
- Skewed and heavy-tailed distributions, including the exponential distribution with scale 1 (Exp) and chi-square distributions with 5 and 10 degrees of freedom ($Chi(5), Chi(10)$)
- Left-skewed and heavy-tailed distribution; specifically, the Beta distribution with the parameters set to 8 and 1, respectively ($B(8,1)$)

In addition, to assess the direct effect of outliers, we generated pairs of samples from contaminated normal distributions defined as

$$CN(p, \sigma) = pN(0,1) + (1 - p)N(0, \sigma)$$

where p is the mixing parameter and $1 - p$ is the proportion of contamination (which equals the proportion of outliers). We selected two contaminated normal populations for the study: $CN(0.9,3)$, where 10% of the population are outliers; and $CN(0.8,3)$, where 20% of the population are outliers. These two distributions are symmetric and have long tails due to the outliers.

We performed a two-sided Bonett test with a target significance level of $\alpha = 0.05$ on each pair of samples from each distribution. Because the simulated significance levels were, in each case, based upon 10,000 pairs of samples replicates, and because we used a target significance level of 5%, the simulation error was $\sqrt{0.95(0.05)/10,000} = 0.2\%$.

The simulation results are summarized in Table 1 below.

Table 1 Simulated significance levels for a two-sided Bonett test in balanced and unbalanced 2-sample designs. The target level of significance is 0.05.

Distribution	n_1, n_2	Simulated level	Distribution	n_1, n_2	Simulated level
N(0,1)	10, 10	0.038	Exp	10, 10	0.052
	20, 10	0.043		20, 10	0.051
	20, 20	0.045		20, 20	0.049
	30, 10	0.044		30, 10	0.044
	30, 20	0.046		30, 20	0.042
	25, 25	0.048		25, 25	0.043
	30, 30	0.048		30, 30	0.042
	40, 40	0.051		40, 40	0.042
	50, 50	0.047		50, 50	0.039
t(5)	10, 10	0.044	Chi(5)	10, 10	0.040
	20, 10	0.042		20, 10	0.043
	20, 20	0.046		20, 20	0.040
	30, 10	0.041		30, 10	0.039
	30, 20	0.046		30, 20	0.043
	25, 25	0.048		25, 25	0.042
	30, 30	0.043		30, 30	0.043
	40, 40	0.046		40, 40	0.040
	50, 50	0.050		50, 50	0.039

Distribution	n_1, n_2	Simulated level	Distribution	n_1, n_2	Simulated level
t(10)	10, 10	0.041	Chi(10)	10, 10	0.044
	20, 10	0.040		20, 10	0.042
	20, 20	0.045		20, 20	0.041
	30, 10	0.046		30, 10	0.043
	30, 20	0.045		30, 20	0.045
	25, 25	0.046		25, 25	0.046
	30, 30	0.048		30, 30	0.038
	40, 40	0.045		40, 40	0.042
	50, 50	0.051		50, 50	0.049
Lpl	10, 10	0.054	B(8,1)	10, 10	0.053
	20, 10	0.056		20, 10	0.045
	20, 20	0.055		20, 20	0.048
	30, 10	0.057		30, 10	0.042
	30, 20	0.058		30, 20	0.047
	25, 25	0.057		25, 25	0.041
	30, 30	0.053		30, 30	0.040
	40, 40	0.047		40, 40	0.042
	50, 50	0.048		50, 50	0.038

Distribution	n_1, n_2	Simulated level	Distribution	n_1, n_2	Simulated level
B(3,3)	10, 10	0.032	CN(0.9,3)	10, 10	0.024
	20, 10	0.037		20, 10	0.022
	20, 20	0.042		20, 20	0.018
	30, 10	0.039		30, 10	0.019
	30, 20	0.038		30, 20	0.020
	25, 25	0.039		25, 25	0.019
	30, 30	0.041		30, 30	0.015
	40, 40	0.044		40, 40	0.020
	50, 50	0.046		50, 50	0.017
U(0,1)	10, 10	0.030	CN(0.8,3)	10, 10	0.022
	20, 10	0.032		20, 10	0.019
	20, 20	0.031		20, 20	0.020
	30, 10	0.034		30, 10	0.017
	30, 20	0.034		30, 20	0.020
	25, 25	0.034		25, 25	0.021
	30, 30	0.037		30, 30	0.017
	40, 40	0.043		40, 40	0.023
	50, 50	0.043		50, 50	0.020

As shown Table 1, when the sample sizes are smaller, the simulated significance levels of the Bonett test are lower than the target level of significance (0.05) for symmetric or nearly symmetric distributions with light to moderate tails. On the other hand, the simulated levels tend to a bit larger than the targeted level when small samples originate from highly skewed distributions.

When the samples are moderately large or large in size, the simulated significance levels are close to the target level for all the distributions. In fact, the test performs reasonably well even for highly skewed distributions, such as the exponential distribution and the Beta(8,1) distribution.

In addition, outliers appear to have more impact in small samples than in large samples. The simulated significance levels for the contaminated normal populations stabilized at approximately 0.020 when the minimum size of the two samples reached 20.

When the minimum size of the two samples is 20, the simulated significance levels consistently fall within the interval [0.038, 0.058], except for the flat uniform distribution and contaminated normal distributions. Although a simulated significance level of 0.040 is slightly conservative for a target level of 0.05, this Type I error rate may be acceptable for most practical purposes. Therefore, we conclude that the Bonett test is valid when the minimum size of the two samples is at least 20.

Simulation B2: Validity of the MC test (multiple-sample models)

Part I: Balanced designs

We performed a simulation to examine the performance of the MC test in multiple-sample models with balanced designs. We generated k samples of equal size from the same distribution, using the set of distributions previously listed in simulation B1. We selected the number of samples in a design to be $k = 3$, $k = 4$, and $k = 6$ and fixed the size of the k samples in each experiment at 10, 15, 20, 25, 50, and 100.

We performed a two-sided MC test with a target significance level of $\alpha = 0.05$ on the same samples of each design case. Because the simulated significance levels were, in each case, based upon 10,000 pairs of sample replicates, and because we used a target significance level of 5%, the simulation error was $\sqrt{0.95(0.05)/10,000} = 0.2\%$.

The simulation results are summarized in Tables 2a and 2b below.

Table 2a Simulated significance levels for a two-sided multiple comparison test in balanced, multi-sample designs. The target level of significance for the test is 0.05.

Distribution	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \dots = n_6$	
	n_i	Simulated level	n_i	Simulated level	n_i	Simulated level
N(0,1)	10	0.038	10	0.038	10	0.036
	15	0.040	15	0.041	15	0.039
	20	0.039	20	0.040	20	0.041
	25	0.045	25	0.047	25	0.047
	50	0.046	50	0.046	50	0.052

Distribution	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \dots = n_6$	
	n_i	Simulated level	n_i	Simulated level	n_i	Simulated level
t(5)	100	0.049	100	0.049	100	0.052
	10	0.042	10	0.044	10	0.042
	15	0.041	15	0.044	15	0.046
	20	0.043	20	0.045	20	0.045
	25	0.046	25	0.048	25	0.046
	50	0.040	50	0.039	50	0.038
	100	0.038	100	0.040	100	0.040
T(10)	10	0.033	10	0.037	10	0.038
	15	0.040	15	0.042	15	0.041
	20	0.042	20	0.043	20	0.043
	25	0.041	25	0.042	25	0.045
	50	0.047	50	0.044	50	0.047
	100	0.048	100	0.046	100	0.047
Lpl	10	0.056	10	0.063	10	0.071
	15	0.056	15	0.061	15	0.063
	20	0.054	20	0.058	20	0.059
	25	0.051	25	0.056	25	0.58
	50	0.045	50	0.051	50	0.049
	100	0.044	100	0.047	100	0.050
B(3,3)	10	0.031	10	0.031	10	0.031
	15	0.037	15	0.036	15	0.034
	20	0.035	20	0.036	20	0.037
	25	0.039	25	0.038	25	0.040
	50	0.044	50	0.044	50	0.044

Distribution	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \dots = n_6$	
	n_i	Simulated level	n_i	Simulated level	n_i	Simulated level
U(0,1)	100	0.044	100	0.046	100	0.043
	10	0.029	10	0.025	10	0.023
	15	0.026	15	0.027	15	0.026
	20	0.028	20	0.030	20	0.028
	25	0.034	25	0.033	25	0.032
	50	0.041	50	0.036	50	0.036
Exp	100	0.048	100	0.047	100	0.045
	10	0.063	10	0.073	10	0.076
	15	0.056	15	0.058	15	0.064
	20	0.051	20	0.053	20	0.057
	25	0.043	25	0.045	25	0.050
	50	0.033	50	0.037	50	0.038
	100	0.033	100	0.035	100	0.035

Table 2b Simulated significance levels for a two-sided multiple comparison test in balanced, multi-sample designs. The target level of significance for the test is 0.05.

Distribution	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \dots = n_6$	
	n_i	Simulated level	n_i	Simulated level	n_i	Simulated level
Chi(5)	10	0.040	10	0.046	10	0.048
	15	0.043	15	0.046	15	0.049
	20	0.040	20	0.040	20	0.042
	25	0.040	25	0.045	25	0.042
	50	0.037	50	0.038	50	0.040
	100	0.036	100	0.037	100	0.038

Distribution	$k = 3$ $n_1 = n_2 = n_3$		$k = 4$ $n_1 = n_2 = n_3 = n_4$		$k = 6$ $n_1 = n_2 = \dots = n_6$	
	n_i	Simulated level	n_i	Simulated level	n_i	Simulated level
Chi(10)	10	0.042	10	0.045	10	0.045
	15	0.038	15	0.044	15	0.047
	20	0.036	20	0.039	20	0.040
	25	0.043	25	0.044	25	0.045
	50	0.041	50	0.040	50	0.042
	100	0.038	100	0.040	100	0.042
B(8,1)	10	0.058	10	0.060	10	0.066
	15	0.057	15	0.061	15	0.064
	20	0.049	20	0.051	20	0.055
	25	0.044	25	0.046	25	0.050
	50	0.037	50	0.037	50	0.039
	100	0.037	100	0.038	100	0.039
CN(0.9,3)	10	0.020	10	0.018	10	0.016
	15	0.022	15	0.020	15	0.017
	20	0.014	20	0.012	20	0.008
	25	0.011	25	0.011	25	0.008
	50	0.009	50	0.007	50	0.006
	100	0.010	100	0.008	100	0.008
CN(0.8, 3)	10	0.017	10	0.015	10	0.011
	15	0.013	15	0.011	15	0.008
	20	0.012	20	0.012	20	0.009
	25	0.013	25	0.010	25	0.009
	50	0.011	50	0.011	50	0.009
	100	0.014	100	0.012	100	0.010

As shown in Tables 2a and 2b, when the sample size is small, the MC test is generally conservative for symmetric and nearly symmetric distributions in balanced designs. On the other hand, the test is liberal for small samples obtained from highly skewed distributions such as the exponential and the beta(8, 1) distributions. As the sample size increases, however, the simulated significance levels approach the target significance level (0.05). In addition, the number of samples does not appear to have a strong effect on the performance of the test for samples that are moderate in sizes. When the data is contaminated with outliers, however, there is a remarkable impact on the performance of the test. The test is consistently and excessively conservative when outliers are present in the data.

Part II: Unbalanced designs

We performed a simulation to examine the performance of the MC test in unbalanced designs. We generated 3 samples from the same distribution, using the set of distributions previously described in Simulation B1. In the first set of experiments, the size of the first two samples was $n_1 = n_2 = 10$ and size of the third sample was $n_3 = 15, 20, 25, 50, 100$. In the second set of experiments, the size of the first two samples was $n_1 = n_2 = 15$ and the size of the third set of samples was $n_3 = 20, 25, 30, 50, 100$. In the third set of experiments, we set the minimum sample size at 20, with the size of the first two samples at $n_1 = n_2 = 20$ and the size of the third sample at $n_3 = 25, 30, 40, 50, 100$.

We performed a two-sided MC test with a target significance level of $\alpha = 0.05$ on the same three samples from each distribution. Because the simulated significance levels were, in each case, based upon 10,000 pairs of samples replicates, and because we used a target significance level of 5%, the simulation error was $\sqrt{0.95(0.05)/10,000} = 0.2\%$.

The simulation results are summarized in Tables 3a and 3b below.

Table 3a Simulated significance levels for the multiple comparison test in multi-sample, unbalanced designs. The target level of significance of the test is 0.05.

Distribution	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Simulated level	n_3	Simulated level	n_3	Simulated level
N(0,1)	15	0.032	20	0.040	25	0.045
	20	0.037	25	0.039	30	0.041
	25	0.038	30	0.037	40	0.043
	50	0.041	50	0.044	50	0.041
	100	0.042	100	0.042	100	0.044

Distribution	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Simulated level	n_3	Simulated level	n_3	Simulated level
t(5)	15	0.040	20	0.042	25	0.043
	20	0.036	25	0.040	30	0.037
	25	0.044	30	0.036	40	0.038
	50	0.033	50	0.036	50	0.035
	100	0.032	100	0.031	100	0.032
t(10)	15	0.039	20	0.042	25	0.042
	20	0.038	25	0.041	30	0.040
	25	0.040	30	0.041	40	0.041
	50	0.037	50	0.043	50	0.042
	100	0.036	100	0.039	100	0.040
Lpl	15	0.059	20	0.060	25	0.054
	20	0.057	25	0.054	30	0.051
	25	0.056	30	0.051	40	0.050
	50	0.049	50	0.051	50	0.050
	100	0.048	100	0.047	100	0.046
B(3,3)	15	0.034	20	0.033	25	0.037
	20	0.031	25	0.035	30	0.039
	25	0.031	30	0.034	40	0.039
	50	0.036	50	0.039	50	0.038
	100	0.035	100	0.039	100	0.039
U(0,1)	15	0.027	20	0.030	25	0.032
	20	0.030	25	0.030	30	0.031
	25	0.028	30	0.032	40	0.036
	50	0.039	50	0.034	50	0.037
	100	0.042	100	0.038	100	0.042

Distribution	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Simulated level	n_3	Simulated level	n_3	Simulated level
Exp	15	0.061	20	0.053	25	0.042
	20	0.060	25	0.052	30	0.047
	25	0.054	30	0.049	40	0.043
	50	0.050	50	0.046	50	0.041
	100	0.044	100	0.040	100	0.040

Table 3b Simulated significance levels for the MC test in multi-sample, unbalanced designs. The target level of significance of the test is 0.05.

Distribution	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Simulated level	n_3	Simulated level	n_3	Simulated level
Chi(5)	15	0.047	20	0.045	25	0.041
	20	0.043	25	0.042	30	0.039
	25	0.043	30	0.039	40	0.040
	50	0.039	50	0.037	50	0.040
	100	0.034	100	0.035	100	0.034
Chi(10)	15	0.043	20	0.042	25	0.042
	20	0.039	25	0.038	30	0.041
	25	0.040	30	0.041	40	0.038
	50	0.038	50	0.041	50	0.042
	100	0.035	100	0.034	100	0.035
B(8,1)	15	0.056	20	0.052	25	0.048
	20	0.054	25	0.046	30	0.044
	25	0.050	30	0.047	40	0.046
	50	0.046	50	0.043	50	0.043
	100	0.043	100	0.042	100	0.044
CN(0.9,3)	15	0.017	20	0.020	25	0.017

Distribution	$n_1 = n_2 = 10$		$n_1 = n_2 = 15$		$n_1 = n_2 = 20$	
	n_3	Simulated level	n_3	Simulated level	n_3	Simulated level
	20	0.020	25	0.019	30	0.012
	25	0.017	30	0.016	40	0.013
	50	0.019	50	0.016	50	0.012
	100	0.014	100	0.016	100	0.010
CN(0.8, 3)	15	0.012	20	0.013	25	0.013
	20	0.016	25	0.012	30	0.012
	25	0.014	30	0.010	40	0.010
	50	0.015	50	0.010	50	0.013
	100	0.012	100	0.011	100	0.010

The simulated significance levels shown in Tables 3a and 3b are consistent with those reported previously for multiple samples with balanced designs. Therefore, the performance of the MC test does not appear to be affected by unbalanced designs. In addition, when the minimum sample size is at least 20, then the simulated levels of significance are close to the target level, except for contaminated data.

In conclusion, when the smallest sample is at least 20, the MC test performs well for multiple (k) samples in both balanced and unbalanced designs. For smaller samples, however, the test is conservative for symmetric and nearly symmetric data and liberal for highly skewed data.

Appendix C: Theoretical power function

The exact theoretical power function of the MC test is not available. However, for 2-sample designs, an approximate power function based on large-sample theory methods can be obtained. For multiple-sample designs, more research efforts are required to derive a similar approximation.

For 2-sample designs, however, the theoretical power function of the Bonett test can be obtained using large-sample theory methods. More specifically, the test statistic, T , given below is asymptotically distributed as a chi-square distribution with 1 degree freedom:

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

In this expression of T , $\hat{\rho} = S_1/S_2$, $\rho = \sigma_1/\sigma_2$, $g_i = (n_i - 3)/n_i$, and γ is the unknown common kurtosis of the two populations.

It follows then that the theoretical power function of a two-sided Bonett test of equality of variances with an approximate level of significance α may be given as

$$\pi(n_1, n_2, \rho) = 1 - \Phi\left(z_{\alpha/2} - \frac{\ln \rho^2}{se}\right) + \Phi\left(-z_{\alpha/2} - \frac{\ln \rho^2}{se}\right)$$

where

$$se = \sqrt{\frac{\gamma - g_1}{n_1 - 1} + \frac{\gamma - g_2}{n_2 - 1}}$$

For one-sided tests, the approximate power function when testing against $\sigma_1 > \sigma_2$ is

$$\pi(n_1, n_2, \rho) = 1 - \Phi\left(z_{\alpha/2} - \frac{\ln \rho^2}{se}\right)$$

and when testing against $\sigma_1 < \sigma_2$, the approximate power function is

$$\pi(n_1, n_2, \rho) = \Phi\left(-z_{\alpha/2} - \frac{\ln \rho^2}{se}\right)$$

Note that during the planning of sample size phase of the data analysis, the common kurtosis of the populations, γ , is unknown. Therefore, the investigator typically must rely upon the opinions of experts or the results of previous experiments to obtain a planning value for γ . If that information is not available, it is often a good practice to perform a small pilot study to develop the plans for the major study. Using the samples from the pilot study, a planning value of γ is obtained as the pooled kurtosis given by

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

In the Assistant Menu, the planning estimate of γ is obtained retrospectively based on the user's data at hand.

Appendix D: Comparison of theoretical and simulated power

Simulation D1: Simulated (actual) power of the Bonett test

We performed a simulation to compare the simulated power levels of the Bonett test to the power levels based upon the approximate power function derived in Appendix C.

We generated 10,000 pairs of samples for each of the distributions described previously (see Simulation B1). In general, the selected sample sizes were large enough for the simulated significance level of the test to be reasonably close to the target significance level, based on our previous results in Simulation B1.

To evaluate the simulated power levels at a ratio of standard deviations $\rho = \sigma_1/\sigma_2 = 1/2$, we multiplied the second sample in every pair of samples by the constant 2. As a result, for a given distribution and for given sample sizes n_1 and n_2 , the simulated power level was calculated as the fraction of the 10,000 pairs of samples replicates for which the two-sided Bonett test was significant. The target significance level of the test was fixed at $\alpha = 0.05$. For comparison, we calculated the corresponding theoretical power levels based on the approximate power function derived in Appendix C.

The results are shown in Tables 4 below.

Table 4 Comparison of simulated power levels to approximate power levels of a two-sided Bonett test. The target significance level is 0.05.

Distribution	n_1, n_2	App. Power	Simulated Power	Distribution	n_1, n_2	App. Power	Simulated Power
N(0,1)	20, 10	0.627	0.527	Exp	20, 10	0.222	0.227
	20, 20	0.830	0.765		20, 20	0.322	0.368
	20, 30	0.896	0.846		20, 30	0.377	0.434
	20, 40	0.925	0.886		20, 40	0.412	0.475
	30, 15	0.825	0.771		30, 15	0.320	0.307
	30, 30	0.954	0.925		30, 30	0.458	0.500
	30, 45	0.980	0.970		30, 45	0.531	0.579

Distribution	n_1, n_2	App. Power	Simulated Power	Distribution	n_1, n_2	App. Power	Simulated Power
	30, 60	0.989	0.984		30, 60	0.575	0.622
t(5)	20, 10	0.222	0.379	Chi(5)	20, 10	0.355	0.347
	20, 20	0.322	0.569		20, 20	0.517	0.530
	20, 30	0.377	0.637		20, 30	0.597	0.616
	20, 40	0.412	0.690		20, 40	0.644	0.661
	30, 15	0.320	0.545		30, 15	0.513	0.510
	30, 30	0.458	0.733		30, 30	0.701	0.711
	30, 45	0.531	0.795		30, 45	0.781	0.793
	30, 60	0.575	0.828		30, 60	0.823	0.833
t(10)	20, 10	0.476	0.450	Chi(10)	20, 10	0.454	0.414
	20, 20	0.673	0.673		20, 20	0.646	0.631
	20, 30	0.756	0.749		20, 30	0.730	0.717
	20, 40	0.800	0.803		20, 40	0.776	0.771
	30, 15	0.668	0.659		30, 15	0.641	0.618
	30, 30	0.850	0.852		30, 30	0.828	0.819
	30, 45	0.910	0.911		30, 45	0.892	0.882
	30, 60	0.936	0.937		30, 60	0.921	0.912
Lpl	20, 10	0.321	0.330	B(8,1)	20, 10	0.363	0.278
	20, 20	0.469	0.519		20, 20	0.528	0.463
	20, 30	0.545	0.585		20, 30	0.609	0.549
	20, 40	0.590	0.632		20, 40	0.655	0.600
	30, 15	0.466	0.475		30, 15	0.524	0.419
	30, 30	0.647	0.673		30, 30	0.713	0.634
	30, 45	0.729	0.758		30, 45	0.792	0.737
	30, 60	0.773	0.800		30, 60	0.833	0.777

Distribution	n_1, n_2	App. Power	Simulated Power	Distribution	n_1, n_2	App. Power	Simulated Power
B(3,3)	20, 10	0.777	0.628	CN(0.9,3)	20, 10	0.238	0.284
	20, 20	0.939	0.869		20, 20	0.346	0.452
	20, 30	0.973	0.936		20, 30	0.405	0.517
	20, 40	0.984	0.964		20, 40	0.442	0.561
	30, 15	0.935	0.871		30, 15	0.343	0.374
	30, 30	0.993	0.980		30, 30	0.491	0.598
	30, 45	0.998	0.995		30, 45	0.567	0.700
	30, 60	0.999	0.999		30, 60	0.612	0.719
U(0,1)	20, 10	0.916	0.740	CN(0.8,3)	20, 10	0.260	0.223
	20, 20	0.992	0.950		20, 20	0.379	0.396
	20, 30	0.998	0.985		20, 30	0.444	0.467
	20, 40	0.999	0.995		20, 40	0.484	0.520
	30, 15	0.991	0.941		30, 15	0.376	0.354
	30, 30	1.0	0.996		30, 30	0.535	0.549
	30, 45	1.0	1.0		30, 45	0.614	0.650
	30, 60	1.0	1.0		30, 60	0.661	0.706

The results show that, in general, the approximate power levels and the simulated power levels are close to each other. They become closer as the samples sizes increase. The approximate power levels are usually slightly larger than the simulated power levels for symmetric and nearly symmetric distributions with moderate to light tails. They are, however, slightly smaller than the simulated power levels for symmetric distributions with heavy tails or for highly skewed distributions. The difference between the two power functions is usually not important, except in the case where the samples are generated from the t distribution with 5 degrees of freedom.

Overall, when the minimum sample size reaches 20, the approximate power levels and the simulated power levels are remarkably close. Therefore, the planning of sample sizes can be based upon the approximate power functions.

© 2015, 2017 Minitab Inc. All rights reserved.

Minitab®, Quality. Analysis. Results.® and the Minitab® logo are all registered trademarks of Minitab, Inc., in the United States and other countries. See minitab.com/legal/trademarks for more information.