MINITAB ASSISTANT WHITE PAPER

This paper explains the research conducted by Minitab statisticians to develop the methods and data checks used in the Assistant in Minitab Statistical Software.

# One-Way ANOVA

## Overview

One-way ANOVA is used to compare the means of three or more groups to determine whether they differ significantly from one another. Another important function is to estimate the differences between specific groups.

The most common method to detect differences among groups in one-way ANOVA is the F-test, which is based on the assumption that the populations for all samples share a common, but unknown, standard deviation. We recognized, in practice, that samples often have different standard deviations. Therefore, we wanted to investigate the Welch method, an alternative to the F-test, which can handle unequal standard deviations. We also wanted to develop a method to calculate multiple comparisons that accounts for samples with unequal standard deviations. With this method, we can graph the individual intervals, which provide an easy way to identify groups that differ from one another.

In this paper, we describe how we developed the methods used in the Minitab Assistant One-Way ANOVA procedure for:

- Welch test
- Multiple comparison intervals

Additionally, we examine conditions that can affect the validity of the one-way ANOVA results, including the presence of unusual data, the sample size and power of the test, and the normality of the data. Based on these conditions, the Assistant automatically performs the following checks on your data and reports the findings in the Report Card:

- Unusual data
- Sample size
- Normality of data

In this paper, we investigate how these conditions relate to one-way ANOVA in practice and we describe how we established the guidelines to check for these conditions in the Assistant.

# One-way ANOVA methods

## The F-test versus the Welch test

The F-test commonly used in one-way ANOVA is based on the assumption that all of the groups share a common, but unknown, standard deviation (σ). In practice, this assumption rarely holds true, which leads to problems controlling the Type I error rate. Type I error is the probability of incorrectly rejecting the null hypothesis (concluding the samples are significantly different when they are not). When the samples have different standard deviations, there is a greater likelihood that the test will reach an incorrect conclusion. To address this problem, the Welch test was developed as an alternative to the F-test (Welch, 1951).

### Objective

We wanted to determine whether to use the F-test or the Welch test for the One-Way ANOVA procedure in the Assistant. To do this, we needed to evaluate how closely the actual test results for the F-test and the Welch test matched the target level of significance (alpha, or Type I error rate) for the test; that is, whether the test incorrectly rejected the null hypothesis more often or less often than intended given different sample sizes and sample standard deviations.

### Method

To compare the F-test and the Welch test, we performed multiple simulations, varying the number of samples, the sample size, and the sample standard deviation. For each condition, we performed 10,000 ANOVA tests using both the F-test and the Welch method. We generated random data so that the means of the samples were the same and thus, for each test, the null hypothesis was true. Then, we performed the tests using target significance levels of 0.05 and 0.01. We counted the number of times out of 10,000 tests the F-test and Welch tests actually rejected the null hypothesis, and compared this proportion to the target significance level. If the test performs well, the estimated Type I error should be very close to the target significance level.

### Results

We found that Welch method performed as well as or better than the F-test under all of the conditions we tested. For example, when comparing 5 samples using the Welch test, the Type I error rates were between 0.0460 and 0.0540, very close to the target significance level of 0.05. This indicates that Type I error rate for the Welch method matches the target value even when sample size and standard deviation varies across samples.

On the other hand, the Type I error rates for the F-test were between 0.0273 and 0.2277. In particular, the F-test did poorly under the following conditions:

- The Type I error rates fell below 0.05 when the largest sample also had the largest standard deviation. This condition results in a more conservative test and demonstrates that simply increasing the sample size is not a viable solution when the standard deviations for the samples are not equal.
- The Type I error rates were above 0.05 when the sample sizes were equal but standard deviations were different. The rates were also greater than 0.05 when the sample with a larger standard deviation was of a smaller size than the other samples. In particular, when smaller samples have larger standard deviations, there is a substantial increase in the risk that this test incorrectly rejects the null hypothesis.

For more information on the simulation methodology and results, see Appendix A.

Because the Welch method performed well when the standard deviations and sizes of the samples were unequal, we use the Welch method for the One-way ANOVA procedure in the Assistant.

# Comparison intervals

When an ANOVA test is statistically significant, indicating that at least one of the sample means 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 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. Although a published method for multiple comparisons exists for samples with equal standard deviations, we needed to extend this method to account for samples with unequal standard deviations.

## Objective

We wanted to develop a method to calculate individual comparison intervals that can be used to make comparisons across samples and that also match the test results as closely as possible. We also wanted to provide a visual method for determining which samples are statistically different from the others.

## Method

Standard multiple comparison methods (Hsu 1996) provide an interval for the difference between each pair of means while controlling for the increased error that occurs when making multiple comparisons. In the special case of equal sample sizes and under the assumption of equal standard deviations, it is possible to display individual intervals for each mean in a way that corresponds exactly to the intervals for the differences of all the pairs. For the case of unequal sample sizes, with the assumption of equal standard deviations, Hochberg, Weiss, and Hart (1982) developed individual intervals that are approximately equivalent to the intervals for differences among pairs, based on the Tukey-Kramer method of multiple comparisons. In the Assistant, we apply the same approach to the Games-Howell method of multiple comparisons, which does not assume equal standard deviations. The approach used in the Assistant in release 16 of Minitab was similar in concept, but was not based directly on the Games-Howell approach. For more details, see Appendix B.

## Results

The Assistant displays the comparison intervals in the Means Comparison Chart in the One-Way ANOVA Summary Report. When the ANOVA test is statistically significant, any comparison interval that does not overlap with at least one other interval is marked in red. It is possible for the test and the comparison intervals to disagree, although this outcome is rare because both methods have the same probability of rejecting the null hypothesis when it is true. If the ANOVA test is significant yet all of the intervals overlap, then the pair with the smallest amount of overlap is marked in red. If the ANOVA test is not statistically significant, then none of the intervals are marked in red, even if some of the intervals do not overlap.

# 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.

### Objective

We wanted to develop a method to check for data values that are very large or very small relative to the overall sample, which 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) 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. |

# Sample size

Power is an important property of any hypothesis test because it indicates the likelihood that you will find a significant effect or difference when one truly exists. Power is the probability that you will reject the null hypothesis in favor of the alternative hypothesis. Often, the easiest way to increase the power of a test is to increase the sample size. In the Assistant, for tests with low power, we indicate how large your sample needs to be to find the difference you specified. If no difference is specified, we report the difference you could detect with adequate power. To provide this information, we needed to develop a method for calculating power because the Assistant uses the Welch method, which does not have an exact formula for power.

## Objective

To develop a methodology for calculating power, we needed to address two questions. First, the Assistant does not require that users enter a full set of means; it only requires that they enter a difference between means that has practical implications. For any given difference, there are an infinite number of possible configurations of means that could produce that difference. Therefore, we needed to develop a reasonable approach to determine which means to use when calculating power, given that we could not calculate power for all possible configurations of means. Second, we needed to develop a method to calculate power because the Assistant uses the Welch method, which does not require equal sample sizes or standard deviations.

## Method

To address the infinite number of possible configurations of means, we developed a method based on the approach used in the standard one-way ANOVA procedure in Minitab (**Stat > ANOVA > One-Way**). We focused on the cases where only two of the means differ by the stated amount and the other means are equal (set to the weighted average of the means). Because we assume that only two means differ from the overall mean (and not more than two), the approach provides a conservative estimate of power. However, because the samples may have different sizes or standard deviations, the power calculation still depends on which two means are assumed to differ.

To solve this problem, we identify the two pairs of means that represent the best and worst cases. The worst case occurs when the sample size is small relative to the sample variance, and power is minimized; the best case occurs when the sample size is large relative to the sample variance, and power is maximized. All of the power calculations consider these two extreme cases, which minimize and maximize the power under the assumption that exactly two means differ from the overall weighted average of means.

To develop the power calculation, we used a method shown in Kulinskaya et al. (2003). We compared the power calculations from our simulation, the method we developed to address the configuration of means and the method shown in Kulinskaya et al. (2003). We also examined another power approximation that shows more clearly how power depends on the configuration of means. For more information on the power calculation, see Appendix C.

## Results

Our comparison of these methods showed that the Kulinskaya method provides a good approximation of power and that our method for handling the configuration of means is appropriate.

When the data does not provide enough evidence against the null hypothesis, the Assistant calculates practical differences that can be detected with an 80% and a 90% probability for the given sample sizes. In addition, if you specify a practical difference, the Assistant calculates the minimum and maximum power values for this difference. When the power values are below 90%, the Assistant calculates a sample size based on the specified difference and the observed sample standard deviations. To ensure that the sample size results in both the minimum and maximum power values being 90% or greater, we assume that the specified difference is between the two means with the greatest variability.

If the user does not specify a difference, the Assistant finds the largest difference at which the maximum of the range of power values is 60%. This value is labeled at the boundary between the red and yellow bars on the Power Report, corresponding to 60% power. We also find the smallest difference at which the minimum of the range of power values is 90%. This value is labeled at the boundary between the yellow and green bars on the Power Report, corresponding to 90% power.

When checking for power and sample size, the Assistant displays the following status indicators in the Report Card:

| Status | Condition |
|---|---|
| ℹ | The data does not provide sufficient evidence to conclude that there are differences among the means. No difference was specified. |
| ✓ | The test finds a difference between the means, so power is not an issue.<br>OR<br>Power is sufficient. The test did not find a difference between the means, 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 means, 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 means, 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. |
| ✖ | Power is not sufficient. The test did not find a difference between the means, 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. |

# Normality

A common assumption in many statistical methods is that the data are normally distributed. Fortunately, even when data are not normally distributed, methods based on the normality assumption can work well. This is in part explained by the central limit theorem, which says that the distribution of any sample mean has an approximate normal distribution and that the approximation becomes almost normal as the sample size gets larger.

## Objective

Our objective was to determine how large the sample needs to be to give a reasonably good approximation of the normal distribution. We wanted to examine the Welch test and comparison intervals with samples of small to moderate size with various nonnormal distributions. We wanted to determine how closely the actual test results for Welch method and the comparison intervals matched the chosen level of significance (alpha, or Type I error rate) for the test; that is, whether the test incorrectly rejected the null hypothesis more often or less often than expected given different sample sizes, numbers of levels, and nonnormal distributions.

## Method

To estimate the Type I error, we performed multiple simulations, varying the number of samples, sample size, and the distribution of the data. The simulations included skewed and heavy-tailed distributions that depart substantially from the normal distribution. The size and standard deviation were constant across samples within each test.

For each condition, we performed 10,000 ANOVA tests using the Welch method and the comparison intervals. We generated random data so that the means of the samples were the same and thus, for each test, the null hypothesis was true. Then, we performed the tests using a target significance level of 0.05. We counted the number of times out of 10,000 when the tests actually rejected the null hypothesis, and compared this proportion to the target significance level. For the comparison intervals, we counted the number of times out of 10,000 when the intervals indicated one or more difference. If the test performs well, the Type I error should be very close to the target significance level.

## Results

Overall, the tests and the comparison intervals perform very well across all conditions with sample sizes as small as 10 or 15. For tests with 9 or fewer levels, in almost every case, the results are all within 3 percentage points of the target significance level for a sample size of 10 and within 2 percentage points for a sample size of 15. For tests that have 10 or more levels, in most cases the results are within 3 percentage points with a sample size of 15 and within 2 percentage points with a sample size of 20. For more information, see Appendix D.

Because the tests perform well with relatively small samples, the Assistant does not test the data for normality. Instead, the Assistant checks the size of the samples and indicates when the samples are less than 15 for 2-9 levels and less than 20 for 10-12 levels. Based on these results, the Assistant displays the following status indicators in the Report Card:

| Status | Condition |
|---|---|
| ✓ | The sample sizes are at least 15 or 20, so normality is not an issue. |
| ⚠ | Because some sample sizes are less than 15 or 20, normality may be an issue. |

# References

Dunnet, C. W. (1980). Pairwise Multiple Comparisons in the Unequal Variance Case. *Journal of the American Statistical Association, 75, 796-800*.

Hoaglin, D. C., Iglewicz, B., and Tukey, J. W. (1986). Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association, 81, 991-999*.

Hochberg, Y., Weiss G., and Hart, S. (1982). On graphical procedures for multiple comparisons. *Journal of the American Statistical Association, 77, 767-772*.

Hsu, J. (1996). *Multiple comparisons: Theory and methods*. Boca Raton, FL: Chapman & Hall.

Kulinskaya, E., Staudte, R. G., and Gao, H. (2003). Power approximations in testing for unequal means in a One-Way ANOVA weighted for unequal variances, *Communication in Statistics, 32 (12), 2353-2371*.

Welch, B.L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika, 34, 28-35*

Welch, B.L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika 38, 330-336*.

# Appendix A: The F-test versus the Welch test

The F-test can result in an increase of the Type I error rate when the assumption of equal standard deviations is violated; the Welch test is designed to avoid these problems.

## Welch test

Random samples of sizes $n_1, \ldots, n_k$ from k populations are observed. Let $\mu_1,\ldots,\mu_k$ denote the population means and let $\sigma_1^2, \ldots, \sigma_k^2$ denote the population variances. Let $\bar{x}_1, \ldots, \bar{x}_k$ denote the sample means and let $s_1^2, \ldots, s_k^2$ denote the sample variances. We are interested in testing the hypotheses:

$H_0$: $\mu_1 = \mu_2 = \cdots = \mu_k$

$H_1$: $\mu_i \neq \mu_j$ for some i, j.

The Welch test for testing the equality of k means compares the statistic

$$W^* = \frac{\sum_{j=1}^{k} w_j\left(\bar{x}_j - \hat{\mu}\right)^2/(k-1)}{1+[2(k-2)/(k^2-1)]\sum_{j=1}^{k} h_j}$$

to the F(k – 1, f) distribution, where

$$w_j = \frac{n_j}{s_j^2},$$

$$W = \sum_{j=1}^{k} w_j ,$$

$$\hat{\mu} = \frac{\sum_{j=1}^{k} w_j \bar{x}_j}{W},$$

$$h_j = \frac{\left(1 - w_j/W\right)^2}{n_j - 1}, \text{ and}$$

$$f = \frac{k^2-1}{3\sum_{j=1}^{k} h_j}.$$

The Welch test rejects the null hypothesis if $W^* \geq F_{k-1,f,1-\alpha}$, the percentile of the F distribution that is exceeded with probability $\alpha$.

## Unequal standard deviations

In this section we demonstrate the sensitivity of the F-test to violations of the assumption of equal standard deviations and compare it to the Welch test.

The results below are for one-way ANOVA tests using 5 samples of N(0, $\sigma^2$). Each row is based on 10,000 simulations using the F-test and the Welch test. We tested two conditions for the standard deviation by increasing the standard deviation of the fifth sample, doubling it and quadrupling it compared to the other samples. We tested three different conditions for the sample size: samples sizes are equal, the fifth sample is greater than the others, and the fifth sample is less than the others.

**Table 1** Type I error rates for simulated F-tests and Welch tests with 5 samples with target significance level α = 0.05

| Standard deviation<br>(**σ**1, **σ**2, **σ**3, **σ**4, **σ**5) | Sample size<br>(n1, n2, n3, n4, n5) | F-test | Welch test |
|---|---|---|---|
| 1, 1, 1, 1, 2 | 10, 10, 10, 10, 20 | .0273 | .0524 |
| 1, 1, 1, 1, 2 | 20, 20, 20, 20, 20 | .0678 | .0462 |
| 1, 1, 1, 1, 2 | 20, 20, 20, 20, 10 | .1258 | .0540 |
| 1, 1, 1, 1, 4 | 10, 10, 10, 10, 20 | .0312 | .0460 |
| 1, 1, 1, 1, 4 | 20, 20, 20, 20, 20 | .1065 | .0533 |
| 1, 1, 1, 1, 4 | 20, 20, 20, 20, 10 | .2277 | .0503 |

When the sample sizes are equal (rows 2 and 5), the probability that the F-test incorrectly rejects the null hypothesis is greater than the target 0.05, and the probability increases when the inequality among standard deviations is greater. The problem is made even worse by decreasing the size of the sample with the largest standard deviation. On the other hand, increasing the size of the sample with the largest standard deviation reduces the probability of rejection. However, increasing the sample size by too much makes the probability of rejection too small, which not only makes the test more conservative than necessary under the null hypothesis, but also adversely affects the power of the test under the alternative hypothesis. Compare these results with the Welch test, which agrees well with the target significance level of 0.05 in every case.

Next we conducted a simulation for cases with k = 7 samples. Each row of the table summarizes 10,000 simulated F-tests. We varied the standard deviations and sizes of the samples. The target significance levels are $\alpha$ = 0.05 and $\alpha$ = 0.01. As above, we see deviations from the target values that can be quite severe. Using a smaller sample size when variability is higher leads to very large Type I error probabilities, while using a larger sample can lead to an extremely conservative test. The results are shown in Table 2 below.

**Table 2** Type I error rates for simulated F-tests with 7 samples

| Standard deviation<br>(**σ**1, **σ**2, **σ**3, **σ**4, **σ**5, **σ**6, **σ**7) | Sample sizes<br>(n1, n2, n3, n4, n5, n6, n7) | Target **α** = 0.05 | Target **α** = 0.01 |
|---|---|---|---|
| 1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9 | 21, 21, 21, 21, 22, 22, 12 | 0.0795 | 0.0233 |
| 1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9 | 20, 21, 21, 21, 21, 24, 12 | 0.0785 | 0.0226 |

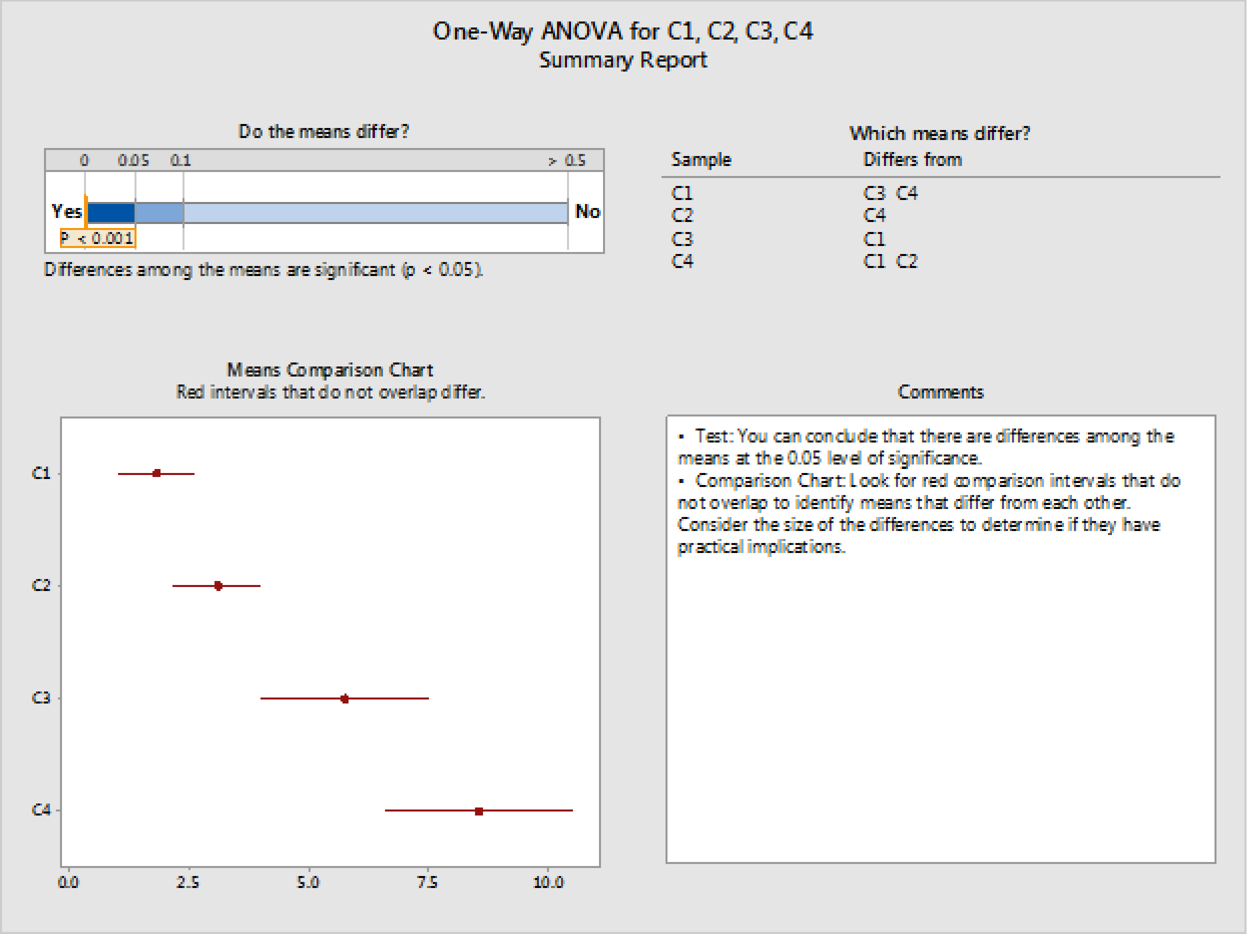
| Standard deviation<br>(**σ**1, **σ**2, **σ**3, **σ**4, **σ**5, **σ**6, **σ**7) | Sample sizes<br>(n1, n2, n3, n4, n5, n6, n7) | Target **α** = 0.05 | Target **α** = 0.01 |
|---|---|---|---|
| 1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9 | 20, 21, 21, 21, 21, 21, 15 | 0.0712 | 0.0199 |
| 1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9 | 20, 20, 20, 21, 21, 23, 15 | 0.0719 | 0.0172 |
| 1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9 | 20, 20, 20, 20, 21, 21, 18 | 0.0632 | 0.0166 |
| 1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9 | 20, 20, 20, 20, 20, 20, 20 | 0.0576 | 0.0138 |
| 1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9 | 18, 19, 19, 20, 20, 20, 24 | 0.0474 | 0.0133 |
| 1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9 | 18, 18, 18, 18, 18, 18, 32 | 0.0314 | 0.0057 |
| 1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9 | 15, 18, 18, 19, 20, 20, 30 | 0.0400 | 0.0085 |
| 1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9 | 12, 18, 18, 18, 19, 19, 36 | 0.0288 | 0.0064 |
| 1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9 | 15, 15, 15, 15, 15, 15, 50 | 0.0163 | 0.0025 |
| 1.85, 1.85, 1.85, 1.85, 1.85, 1.85, 2.9 | 12, 12, 12, 12, 12, 12, 68 | 0.0052 | 0.0002 |
| 1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5 | 21, 21, 21, 21, 22, 22, 12 | 0.1097 | 0.0436 |
| 1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5 | 20, 21, 21, 21, 21, 24, 12 | 0.1119 | 0.0452 |
| 1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5 | 20, 21, 21, 21, 21, 21, 15 | 0.0996 | 0.0376 |
| 1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5 | 20, 20, 20, 21, 21, 23, 15 | 0.0657 | 0.0345 |
| 1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5 | 20, 20, 20, 20, 21, 21, 18 | 0.0779 | 0.0283 |
| 1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5 | 20, 20, 20, 20, 20, 20, 20 | 0.0737 | 0.0264 |
| 1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5 | 18, 19, 19, 20, 20, 20, 24 | 0.0604 | 0.0204 |
| 1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5 | 18, 18, 18, 18, 18, 18, 32 | 0.0368 | 0.0122 |
| 1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5 | 15, 18, 18, 19, 20, 20, 30 | 0.0390 | 0.0117 |
| 1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5 | 12, 18, 18, 18, 19, 19, 36 | 0.0232 | 0.0046 |
| 1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5 | 15, 15, 15, 15, 15, 15, 50 | 0.0124 | 0.0026 |
| 1.75, 1.75, 1.75, 1.75, 1.75, 1.75, 3.5 | 12, 12, 12, 12, 12, 12, 68 | 0.0027 | 0.0004 |
| 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9 | 21, 21, 21, 21, 22, 22, 12 | 0.1340 | 0.0630 |
| 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9 | 20, 21, 21, 21, 21, 24, 12 | 0.1329 | 0.0654 |

| Standard deviation<br>(σ1, σ2, σ3, σ4, σ5, σ6, σ7) | Sample sizes<br>(n1, n2, n3, n4, n5, n6, n7) | Target α = 0.05 | Target α = 0.01 |
|---|---|---|---|
| 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9 | 20, 21, 21, 21, 21, 21, 15 | 0.1101 | 0.0484 |
| 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9 | 20, 20, 20, 21, 21, 23, 15 | 0.1121 | 0.0495 |
| 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9 | 20, 20, 20, 20, 21, 21, 18 | 0.0876 | 0.0374 |
| 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9 | 20, 20, 20, 20, 20, 20, 20 | 0.0808 | 0.0317 |
| 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9 | 18, 19, 19, 20, 20, 20, 24 | 0.0606 | 0.0243 |
| 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9 | 18, 18, 18, 18, 18, 18, 32 | 0.0356 | 0.0119 |
| 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9 | 15, 18, 18, 19, 20, 20, 30 | 0.0412 | 0.0134 |
| 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9 | 12, 18, 18, 18, 19, 19, 36 | 0.0261 | 0.0068 |
| 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9 | 15, 15, 15, 15, 15, 15, 50 | 0.0100 | 0.0023 |
| 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 1.68333, 3.9 | 12, 12, 12, 12, 12, 12, 68 | 0.0017 | 0.0003 |
| 1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7 | 21, 21, 21, 21, 22, 22, 12 | 0.1773 | 0.1006 |
| 1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7 | 20, 21, 21, 21, 21, 24, 12 | 0.1811 | 0.1040 |
| 1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7 | 20, 21, 21, 21, 21, 21, 15 | 0.1445 | 0.0760 |
| 1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7 | 20, 20, 20, 21, 21, 23, 15 | 0.1448 | 0.0786 |
| 1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7 | 20, 20, 20, 20, 21, 21, 18 | 0.1164 | 0.0572 |
| 1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7 | 20, 20, 20, 20, 20, 20, 20 | 0.1020 | 0.0503 |
| 1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7 | 18, 19, 19, 20, 20, 20, 24 | 0.0834 | 0.0369 |
| 1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7 | 18, 18, 18, 18, 18, 18, 32 | 0.0425 | 0.0159 |
| 1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7 | 15, 18, 18, 19, 20, 20, 30 | 0.0463 | 0.0168 |
| 1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7 | 12, 18, 18, 18, 19, 19, 36 | 0.0305 | 0.0103 |
| 1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7 | 15, 15, 15, 15, 15, 15, 50 | 0.0082 | 0.0021 |

| Standard deviation<br>(**σ**1, **σ**2, **σ**3, **σ**4, **σ**5, **σ**6, **σ**7) | Sample sizes<br>(n1, n2, n3, n4, n5, n6, n7) | Target **α** =<br>0.05 | Target **α** =<br>0.01 |
|---|---|---|---|
| 1.55, 1.55, 1.55, 1.55, 1.55, 1.55, 4.7 | 12, 12, 12, 12, 12, 12, 68 | 0.0013 | 0.0001 |

# Appendix B: Comparison intervals

The means comparison chart allows you to evaluate the statistical significance of differences among the population means.



One-Way ANOVA for C1, C2, C3, C4
Summary Report

Do the means differ?

Differences among the means are significant (p < 0.05).

Which means differ?

| Sample | Differs from |
|---|---|
| C1 | C3 C4 |
| C2 | C4 |
| C3 | C1 |
| C4 | C1 C2 |

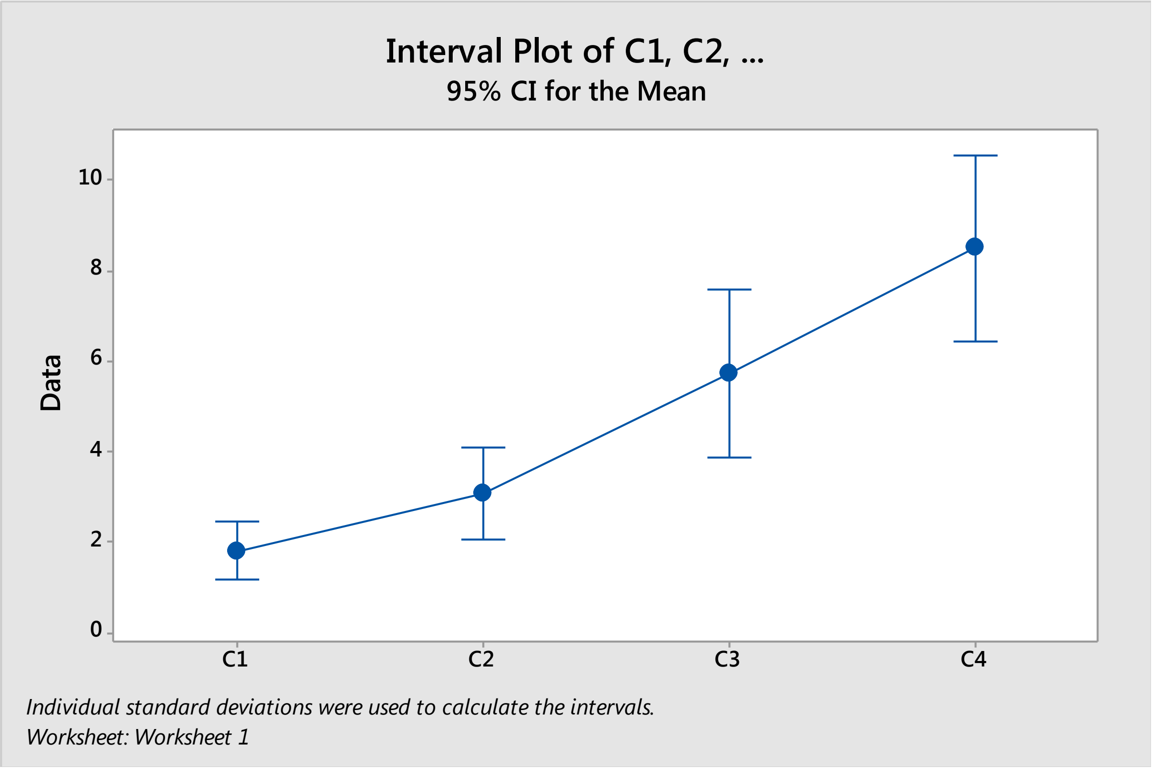
Means Comparison Chart
Red intervals that do not overlap differ.

Comments

- Test: You can conclude that there are differences among the means at the 0.05 level of significance.
- Comparison Chart: Look for red comparison intervals that do not overlap to identify means that differ from each other. Consider the size of the differences to determine if they have practical implications.

**Figure 1** The Means Comparison Chart in the Assistant One-Way ANOVA Summary Report

A similar set of intervals appears in the output for the standard one-way ANOVA procedure in Minitab (**Stat** > **ANOVA** > **One-Way**):



However, note that the intervals above are simply individual confidence intervals for the means. When the ANOVA test (either F or Welch) concludes that some means are different, there is a natural tendency to look for intervals that do not overlap and draw conclusions about which means differ. This informal analysis of the individual confidence intervals will often lead to reasonable conclusions, but it does not control for the probability of error the same way the ANOVA test does. Depending on the number of populations, the intervals may be substantially more or less likely than the test to conclude that there are differences. As a result, the two methods can easily reach inconsistent conclusions. The comparison chart is designed to more consistently match the Welch test results when making multiple comparisons, although it is not always possible to achieve complete consistency.

Multiple comparison methods, such as such as the Tukey-Kramer and Games-Howell comparisons in Minitab (**Stat** > **ANOVA** > **One-Way**), allow you to draw statistically valid conclusions about differences among the individual means. These two methods are pairwise comparison methods, which provide an interval for the difference between each pair of means. The probability that all intervals simultaneously contain the differences they are estimating is at least $1 - \alpha$. The Tukey-Kramer method depends on the assumption of equal variances, while the Games-Howell method does not require equal variances. If the null hypothesis of equal means is true, then all the differences are zero, and the probability that any of the Games-Howell intervals will fail to contain zero is at most $\alpha$. So we can use the intervals to perform a hypothesis test with significance level $\alpha$. We use Games-Howell intervals as the starting point for deriving the comparison chart intervals in the Assistant.

Given a set of intervals [$L_{ij}$, $U_{ij}$] for all the differences $\mu_i - \mu_j$, $1 \le i < j \le k$, we want to find a set of intervals [$L_i$, $U_i$] for the individual means $\mu_i$, $1 \le i \le k$, that conveys the same information. This requires that any difference d is in the interval [$L_{ij}$, $U_{ij}$] if, and only if, there exists $\mu_i \in [L_i, U_i]$ and $\mu_j \in [L_j, U_j]$ such that $\mu_i - \mu_j = d$. The endpoints of the intervals must be related by the equations

$$U_i - L_j = U_{ij} \; and$$

$$L_i - U_j = L_{ij}.$$

For k = 2, we have only one difference but two individual intervals, so it is possible to obtain exact comparison intervals. In fact, there is quite a bit of flexibility in the width of the intervals that satisfy this condition. For k = 3, there are three differences and three individual intervals, so again it is possible to satisfy the condition, but now without the flexibility in setting the width of the intervals. For k = 4, there are six differences but only four individual intervals. Comparison intervals must try to convey the same information using fewer intervals. In general, for k ≥ 4, there are more differences than individual means, so there is not an exact solution unless additional conditions are imposed on the intervals for differences, such as equal widths.

Tukey-Kramer intervals have equal widths only if all the sample sizes are the same. The equal widths are also a consequence of assuming equal variances. Games-Howell intervals do not assume equal variances, and so do not have equal widths. In the Assistant, we will have to rely on approximate methods to define comparison intervals.

The Games-Howell interval for $\mu_i - \mu_j$ is

$$\bar{x}_i - \bar{x}_j \pm |q^*(k, \hat{v}_{ij})| \sqrt{s_i^2/n_i + s_j^2/n_j}$$

where $q^*(k, \hat{v}_{ij})$ is the appropriate percentile of the studentized range distribution, which depends on k, the number of means being compared, and on

$\nu_{ij}$, the degrees of freedom associated with the pair (i, j):

$$\hat{v}_{ij} = \frac{\left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}\right)^2}{\left(\frac{s_i^2}{n_i}\right)^2 \frac{1}{n_i - 1} + \left(\frac{s_j^2}{n_j}\right)^2 \frac{1}{n_j - 1}}.$$

Hochberg, Weiss, and Hart (1982) obtained individual intervals that are approximately equivalent to these pairwise comparisons by using:

$$\bar{x}_i \pm |q^*(k, \nu)| s_p X_i.$$

The values $X_i$ are selected to minimize

$$\sum \sum_{i \neq j} (X_i + X_j - a_{ij})^2,$$

Where:

$$a_{ij} = \sqrt{1/n_i + 1/n_j}\ .$$

We adapt this approach to the case of unequal variances by deriving intervals from Games-Howell comparisons of the form

$$\bar{x}_i \pm d_i.$$

The values $d_i$ are selected to minimize

$$\sum\sum_{i \neq j}\left(d_i + d_j - b_{ij}\right)^2,$$

Where:

$$b_{ij} = \left|q^*(k, \hat{v}_{ij})\right|\sqrt{s_i^2/n_i + s_j^2/n_j}\ .$$

The solution is

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

The graphs below compare simulation results for the Welch test with the results for comparison intervals using two methods: the Games-Howell based method we use now, and the method used in release 16 of Minitab based on an averaging of degrees of freedom. The vertical axis is the proportion of times out of 10,000 simulations that the Welch test incorrectly rejects the null hypotheses or that not all the comparison intervals overlap. The target alpha is $\alpha = 0.05$ in these examples. These simulations cover various cases of unequal standard deviations and sample sizes; each position along the horizontal axis represents a different case.

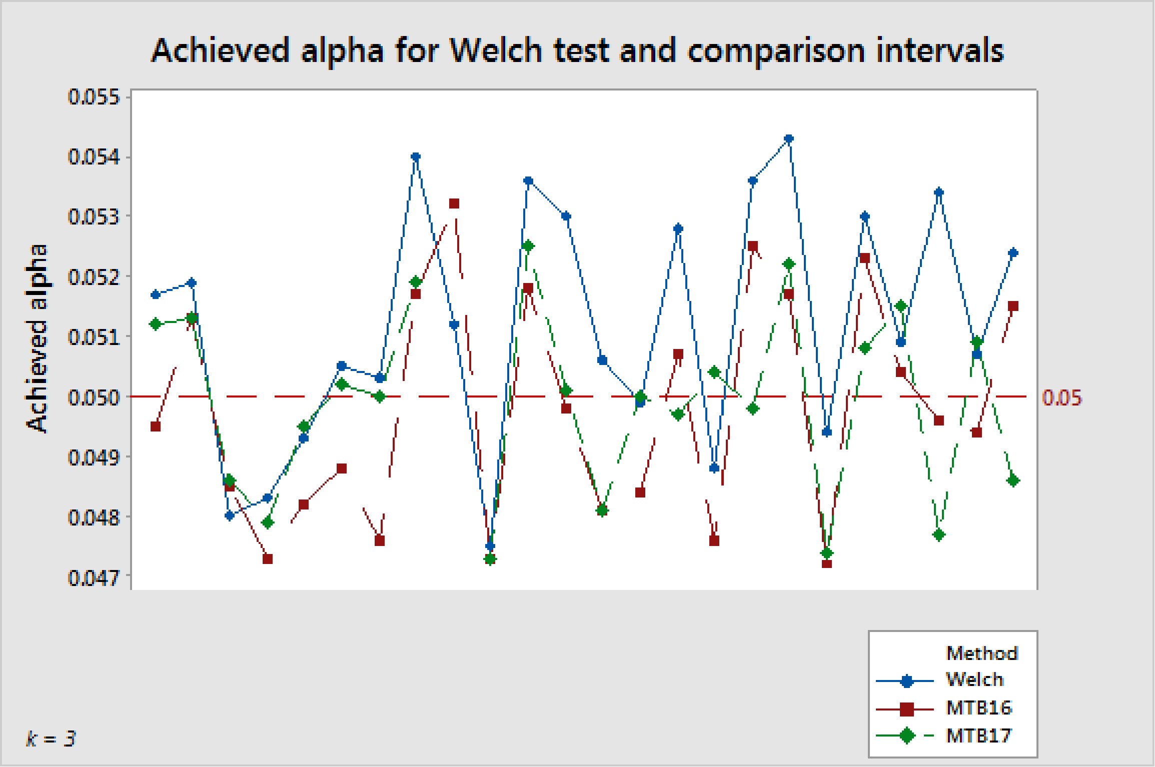


**Figure 2** Welch test compared with two methods of calculating comparison intervals for 3 samples

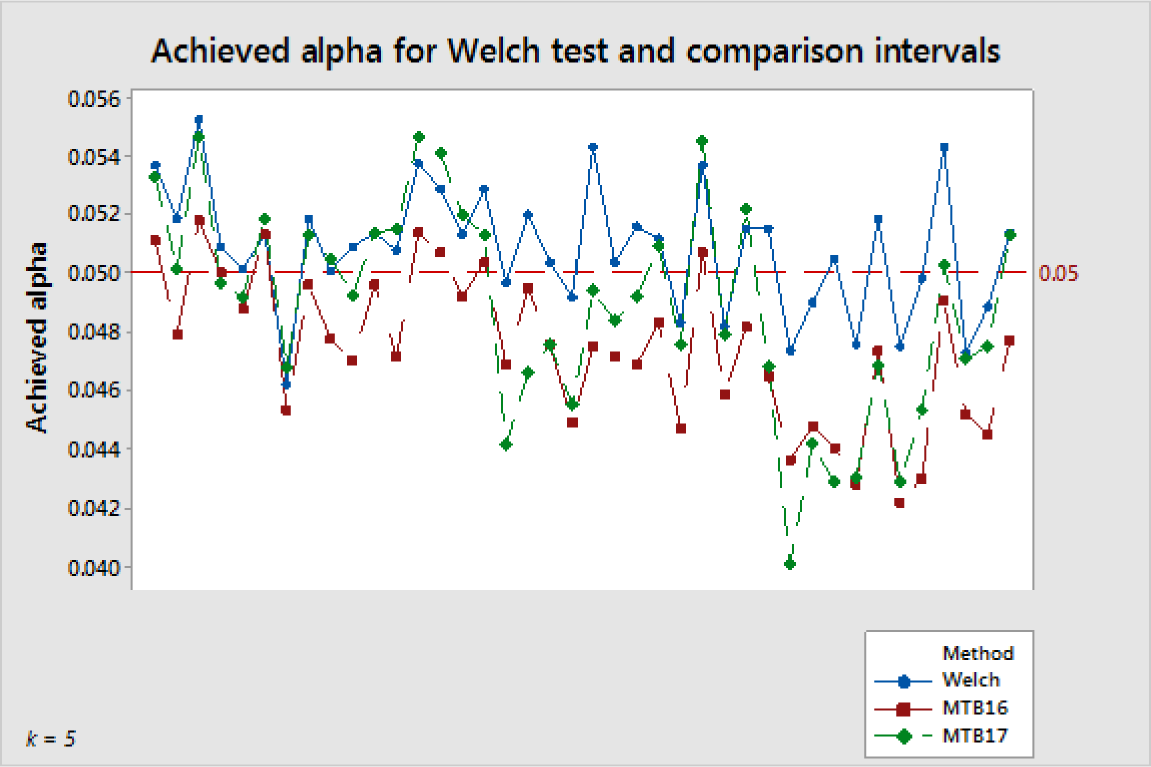


**Figure 3** Welch test compared with two methods of calculating comparison intervals for 5 samples

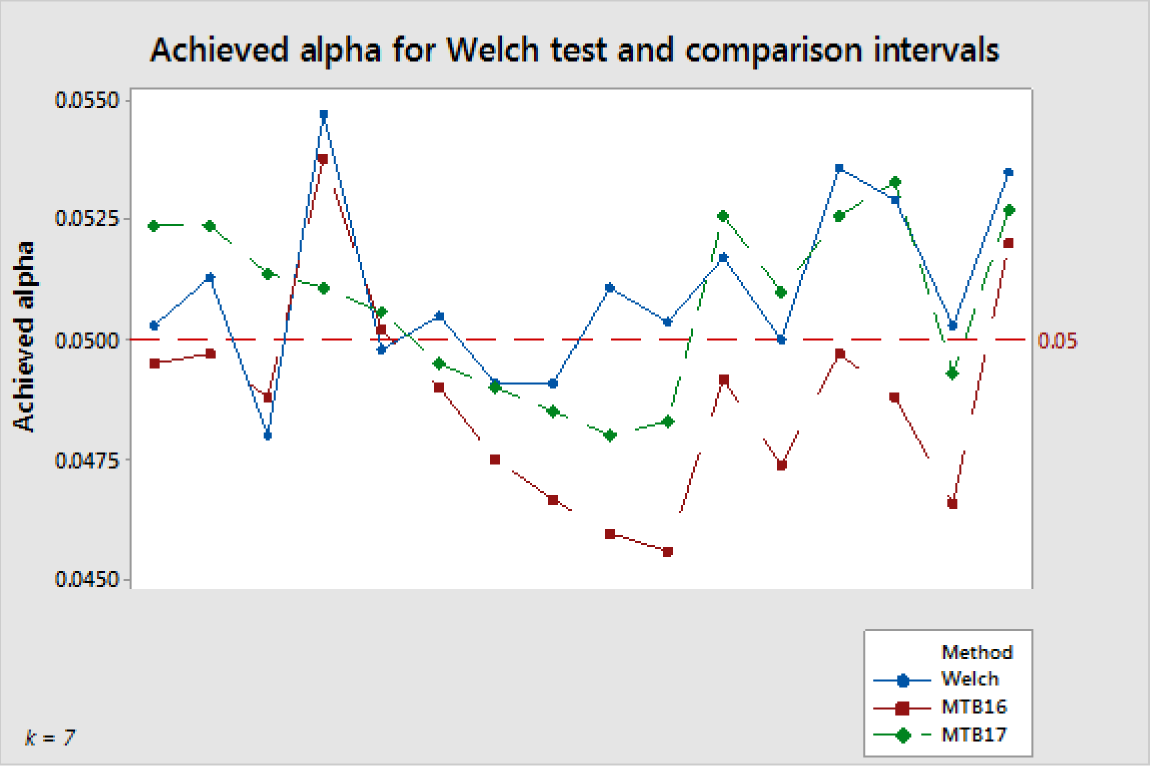


**Figure 4** Welch test compared with two methods of calculating comparison intervals for 7 samples

These results show simulated alpha values in a narrow range around the target value of 0.05. Also, the results using the Games-Howell-based method implemented in release 17 of Minitab are arguably more closely aligned with the results for the Welch test than was the method used in release 16 of Minitab.

There is evidence that the coverage probability of intervals may be sensitive to unequal standard deviations. But the sensitivity is not nearly as extreme as that of the F-test. The graph below illustrates this dependence in the case of k = 5.

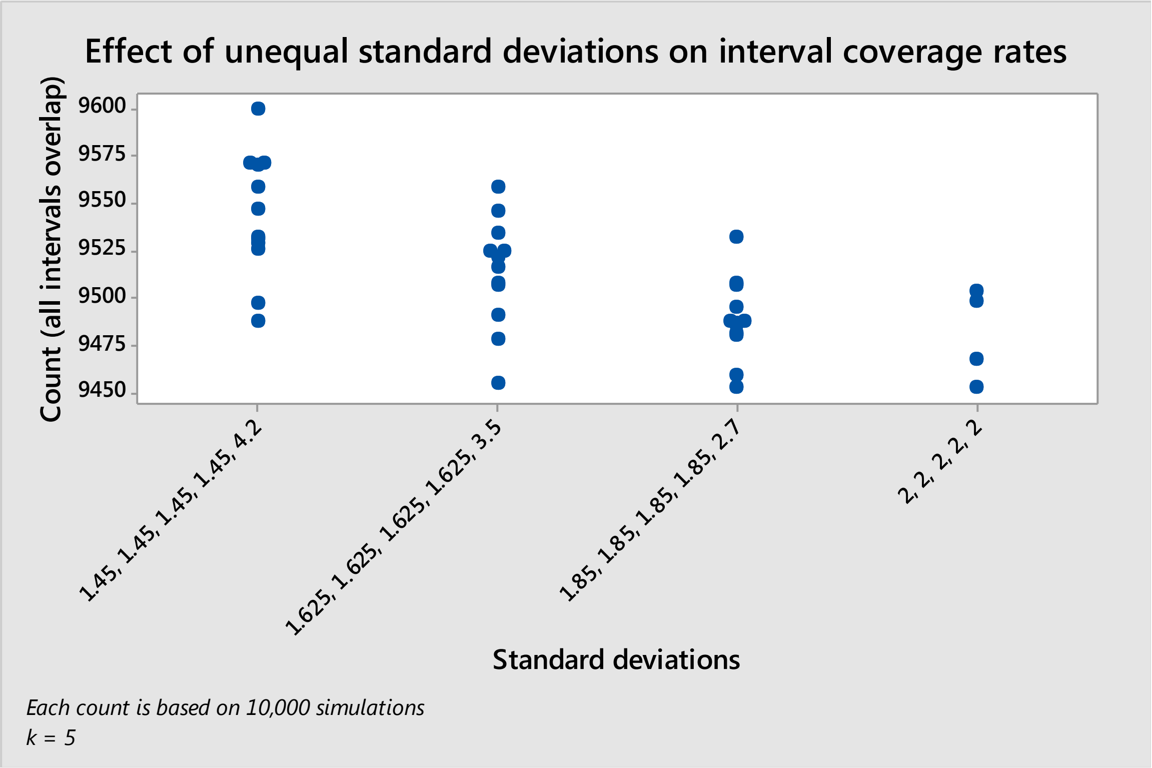
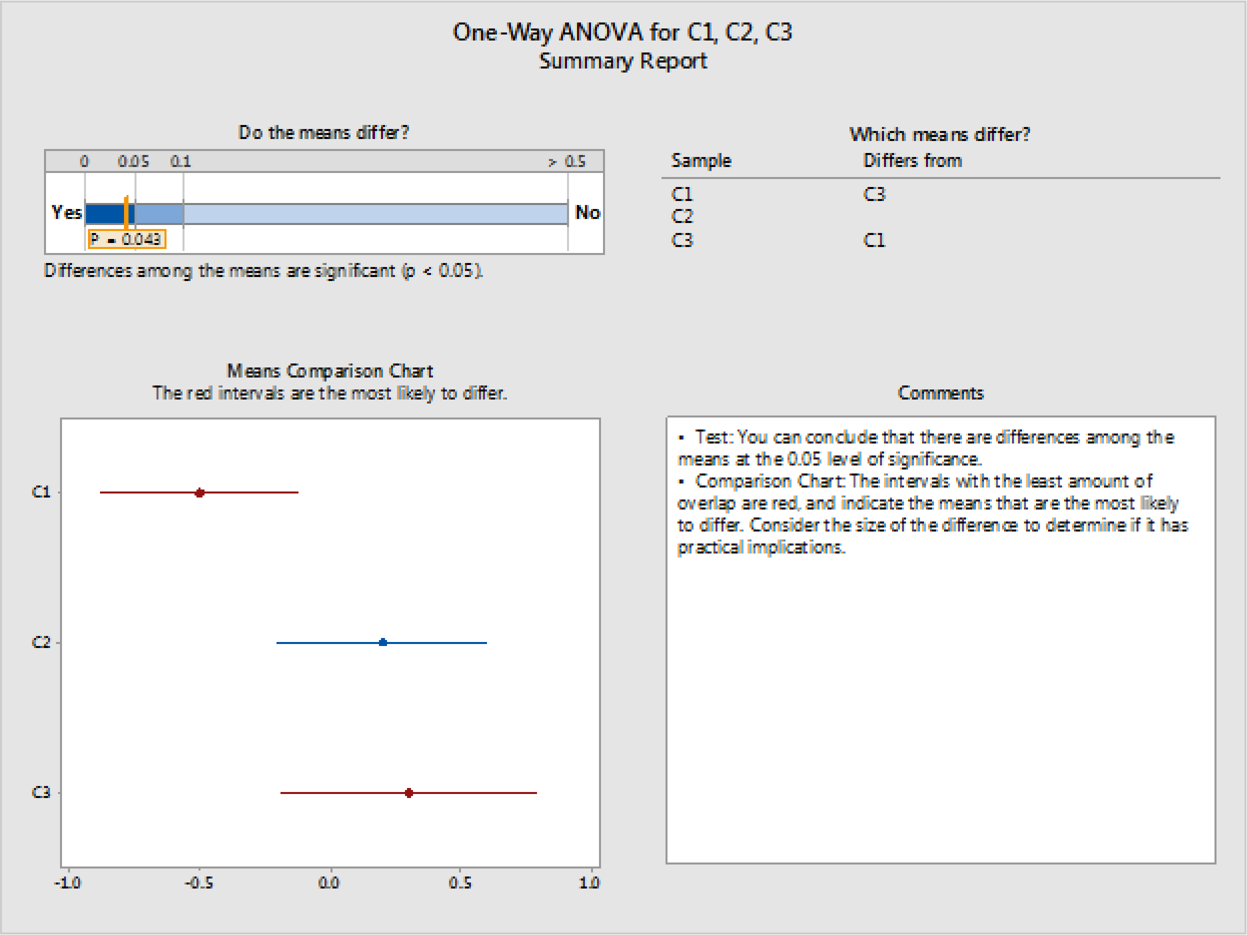


**Figure 5** Results of simulation with unequal standard deviations

# Using the hypothesis test and comparison intervals together

In rare cases, it is possible that the hypothesis test and the comparison will not agree about rejecting the null hypothesis. The test can reject the null hypothesis while the comparison intervals all still overlap. Conversely, the test can fail to reject the null hypothesis, while there are intervals that do not overlap. These disagreements are rare because both methods have the same probability of rejecting the null hypothesis when it is true.

When this happens, we first consider the test results and use the comparisons to investigate further in the event of a significant test. If the test rejects the null hypothesis at significance level $\alpha$, then any comparison interval that fails to overlap with at least one other is marked in red. This is used as a visual cue that the corresponding group mean differs from at least one other. Even if all the intervals overlap, the pair with the smallest amount of overlap is colored red if the test is significant to indicate the "most likely" difference (see Figure 6 below). This is a somewhat arbitrary choice, especially if there are other pairs that have very little overlap. But no other pair has a bound on its difference that is closer to zero.



One-Way ANOVA for C1, C2, C3
Summary Report

Differences among the means are significant (p < 0.05).

Which means differ?

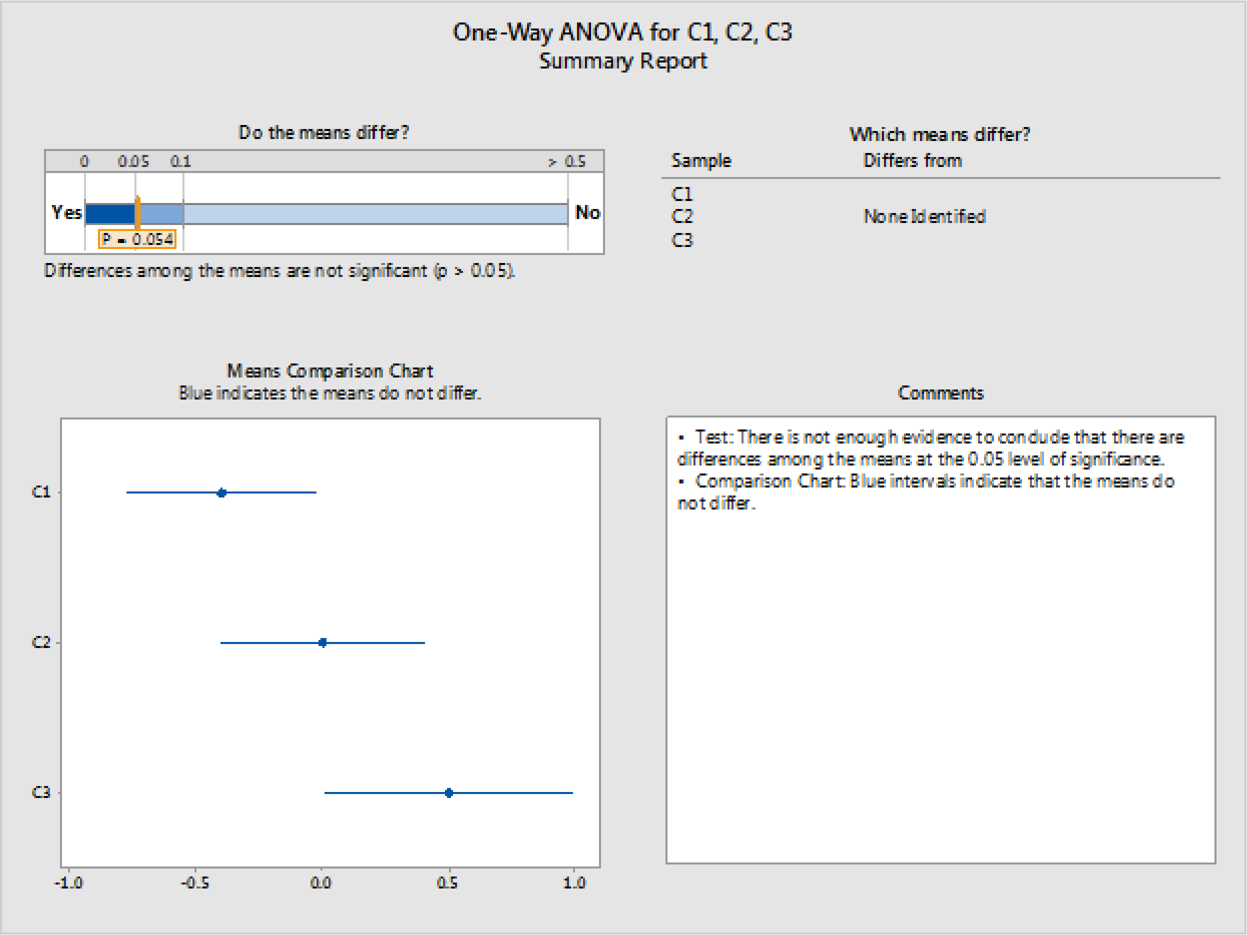
| Sample | Differs from |
|---|---|
| C1 | C3 |
| C2 | |
| C3 | C1 |

Comments

- Test: You can conclude that there are differences among the means at the 0.05 level of significance.
- Comparison Chart: The intervals with the least amount of overlap are red, and indicate the means that are the most likely to differ. Consider the size of the difference to determine if it has practical implications.

**Figure 6** Significant test, intervals marked in red even when they overlap among samples

If the test fails to reject the null hypothesis, then none of the intervals are marked in red, even if there are intervals that do not overlap (see Figure 7 below). Although the intervals imply that there are differences among the means, keep in mind that failure to reject the null hypothesis is not the same as concluding that the null hypothesis is true. It only indicates that the observed differences are not large enough to rule out chance as the cause. It is also worth noting that the gap between the non-overlapping intervals will typically be very small in this situation, so that very small differences are still consistent with the intervals, and do not necessarily indicate that there is a difference with practical implications.



One-Way ANOVA for C1, C2, C3
Summary Report

Do the means differ?

Differences among the means are not significant (p > 0.05).

Which means differ?

| Sample | Differs from |
|---|---|
| C1 | |
| C2 | None Identified |
| C3 | |

Means Comparison Chart
Blue indicates the means do not differ.

Comments

- Test: There is not enough evidence to conclude that there are differences among the means at the 0.05 level of significance.
- Comparison Chart: Blue intervals indicate that the means do not differ.

**Figure 7** Test fails, no intervals marked in red even when there is no overlap among samples

# Appendix C: Sample size

In one-way ANOVA, the parameters being tested are the population means $\mu_1$, $\mu_2$, … $\mu_k$ of the different groups or populations. The parameters satisfy the null hypothesis if they are all equal. If there are any differences among the means, they satisfy the alternative hypothesis. The probability of rejecting the null hypothesis should be no more than $\alpha$ for means that satisfy the null hypothesis. The actual probabilities depend on the standard deviation of the distributions and on the size of the samples. The power to detect any deviation from the null hypothesis increases with smaller standard deviations or larger samples.

We can compute the power of the F-test under the assumption of normal distributions with equal standard deviations using a non-central F distribution. The non-centrality parameter is:

$$\theta_F = \sum_{i=1}^{k} n_i \, (\mu_i - \mu)^2 / \sigma^2$$

where μ is the weighted average of the means:

$$\mu \; = \; \sum_{i=1}^{k} n_i \mu_i \, / \sum_{i=1}^{k} n_i \; ,$$

and σ is the standard deviation, which is assumed to be constant. All other things being equal, the power increases with $\theta_F$. This is the precise sense in which the power increases as the means deviate farther from the null hypothesis.

Unlike the F-test, the Welch test does not have a simple exact formula for power. But we will look at two reasonably good approximate formulas. The first uses a noncentral F-distribution in a way similar to the power of the F-test. The noncentrality parameter will use is still of the form:

$$\theta_W \; = \; \sum_{i=1}^{k} w_i (\mu_i - \mu)^2$$

where μ is the weighted average:

$$\mu \; = \sum_{i=1}^{k} w_i \mu_i / \sum_{j=1}^{k} w_j$$

but the weights will depend on the standard deviations as well as the sample sizes, i.e. $w_i = n_i/\sigma_i^2$ or $w_i = \; n_i/s_i^2$, depending on whether we are simulating the results for known standard deviations $\sigma_i^2$ or estimating the power based on sample standard deviations $s_i^2$. The approximate power is then computed as:

$$P\left(F_{k-1,f,\theta_w} \geq \; F_{k-1,f,1-\alpha}\right)$$

where the denominator degrees of freedom are

$$f \; = \; \frac{k^2 - 1}{3 \sum_{i=1}^{k} \left(1 - w_i / \sum_{j=1}^{k} w_j\right) / (n_i - 1)}.$$

As we show below, this provides reasonably good approximations to the power observed in simulations. And while we use a different approximation to compute the power in the Assistant menu, this one provides good insight, and is the basis for selecting the configuration of means at which we compute the power in the Assistant menu.

# Configuration of means

In keeping with the approach used for power and sample size in Minitab (**Stat** > **ANOVA** > **One-Way**), the Assistant does not ask the user for a full set of means at which to evaluate power. Instead, it asks the user for a difference between means that has practical implications. For a given difference, there are an infinite number of possible configurations of means in which the largest and smallest means differ by that amount. For example, all of the following have a maximum difference of 10 among a set of five means:

$\mu_1 = 0, \mu_2 = 5, \mu_3 = 5, \mu_4 = 5, \mu_5 = 10;$

$\mu_1 = 5, \mu_2 = 0, \mu_3 = 10, \mu_4 = 10, \mu_5 = 0;$

$\mu_1 = 0, \mu_2 = 10, \mu_3 = 0, \mu_4 = 0, \mu_5 = 0;$

and there are infinitely many more.

We follow the approach used for power and sample size in Minitab (**Stat** > **Power and Sample Size** > **One-Way ANOVA**), namely picking a case where all but two of the means are at the (weighted) average of the means, and the remaining two means differ by the stated amount. However, because of the possibility of unequal variances and sample sizes, the non-centrality parameter (and hence the power) still depend on which two means are assumed to differ.

Consider the configuration of means $\mu_1, \dots, \mu_k$ in which all but two of the means are equal to the overall weighted mean $\mu$, and two means, say $\mu_i > \mu_j$, differ from each other and from the overall mean. Let $\Delta = \mu_i - \mu_j$ denote the difference between the two means. Let $\Delta_i = \mu_i - \mu$ and $\Delta_j = \mu - \mu_j$. Hence $\Delta = \Delta_i + \Delta_j$. Also, since $\mu$ represents the weighted mean of all k means, and (k – 2) of the means are assumed to equal $\mu$, we have:

$$\mu = \left[\sum_{l \neq i,j} w_l \mu_l + w_i(\mu + \Delta_i) + w_j(\mu - \Delta_j)\right] \Big/ \sum_{l=1}^{k} w_l = \mu + (w_i\Delta_i - w_j\Delta_j) \Big/ \sum_{l=1}^{k} w_l \,.$$

Hence:

$$w_i\Delta_i = w_j\Delta_j = w_j(\Delta - \Delta_i)\,,$$

and therefore,

$$\Delta_i = \frac{w_j}{w_i + w_j}\Delta$$

$$\Delta_j = \frac{w_i}{w_i + w_j}\Delta$$

For this particular configuration of means, we can compute the noncentrality parameter related to the Welch test:

$$\theta_W = w_i(\mu_i - \mu)^2 + w_j\left(\mu_j - \mu\right)^2$$

$$= \frac{w_i w_j^2 \Delta^2 + w_j w_i^2 \Delta^2}{\left(w_i + w_j\right)^2} = \frac{w_i w_j \Delta^2}{w_i + w_j}$$

This quantity is increasing in $w_i$ for fixed $w_j$ and vice versa. Therefore it is maximized at the pair (i, j) with the two largest weights and minimized at the pair with the two smallest weights. All the power computations consider these two extreme cases, which maximize and minimize the power under the assumption that exactly two means differ from the overall weighted average of means.

If you specify a difference for the test, the minimum and maximum power values are evaluated for this difference. The range of these powers is indicated on the reports relative to a color-coded bar on which powers at or below 60% are in red, powers at or above 90% are in green, and powers between 60% and 90% are in yellow. The Report Card results depend on where the range of powers falls relative to this color-coded scale. If the entire range is in the red, then the power for any pair of groups is less than or equal to 60%, and the red icon appears on the report card to indicate a problem of insufficient power. If the entire range is in the green, the power for any group is at least 90%, and the green icon on the Report Card indicates the condition of sufficient power. All other conditions are treated as intermediate situations indicated by a yellow icon on the Report Card.

In cases where the green condition is not met, the Assistant computes a sample size which would lead to the green condition given the user-specified difference and the observed sample standard deviations. Estimated power depends on the sample sizes via the weights $w_i = n_i/s_i^2$. If all samples are assumed to have the same sample size, then the two smallest weights correspond to the two groups with the largest sample standard deviations. The Assistant finds a sample size that gives power of at least 90% if the specified difference is between the two groups with the greatest variability. Hence, taking a sample size at least this large for all groups would result in the full range of power values being at least 90%, which satisfies the green condition.

If the user does not specify a difference for the power computation, then the Assistant finds the largest difference at which the maximum of the range of computed powers would be 60%. This value is labeled at the boundary between the red and yellow sections of the bar, corresponding to 60% power. It also finds the smallest difference at which the minimum of the range of computed powers would be 90%. This value is labeled at the boundary between the yellow and green sections of the bar, corresponding to 90% power.

# Power calculation

The power is computed using the approximation due to Kulinskaya et al. (2003):

Define:

$\lambda = \sum_{i=1}^{k} w_i(\mu_i - \mu)^2$ ,

$A = \sum_{i=1}^{k} h_i$ ,

$B = \sum_{i=1}^{k} w_i(\mu_i - \mu)^2(1 - w_i/W)/(n_i - 1)$ ,

$D = \sum_{i=1}^{k} w_i^2(\mu_i - \mu)^4/(n_i - 1)$ ,

$E = \sum_{i=1}^{k} w_i^3(\mu_i - \mu)^6/(n_i - 1)^2$.

The first three cumulants of the numerator $\sum_{i=1}^{k} w_i(\bar{x}_i - \hat{\mu})^2$ of the Welch statistic can be estimated as:

$\kappa_1 = k - 1 + \lambda + 2A + 2B$,

$\kappa_2 = 2(k - 1 + 2\lambda + 7A + 14B + D)$,

$\kappa_3 = 8(k - 1 + 3\lambda + 15A + 45B + 6D + 2E)$.

Let $F_{k-1, f, 1-\alpha}$ denote the (1 – α) quantile of the F(k – 1, f) distribution. Recall that W* ≥ $F_{k-1, f, 1-\alpha}$ is the criterion for rejecting the null hypothesis in a size α Welch test.

Let

$q = (k - 1)\left[1 + \frac{2(k-2)A}{k^2 - 1}\right] F_{k-1, f, 1-\alpha}$ ,

$b = \kappa_1 - 2\kappa_2^2/\kappa_3$,

$c = \kappa_3/(4\kappa_2)$ [Note: the expression for c is shown in Kulinskaya et al. (2003) without the parentheses.]

$\nu = 8\kappa_2^3/\kappa_3^2$.

Then the estimated approximate power of the Welch test is:

$$P(\chi_\nu^2 \geq \frac{q - b}{c})$$

where $\chi_\nu^2$ is a chi-square random variable with ν degrees of freedom.

The following results compare the power for the two approximation methods and the simulated power for a range of examples, based on 10,000 simulations.

**Table 3** Power calculations for the two approximation methods compared to simulated power

| Example | alpha | Simulated power | Noncentral F | Kulinskaya et al. |
|---|---|---|---|---|
| **μ’s: 0, 0, 0, -0.1724, 0.8276** | 0.10 | 0.1372 | 0.135702 | 0.135795 |
| σ’s: 2, 2, 2, 2, 4 | 0.05 | 0.0739 | 0.072563 | 0.069512 |
| n’s: 12, 12, 12, 12, 10 | 0.01 | 0.0195 | 0.016587 | 0.012538 |
| **μ’s: 0, 0, 0, -0.3448, 1.6552** | 0.10 | 0.2498 | 0.251064 | 0.257455 |
| σ’s: 2, 2, 2, 2, 4 | 0.05 | 0.1574 | 0.153128 | 0.156215 |
| n’s: 12, 12, 12, 12, 10 | 0.01 | 0.0541 | 0.045211 | 0.042195 |
| **μ’s: 0, 0, 0, -0.5172, 2.4828** | 0.10 | 0.4534 | 0.445570 | 0.453506 |
| σ’s: 2, 2, 2, 2, 4 | 0.05 | 0.3211 | 0.311994 | 0.321575 |
| n’s: 12, 12, 12, 12, 10 | 0.01 | 0.1273 | 0.121225 | 0.125065 |
| **μ’s: 0, 0, 0, -0.6896, 3.3104** | 0.10 | 0.6620 | 0.671317 | 0.670296 |
| σ’s: 2, 2, 2, 2, 4 | 0.05 | 0.5219 | 0.533819 | 0.538617 |
| n’s: 12, 12, 12, 12, 10 | 0.01 | 0.2842 | 0.271316 | 0.282759 |
| **μ’s: 0, 0, 0, -0.8620, 4.1380** | 0.10 | 0.8417 | 0.852589 | 0.846697 |
| σ’s: 2, 2, 2, 2, 4 | 0.05 | 0.7382 | 0.752173 | 0.746121 |
| n’s: 12, 12, 12, 12, 10 | 0.01 | 0.4883 | 0.487601 | 0.493230 |
| **μ’s: 0, 0, 0, -1.0344, 4.9656** | 0.10 | 0.9429 | 0.952077 | 0.954929 |
| σ’s: 2, 2, 2, 2, 4 | 0.05 | 0.8866 | 0.901485 | 0.897937 |
| n’s: 12, 12, 12, 12, 10 | 0.01 | 0.6910 | 0.711055 | 0.703379 |
| **μ’s: 0, 0, 0, 0, 0, -0.148148, 1.85185** | 0.10 | 0.2011 | 0.189392 | 0.200114 |
| σ’s: 2, 2, 2, 2, 2, 2, 5 | 0.05 | 0.1201 | 0.108986 | 0.117420 |
| n’s: 20, 20, 20, 20, 20, 20, 10 | 0.01 | 0.0385 | 0.028986 | 0.031456 |
| **μ’s: 0, 0, 0, 0, 0, -0.296296, 3.70370** | 0.10 | 0.4942 | 0.485917 | 0.500143 |
| σ’s: 2, 2, 2, 2, 2, 2, 5 | 0.05 | 0.3677 | 0.351593 | 0.375296 |
| n’s: 20, 20, 20, 20, 20, 20, 10 | 0.01 | 0.1770 | 0.149041 | 0.177189 |
| **μ’s: 0, 0, 0, 0, 0, -0.444444, 5.55556** | 0.10 | 0.8125 | 0.829702 | 0.819542 |
| σ’s: 2, 2, 2, 2, 2, 2, 5 | 0.05 | 0.7131 | 0.727384 | 0.720807 |
| n’s: 20, 20, 20, 20, 20, 20, 10 | 0.01 | 0.4876 | 0.474291 | 0.494690 |

| Example | alpha | Simulated power | Noncentral F | Kulinskaya et al. |
|---|---|---|---|---|
| **μ's: 0, 0, 0, 0, 0, -0.592593, 7.40741** | 0.10 | 0.9645 | 0.977211 | 0.984213 |
| σ's: 2, 2, 2, 2, 2, 2, 5 | 0.05 | 0.9286 | 0.949997 | 0.949239 |
| n's: 20, 20, 20, 20, 20, 20, 10 | 0.01 | 0.7938 | 0.831174 | 0.814067 |
| **μ's: 0, 0, 0, 0, 0, -0.740741, 9.25926** | 0.10 | 0.9961 | 0.998947 | 1.00000 |
| σ's: 2, 2, 2, 2, 2, 2, 5 | 0.05 | 0.9895 | 0.996653 | 1.00000 |
| n's: 20, 20, 20, 20, 20, 20, 10 | 0.01 | 0.9528 | 0.977536 | 0.98705 |
| **μ's: 0, 0, 0, 0, 0, -0.888889, 11.1111** | 0.10 | 0.9999 | 0.999985 | 1.00000 |
| σ's: 2, 2, 2, 2, 2, 2, 5 | 0.05 | 0.9995 | 0.999926 | 1.00000 |
| n's: 20, 20, 20, 20, 20, 20, 10 | 0.01 | 0.9943 | 0.998910 | 1.00000 |
| **μ's: 0, 0, 0, 0, 0, -0.518519, 6.48148** | 0.10 | 0.9059 | 0.929392 | 0.924696 |
| σ's: 2, 2, 2, 2, 2, 2, 5 | 0.05 | 0.8403 | 0.868721 | 0.856720 |
| n's: 20, 20, 20, 20, 20, 20, 10 | 0.01 | 0.6511 | 0.671210 | 0.666520 |
| **μ's: 0, 0, 0, 0, 0, -.5, .5** | 0.10 | 0.1870 | 0.186658 | 0.183290 |
| σ's: 2, 2, 2, 2, 2, 2, 2 | 0.05 | 0.1098 | 0.106600 | 0.100189 |
| n's: 12, 12, 12, 12, 12, 12, 12 | 0.01 | 0.0315 | 0.027773 | 0.021332 |
| **μ's: 0, 0, 0, 0, 0, -1, 1** | 0.10 | 0.4734 | 0.474736 | 0.472469 |
| σ's: 2, 2, 2, 2, 2, 2, 2 | 0.05 | 0.3394 | 0.338655 | 0.334430 |
| n's: 12, 12, 12, 12, 12, 12, 12 | 0.01 | 0.1378 | 0.137788 | 0.128693 |
| **μ's: 0, 0, 0, 0, 0, -1.5, 1.5** | 0.10 | 0.8228 | 0.817355 | 0.810181 |
| σ's: 2, 2, 2, 2, 2, 2, 2 | 0.05 | 0.7112 | 0.707319 | 0.698461 |
| n's: 12, 12, 12, 12, 12, 12, 12 | 0.01 | 0.4391 | 0.441154 | 0.431868 |
| **μ's: 0, 0, 0, 0, 0, -2, 2** | 0.10 | 0.9691 | 0.973246 | 0.973319 |
| σ's: 2, 2, 2, 2, 2, 2, 2 | 0.05 | 0.9312 | 0.940585 | 0.936546 |
| n's: 12, 12, 12, 12, 12, 12, 12 | 0.01 | 0.7817 | 0.799339 | 0.785099 |
| **μ's: 0, 0, 0, 0, 0, -2.5, 2.5** | 0.10 | 0.9984 | 0.998579 | 0.999763 |
| σ's: 2, 2, 2, 2, 2, 2, 2 | 0.05 | 0.9936 | 0.995330 | 0.997481 |
| n's: 12, 12, 12, 12, 12, 12, 12 | 0.01 | 0.9587 | 0.967674 | 0.966249 |
| **μ's: 0, 0, 0, 0, 0, -3, 3** | 0.10 | 1.0000 | 0.999975 | 1.00000 |
| σ's: 2, 2, 2, 2, 2, 2, 2 | 0.05 | 0.9997 | 0.999870 | 1.00000 |
| n's: 12, 12, 12, 12, 12, 12, 12 | 0.01 | 0.9959 | 0.997927 | 0.99961 |

| Example | alpha | Simulated power | Noncentral F | Kulinskaya et al. |
|---|---|---|---|---|
| **μ's: 0, 0, 0, 0, 0, -3.5, 3.5** | 0.10 | 1.00000 | 1.00000 | 1.00000 |
| σ's: 2, 2, 2, 2, 2, 2, 2 | 0.05 | 1.00000 | 1.00000 | 1.00000 |
| n's: 12, 12, 12, 12, 12, 12, 12 | 0.01 | 0.99998 | 0.99995 | 1.00000 |
| **μ's: 0, 0, 0, 0, 0, -1.75, 1.75** | 0.10 | 0.9140 | 0.921225 | 0.916652 |
| σ's: 2, 2, 2, 2, 2, 2, 2 | 0.05 | 0.8418 | 0.852755 | 0.843856 |
| n's: 12, 12, 12, 12, 12, 12, 12 | 0.01 | 0.6190 | 0.633815 | 0.620704 |
| **μ's: 0, -0.5, 0.5** | 0.10 | 0.2548 | 0.259249 | 0.257149 |
| σ's: 2, 2, 2 | 0.05 | 0.1549 | 0.160861 | 0.156251 |
| n's: 12, 12, 12 | 0.01 | 0.0470 | 0.049045 | 0.042292 |
| **μ's: 0, -1, 1** | 0.10 | 0.6540 | 0.659073 | 0.654105 |
| σ's: 2, 2, 2 | 0.05 | 0.5205 | 0.522885 | 0.515816 |
| n's: 12, 12, 12 | 0.01 | 0.2612 | 0.263550 | 0.252469 |
| **μ's: 0, -1.5, 1.5** | 0.10 | 0.9364 | 0.935939 | 0.937768 |
| σ's: 2, 2, 2 | 0.05 | 0.8747 | 0.875620 | 0.872608 |
| n's: 12, 12, 12 | 0.01 | 0.6614 | 0.664478 | 0.652563 |
| **μ's: 0, -1.75, 1.75** | 0.10 | 0.9810 | 0.981434 | 0.986815 |
| σ's: 2, 2, 2 | 0.05 | 0.9522 | 0.956100 | 0.959796 |
| n's: 12, 12, 12 | 0.01 | 0.8251 | 0.830726 | 0.823624 |
| **μ's: 0, -2, 2** | 0.10 | 0.9953 | 0.995969 | 0.999332 |
| σ's: 2, 2, 2 | 0.05 | 0.9878 | 0.988175 | 0.993705 |
| n's: 12, 12, 12 | 0.01 | 0.9308 | 0.931922 | 0.933446 |
| **μ's: 0, -2.5, 2.5** | 0.10 | 0.9999 | 0.999923 | 1.00000 |
| σ's: 2, 2, 2 | 0.05 | 0.9997 | 0.999634 | 1.00000 |
| n's: 12, 12, 12 | 0.01 | 0.9949 | 0.994725 | 0.99909 |
| **μ's: 0, -3, 3** | 0.10 | 1.0000 | 1.00000 | 1.00000 |
| σ's: 2, 2, 2 | 0.05 | 1.0000 | 1.00000 | 1.00000 |
| n's: 12, 12, 12 | 0.01 | 0.9999 | 0.99985 | 1.00000 |
| **μ's: 0, -3.5, 3.5** | 0.10 | 1.0000 | 1.00000 | 1.00000 |
| σ's: 2, 2, 2 | 0.05 | 1.0000 | 1.00000 | 1.00000 |
| n's: 12, 12, 12 | 0.01 | 0.9999 | 1.00000 | 1.00000 |

| Example | alpha | Simulated power | Noncentral F | Kulinskaya et al. |
|---|---|---|---|---|
| **μ's: 0, -0.142857, 0.857143** | 0.10 | 0.1452 | 0.143156 | 0.146824 |
| σ's: 2, 2, 4 | 0.05 | 0.0790 | 0.077699 | 0.077538 |
| n's: 14, 12, 8 | 0.01 | 0.0223 | 0.018200 | 0.014338 |
| **μ's: 0, -0.285714, 1.71429** | 0.10 | 0.2765 | 0.274240 | 0.286222 |
| σ's: 2, 2, 4 | 0.05 | 0.1787 | 0.170628 | 0.179469 |
| n's: 14, 12, 8 | 0.01 | 0.0624 | 0.051588 | 0.050335 |
| **μ's: 0, -0.428571, 2.57143** | 0.10 | 0.4861 | 0.476925 | 0.490018 |
| σ's: 2, 2, 4 | 0.05 | 0.3487 | 0.338626 | 0.355743 |
| n's: 14, 12, 8 | 0.01 | 0.1467 | 0.132405 | 0.141352 |
| **μ's: 0, -0.50000, 3** | 0.10 | 0.5846 | 0.588533 | 0.596795 |
| σ's: 2, 2, 4 | 0.05 | 0.4425 | 0.444491 | 0.460707 |
| n's: 14, 12, 8 | 0.01 | 0.2107 | 0.197290 | 0.212798 |
| **μ's: 0, -0.571429, 3.42857** | 0.10 | 0.6933 | 0.694684 | 0.696773 |
| σ's: 2, 2, 4 | 0.05 | 0.5631 | 0.555731 | 0.567129 |
| n's: 14, 12, 8 | 0.01 | 0.3052 | 0.279131 | 0.299302 |
| **μ's: 0, -0.714286, 4.28571** | 0.10 | 0.8480 | 0.861469 | 0.859329 |
| σ's: 2, 2, 4 | 0.05 | 0.7402 | 0.759703 | 0.759762 |
| n's: 14, 12, 8 | 0.01 | 0.4871 | 0.480052 | 0.497421 |
| **μ's: 0, -0.857143, 5.14286** | 0.10 | 0.9434 | 0.952562 | 0.961913 |
| σ's: 2, 2, 4 | 0.05 | 0.8869 | 0.898817 | 0.902716 |
| n's: 14, 12, 8 | 0.01 | 0.6649 | 0.687058 | 0.692591 |
| **μ's: 0, -1, 6** | 0.10 | 0.9849 | 0.987981 | 0.999989 |
| σ's: 2, 2, 4 | 0.05 | 0.9609 | 0.967589 | 0.985049 |
| n's: 14, 12, 8 | 0.01 | 0.8294 | 0.847436 | 0.853787 |
| **μ's: 0, -1.14286, 6.85714** | 0.10 | 0.9976 | 0.997776 | 1.00000 |
| σ's: 2, 2, 4 | 0.05 | 0.9890 | 0.992220 | 1.00000 |
| n's: 14, 12, 8 | 0.01 | 0.9222 | 0.940972 | 0.96383 |

| Example | alpha | Simulated power | Noncentral F | Kulinskaya et al. |
|---|---|---|---|---|
| **μ's: 1, 2, 3**<br>σ's: 0.3, 2.4, 3.6<br>n's: 13, 19, 25 | 0.10<br>0.05<br>0.01 | 0.8838<br>0.7995<br>0.5632 | 0.882194<br>0.797869<br>0.556486 | 0.884649<br>0.802137<br>0.563208 |
| **μ's: 1, 2, 3**<br>σ's: 2.77489, 2.77489, 2.77489<br>n's: 13, 19, 25 | 0.10<br>0.05<br>0.01 | 0.5649<br>0.4305<br>0.1994 | 0.566831<br>0.431302<br>0.201329 | 0.565141<br>0.428126<br>0.195734 |

The above results are summarized in the below graph, which shows the discrepancies between each approximation and the value of power estimated by simulation.

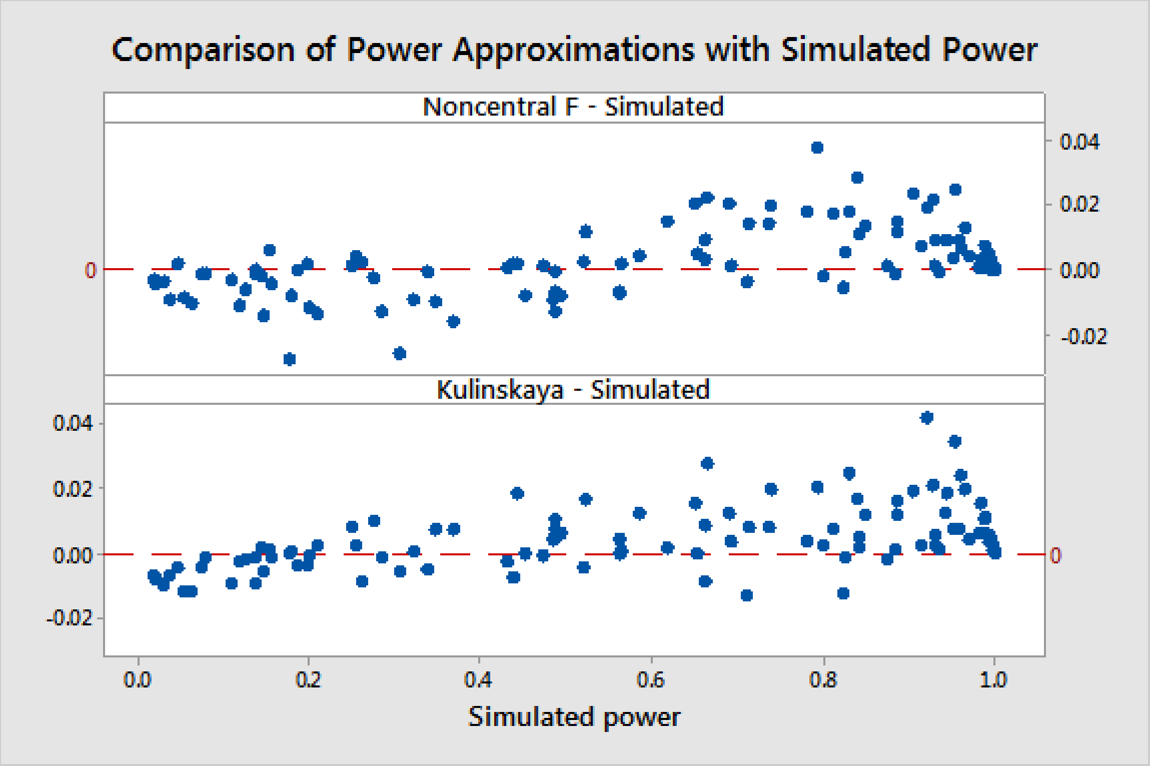


**Figure 8** Comparison of two power approximations and the power estimated by the simulation

# Appendix D: Normality

In this section, we present the simulations that examine the performance of the Welch test and comparison intervals with samples of small to moderate size from several nonnormal distributions.

The tables below summarize simulation results for different types of distributions under the null hypotheses of equal means. For these examples, all standard deviations are also equal and all samples are of equal size. The number of samples is k = 3, 5, or 7.

Each cell shows the estimate of the Type I error based on 10,000 simulations. The target significance level (target $\alpha$) is 0.05.

**Table 4** Simulation results of the Welch test with equal mean for different distributions

| | Sample size n = 10 | | | Sample size n = 15 | | |
|---|---|---|---|---|---|---|
| **Distribution** | **k = 3** | **k = 5** | **k = 7** | **k = 3** | **k = 5** | **k = 7** |
| **N(0,1)** | 0.0490 | 0.0486 | 0.0512 | 0.0534 | 0.0522 | 0.0550 |
| **T(3)** | 0.0371 | 0.0361 | 0.0348 | 0.0353 | 0.0385 | 0.0365 |
| **T(5)** | 0.0440 | 0.0425 | 0.0439 | 0.0435 | 0.0428 | 0.0428 |
| **Laplace(0,1)** | 0.0433 | 0.0354 | 0.0345 | 0.0445 | 0.0397 | 0.0407 |
| **Uniform(-1, 1)** | 0.0544 | 0.0640 | 0.0718 | 0.0517 | 0.0573 | 0.0585 |
| **Beta(3, 3)** | 0.0504 | 0.0577 | 0.0622 | 0.0501 | 0.0538 | 0.0564 |
| **Exponential** | 0.0508 | 0.0621 | 0.0748 | 0.0483 | 0.0633 | 0.0779 |
| **Chi-square(3)** | 0.0473 | 0.0579 | 0.0753 | 0.0499 | 0.0588 | 0.0703 |
| **Chi-square(5)** | 0.0458 | 0.0594 | 0.0643 | 0.0504 | 0.0606 | 0.0679 |
| **Chi-square(10)** | 0.0463 | 0.0510 | 0.0585 | 0.0463 | 0.0552 | 0.0567 |
| **Beta(8, 1)** | 0.0500 | 0.0622 | 0.0775 | 0.0549 | 0.0653 | 0.0760 |

The Type I error rates are all within 3 percentage points of the target $\alpha$ even with samples of size 10. Larger deviations tend to occur with more groups and with distributions that are far from normal. At sample sizes of 10, the only cases where the acceptance probability is off by more than 2 percentage points are for k = 7. These occur for the uniform distribution, which has much shorter tails than the normal, and for the highly skewed exponential, chi-square(3), and beta(8, 1) distributions. Increasing the sample sizes to 15 markedly improves the results for the uniform distribution, but not for the two highly skewed distributions.

We performed a similar simulation for comparison intervals. The simulated $\alpha$ in this case is the number of simulations out of 10,000 in which some intervals do not overlap. The target $\alpha = 0.05$.

**Table 5** Simulation results of comparison intervals with equal means for different distributions

| | Sample size n = 10 | | | Sample size n = 15 | | |
|---|---|---|---|---|---|---|
| Distribution | k = 3 | k = 5 | k = 7 | k = 3 | k = 5 | k = 7 |
| N(0,1) | 0.0493 | 0.0494 | 0.0469 | 0.0538 | 0.0518 | 0.0561 |
| t(3) | 0.0378 | 0.0321 | 0.0254 | 0.0347 | 0.0343 | 0.0289 |
| t(5) | 0.0449 | 0.0399 | 0.0361 | 0.0447 | 0.0444 | 0.0412 |
| Laplace(0,1) | 0.0438 | 0.0305 | 0.0246 | 0.0456 | 0.0366 | 0.0348 |
| Uniform(-1, 1) | 0.0559 | 0.0605 | 0.0699 | 0.0534 | 0.0607 | 0.0590 |
| Beta(3, 3) | 0.0515 | 0.0569 | 0.0615 | 0.0510 | 0.0553 | 0.0568 |
| Exponential | 0.0353 | 0.0254 | 0.0207 | 0.0346 | 0.0310 | 0.0275 |
| Chi-square(3) | 0.0375 | 0.0305 | 0.0296 | 0.0384 | 0.0359 | 0.0339 |
| Chi-square(5) | 0.0405 | 0.0390 | 0.0353 | 0.0417 | 0.0433 | 0.0416 |
| Chi-square(10) | 0.0425 | 0.0428 | 0.0447 | 0.0435 | 0.0476 | 0.0464 |
| Beta(8, 1) | 0.0381 | 0.0352 | 0.0287 | 0.0459 | 0.0428 | 0.0403 |

As with the Welch test, the Type I error rates are all within 3 percentage points of the target $\alpha$ even with samples of size 10. Larger deviations tend to occur with more samples and with distributions that are far from normal. At sample sizes of 10, the error rates are sometimes off by more than 2 percentage points for k = 7 (and in one case, for k = 5). These cases occur for the extremely heavy-tailed t distribution with 3 degrees of freedom, the Laplace distribution, and the highly skewed exponential and Chi-square (3) distributions. Increasing the sample sizes to 15 improves the results, leaving only the t(3) and exponential distributions with simulated $\alpha$ values that are off-target by more than 2 percentage points. Note that unlike the results for the Welch test, the larger deviations for comparison intervals are on the conservative side.

One-way ANOVA in the Assistant allows up to k = 12 samples, so next we consider results for more than 7 samples. The table below shows the Type I error rates using the Welch test for nonnormal data in k = 9 groups. Again, the target $\alpha = 0.05$.

**Table 6** Simulation results of Welch test for different distributions with 9 samples

| Distribution | k = 9 |
|---|---|
| t(3) | 0.0362 |
| t(5) | 0.0426 |
| Laplace(0,1) | 0.0402 |
| Uniform(-1, 1) | 0.0625 |
| Beta(3, 3) | 0.0584 |
| Exponential | 0.0885 |
| Chi-square(3) | 0.0774 |
| Chi-square(5) | 0.0686 |
| Chi-square(10) | 0.0581 |
| Beta(8, 1) | 0.0863 |

As might be expected, the highly skewed distributions show the largest deviations from the target $\alpha$. Even so, none of the error rates deviate from the target by more than 4 percentage points, although the deviation for the exponential distribution is close. The Report Card treats samples of size 15 sufficient not to flag a problem for nonnormal data because all the results are at least reasonably close to the target $\alpha$.

Samples of size n = 15 do not perform as well when we get to k = 12 samples. Below we consider the simulated results for the Welch test for a range of sample sizes using extremely nonnormal distributions, which will assist us in developing a reasonable criterion for the sample size.

**Table 7** Simulation results of Welch test for different distributions with 12 samples

| n | T(3) | Uniform | Chi-square(5) |
|---|---|---|---|
| 10 | 0.0397 | 0.0918 | 0.0792 |
| 15 | 0.0351 | 0.0695 | 0.0717 |
| 20 | 0.0362 | 0.0622 | 0.0671 |
| 30 | 0.0408 | 0.0573 | 0.0657 |

For these distributions n = 15 is acceptable if we are willing to accept a deviation of slightly over 2 percentage points from the target $\alpha$. To keep the deviation below 2 percentage points the sample size should be 20. Now, we consider the results from the more skewed chi-square (3) and exponential distributions.

**Table 8** Simulation results of Welch test for chi-square and exponential distributions with 12 samples

| n | Chi-square(3) | Exponential |
|---|---|---|
| 10 | 0.1013 | 0.1064 |
| 15 | 0.0854 | 0.1079 |
| 20 | 0.0850 | 0.0951 |
| 30 | 0.0746 | 0.0829 |
| 40 | 0.0727 | 0.0735 |
| 50 | 0.0675 | 0.0694 |

These highly skewed distributions present more of a challenge. If we are willing to accept a deviation of well over 3 percentage points from the target $\alpha = 0.05$, then n = 15 could be considered sufficient even for the chi-square (3) distribution, but the exponential distribution would require something closer to n = 30. While the criterion of a specific sample size is somewhat arbitrary, and that n = 20 works quite well for a wide range of distributions and marginally well for extremely skewed distributions, we use n = 20 as the minimum recommended sample size for 10 to 12 samples. Clearly, if there is a need to keep the deviation small even for extremely skewed distributions, then larger samples are recommended.