

## Chapter 202

# Paired T-Test for Equivalence

---

### Introduction

This procedure provides reports for making inference about the equivalence of two variables based on a paired sample. The question of interest is whether two variables, each measured on the same subject, are actually equivalent, that is, differ on average by a small margin, at most. This is tested by the TOST (two one-sided tests) equivalence test. This test allows you to obtain a p-value so that you can quantify the results of hypothesis test.

---

### Technical Details

Suppose you want to evaluate the equivalence of a continuous random variable  $X_T$  as compared to a second paired random variable  $X_C$ . Assume that  $n$  paired observations  $(X_{Tk}, X_{Ck})$ ,  $k = 1, 2, \dots, n$  are available. The  $D$ 's are the differences formed as  $D = X_T - X_C$ .

---

### TOST Equivalence Test

Schuirmann's (1987) two one-sided tests (TOST) approach is used to test equivalence. The equivalence test essentially reverses the roles of the null and alternative hypothesis. Assume that  $\mu_D = \mu_{T-C}$  represent the mean of the differences between the two variables and  $M$  to represent the so-called *margin of equivalence*, the null and alternative hypotheses are

$$H_0: \mu_D < -M \text{ or } \mu_D > M$$

$$H_1: -M < \mu_D < M$$

The null hypothesis is made up of two simple one-sided hypotheses:

$$H_{0_1}: \mu_D < -M$$

$$H_{0_2}: \mu_D > M$$

If both of these one-sided tests are rejected, we conclude  $H_1$  that the paired variables are equivalent (their average difference is confined within a small margin). Schuirmann showed that if we want the alpha level of the equivalence test to be  $\alpha$ , then each of the one-sided tests should be  $\alpha$  as well (not  $\alpha/2$  as you might expect). The probability level (p-value) of the equivalence test is equal to the maximum of the probability levels of the two one-sided tests. These tests are conducted using the standard formulas for the paired t-test.

## Paired T-Test for Equivalence

---

### Assumptions

This section describes the assumptions that are made when you use each of the tests of this procedure. The key assumption relates to normality or non-normality of the data. One of the reasons for the popularity of the t-test is its robustness in the face of assumption violation. Unfortunately, in practice it often happens that more than one assumption is not met. Hence, take the steps to check the assumptions before you make important decisions based on these tests. There are reports in this procedure that permit you to examine the assumptions, both visually and through assumptions tests.

### Paired T-Test Assumptions

The assumptions of the paired t-test are:

1. The data are continuous (not discrete).
2. The data, i.e., the differences for the matched-pairs, follow a normal probability distribution.
3. The sample of pairs is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample.

### Wilcoxon Signed-Rank Test Assumptions

The assumptions of the Wilcoxon signed-rank test are as follows (note that the difference is between the two data values of a pair):

1. The differences are continuous (not discrete).
2. The distribution of these differences is symmetric.
3. The differences are mutually independent.
4. The differences all have the same median.
5. The measurement scale is at least interval.

---

## Data Structure

For this procedure, the data are entered in two columns.

| X1 | X2 |
|----|----|
| 57 | 53 |
| 63 | 69 |
| 66 | 63 |
| 74 | 76 |
| 77 | 75 |
| 77 | 79 |
| 78 | 77 |
| 79 | 77 |
| 80 | 81 |

---

## Procedure Options

This section describes the options available in this procedure.

---

### Variables Tab

This panel specifies the variables that will be used in the analysis.

---

#### Select Paired Variables

##### Treatment Variable

Select (or enter) the name or column number of the treatment variable. Each paired difference is calculated as Treatment - Control.

Only one variable may be entered. It is assumed that the two values on each row come from the same subject. Each row represents a different subject.

##### Control Variable

Select (or enter) the name or column number of the control variable. Each paired difference is calculated as Treatment - Control.

Only one variable may be entered. It is assumed that the two values on each row come from the same subject. Each row represents a different subject.

---

#### Equivalence Test Options

##### Equivalence Bounds

This option allows the user to specify how the equivalence bounds are created. The choices are

- **Symmetric**

The equivalence bounds are symmetric about zero. In this case you enter a single margin value and the equivalence bounds are computed as

$$\text{Upper Equivalence Bound} = (\text{Equivalence Margin})$$

$$\text{Lower Equivalence Bound} = -(\text{Equivalence Margin})$$

- **Asymmetric**

The upper and lower equivalence bounds are entered separately.

##### Equivalence Margin (*Shown when Equivalence Bounds = "Symmetric"*)

Enter the desired value for the margin of equivalence. The scale of this value is the same as the data values. The equivalence bounds are computed as

$$\text{Upper Equivalence Bound} = (\text{Equivalence Margin})$$

$$\text{Lower Equivalence Bound} = -(\text{Equivalence Margin})$$

##### Range

This value should be positive.

##### Upper and Lower Equivalence Bounds (*Shown when Equivalence Bounds = "Asymmetric"*)

Enter the values for the upper and lower equivalence bounds. The Upper Equivalence Bound should be positive and the Lower Equivalence Bound should be negative.

## Reports Tab

The options on this panel specify which reports will be included in the output.

---

### Select Reports

#### Descriptive Statistics

This section reports the means, standard deviations, standard errors, and confidence intervals of each variable and their difference.

#### Confidence Level

This confidence level is used for the confidence intervals of the means and difference. Typical confidence levels are 90%, 95%, and 99%, with 95% being the most common.

---

### Tests

#### Alpha

This is the significance level of the equivalence tests. A value of 0.05 is recommended. Typical values range from 0.001 to 0.200.

---

### Tests – Parametric

#### Paired T-Test

This report provides the results of the common (parametric) paired-difference T-Test. This test should only be used when the differences follow the normal distribution.

---

### Tests – Nonparametric

#### Wilcoxon Signed-Rank Test

This nonparametric test makes use of the sign and the magnitude of the rank of the differences (paired differences minus the hypothesized difference). It is one of the most commonly used nonparametric alternatives to the paired t-test.

There are 3 different tests that can be conducted:

- **Exact Test**

The exact test can be calculated if there are no ties. This test is recommended when there are no ties.

- **Normal Approximation Test**

The normal approximation method may be used to approximate the distribution of the sum of ranks when the sample size  $\geq 10$ .

- **Normal Approximation Test with Continuity Correction**

The normal approximation with continuity correction may be used to approximate the distribution of the sum of ranks when the sample size  $\geq 10$ .

## Paired T-Test for Equivalence

---

### Assumptions

#### Shapiro-Wilk Test for Normality of the Differences

This section reports a Shapiro-Wilk normality test.

#### Assumptions Alpha

This is the significance level of the Shapiro-Wilk normality test. A value of 0.05 is recommended. Typical values range from 0.001 to 0.200.

---

### Report Options Tab

The options on this panel control the label and decimal options of the report.

---

#### Report Options

##### Variable Names

This option lets you select whether to display only variable names, variable labels, or both.

---

##### Decimal Places

##### Means, Differences, and C.I. Limits – Test Statistics

These options specify the number of decimal places used in the reports. If one of the Auto options is used, the ending zero digits are not shown. For example, if 'Significant Digits (Up to 7)' is chosen,

0.0500 is displayed as 0.05 and 1.314583689 is displayed as 1.314584.

The output formatting system is not designed to accommodate (Up to 13), and if chosen, this will likely lead to lines that run on to a second line. This option is included, however, for the rare case when a very large number of decimals is needed.

---

### Plots Tab

The options on this panel control the inclusion and appearance of the plots.

---

#### Select Plots

##### Histogram ... Average-Difference Plot

Check the boxes to display the plot. Click the plot format button to change the plot settings.

## Paired T-Test for Equivalence

## Example 1 – TOST Equivalence Test using Paired Data

This section presents an example of how to test the equivalence of two measurement methods. Suppose two measurements were made on each of 100 subjects. The first measurement was made by a lengthy, invasive method and the second measurement was made by a second, much less invasive method. The data are in the **Bland-Altman** dataset. The researchers wish to determine if the difference between measurements are within 1 of each other on average.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Paired T-Test for Equivalence window.

### 1 Open the Bland-Altman dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Bland-Altman.NCSS**.
- Click **Open**.

### 2 Open the Paired T-Test for Equivalence window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Paired T-Test for Equivalence** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

### 3 Specify the variables.

- Select the **Variables tab**. (This is the default.)
- Double-click in the **Treatment Variable** text box. This will bring up the variable selection window.
- Select **Method1** from the list of variables and then click **Ok**. “Method1” will appear in this box.
- Double-click in the **Control Variable** text box. This will bring up the variable selection window.
- Select **Method2** from the list of variables and then click **Ok**. “Method2” will appear in this box.
- Leave the **Equivalence Bounds** at **Symmetric**.
- Set the **Equivalence Margin** to **1**.

### 4 Specify the reports.

- Select the **Reports tab**.
- Leave all checked reports as checked.
- Check the **Wilcoxon Signed-Rank Test** check box. Leave all sub-boxes checked.

### 5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

The following reports and charts will be displayed in the Output window.

## Descriptive Statistics

| Descriptive Statistics |       |       | Standard Deviation of Data | Standard Error of Mean | T*     | 95.0% LCL of Mean | 95.0% UCL of Mean |
|------------------------|-------|-------|----------------------------|------------------------|--------|-------------------|-------------------|
| Variable               | Count | Mean  |                            |                        |        |                   |                   |
| Method1                | 100   | 50.72 | 28.30893                   | 2.830893               | 1.9842 | 45.10289          | 56.3371           |
| Method2                | 100   | 50.62 | 28.07701                   | 2.807702               | 1.9842 | 45.04891          | 56.19109          |
| Difference             | 100   | 0.1   | 2.787055                   | 0.2787055              | 1.9842 | -0.4530122        | 0.6530122         |

Correlation Coefficient = 0.995147

This report provides basic descriptive statistics and confidence intervals for the two variables and their difference.

## Paired T-Test for Equivalence

### Variable

These are the names of the variables or groups.

### Count

The count gives the number of non-missing values. This value is often referred to as the group sample size or  $n$ .

### Mean

This is the average for each group.

### Standard Deviation

The sample standard deviation is the square root of the sample variance. It is a measure of spread.

### Standard Error

This is the estimated standard deviation for the distribution of sample means for an infinite population. It is the sample standard deviation divided by the square root of sample size,  $n$ .

### T\*

This is the t-value used to construct the confidence interval. If you were constructing the interval manually, you would obtain this value from a table of the Student's t distribution with  $n - 1$  degrees of freedom.

### LCL and UCL of the Mean

This is the lower limit of an interval estimate of the mean based on a Student's t distribution with  $n - 1$  degrees of freedom. This interval estimate assumes that the population standard deviation is not known and that the data are normally distributed. The confidence interval formula is

$$\bar{x} \pm T_{df} \cdot SE_{\bar{x}}$$

## Paired T-Test for Equivalence using TOST (Two One-Sided Tests)

### Paired T-Test for Equivalence using TOST (Two One-Sided Tests)

Equivalence Hypothesis:  $-1 < [\text{Mean of (Method1)} - (\text{Method2})] < 1$

| Test           | Alternative Hypothesis† | Mean Difference | Standard Error | T-Statistic | DF | Prob Level | Reject H0 at $\alpha = 0.050?$ |
|----------------|-------------------------|-----------------|----------------|-------------|----|------------|--------------------------------|
| Lower Boundary | Diff > -1               | 0.1             | 0.2787055      | 3.9468      | 99 | 0.00007    | Yes                            |
| Upper Boundary | Diff < 1                | 0.1             | 0.2787055      | -3.2292     | 99 | 0.00084    | Yes                            |
| Equivalence    | -1 < Diff < 1           | 0.1             |                |             |    | 0.00084    | Yes                            |

† "Diff" refers to the Mean Difference.

This report shows the equivalence test (the third row) and the two one-sided tests (first and second rows). Since the Prob Level for the equivalence test (the maximum of the two one-sided tests) is less than the designated value of alpha (0.05), we can reject the null hypothesis and conclude that the means are equivalent for an equivalence margin of 1. This test assumes that the differences are normal.

### Test

The two one-sided tests are shown on the first two rows and the equivalence test is shown at the end.

### Alternative Hypothesis

Assume that  $\mu_D = \mu_{T-C}$  represents the mean of the differences between the two variables and that  $M$  is the positive *equivalence margin*. The null and alternative hypotheses are

$$\begin{aligned} H_0: \mu_D < -M \text{ or } \mu_D > M \\ H_1: -M < \mu_D < M \end{aligned}$$

## Paired T-Test for Equivalence

### Mean Difference

This is the average of the paired differences,  $\bar{x}_D$ .

### Standard Error

This is the estimated standard deviation of the distribution of sample means for an infinite population.

$$SE_{\bar{x}_D} = \frac{s_D}{\sqrt{n}}$$

### T-Statistic

The T-Statistic is the value used to produce the  $p$ -value (Prob Level) based on the  $T$  distribution. The formula for the T-Statistic is:

$$T = \frac{\bar{x}_D - M}{SE_{\bar{x}_D}}$$

### DF

The degrees of freedom define the  $T$  *distribution* upon which the probability values are based. The formula for the degrees of freedom is the number of pairs minus one:

$$df = n - 1$$

### Prob Level

The probability level, also known as the  $p$ -value or significance level, is the probability that the test statistic will take a value at least as extreme as the observed value, assuming that the null hypothesis is true. If the  $p$ -value is less than the prescribed  $\alpha$ , in this case 0.05, the null hypothesis is rejected in favor of the alternative hypothesis. Otherwise, there is not sufficient evidence to reject the null hypothesis. The Prob Level for the equivalence test is equal to the maximum of the two one-sided tests.

### Reject H0 at $\alpha = 0.050$ ?

This column indicates whether or not the null hypothesis is rejected, in favor of the alternative hypothesis, based on the  $p$ -value and chosen  $\alpha$ . A test in which the null hypothesis is rejected is sometimes called *significant*.

## Wilcoxon Signed-Rank Test for Equivalence using TOST (Two One-Sided Tests)

### Wilcoxon Signed-Rank Test for Equivalence using TOST (Two One-Sided Tests)

Equivalence Hypothesis:  $-1 < [\text{Median of (Method1)} - (\text{Method2})] < 1$

| Test                     | Sum of Ranks (W) | Mean of W               | Std Dev of W | Number of Zeros | Number Sets of Ties             | Multiplicity Factor |
|--------------------------|------------------|-------------------------|--------------|-----------------|---------------------------------|---------------------|
| Lower Boundary           | 3807.5           | 2486                    | 288.7449     | 12              | 5                               | 50466               |
| Upper Boundary           | 1464.5           | 2448.5                  | 288.4896     | 17              | 5                               | 43920               |
| Test Type                | Test             | Alternative Hypothesis† | Z-Value      | Prob Level      | Reject H0 at $\alpha = 0.050$ ? |                     |
| Exact*                   | Lower Boundary   | Diff > -1               |              |                 |                                 |                     |
| Exact*                   | Upper Boundary   | Diff < 1                |              |                 |                                 |                     |
| Exact*                   | Equivalence      | -1 < Diff < 1           |              |                 |                                 |                     |
| Normal Approximation     | Lower Boundary   | Diff > -1               | 4.5767       | 0.00000         | Yes                             |                     |
| Normal Approximation     | Upper Boundary   | Diff < 1                | -3.4109      | 0.00032         | Yes                             |                     |
| Normal Approximation     | Equivalence      | -1 < Diff < 1           |              | 0.00032         | Yes                             |                     |
| Normal Approx. with C.C. | Lower Boundary   | Diff > -1               | 4.5750       | 0.00000         | Yes                             |                     |
| Normal Approx. with C.C. | Upper Boundary   | Diff < 1                | -3.4091      | 0.00033         | Yes                             |                     |
| Normal Approx. with C.C. | Equivalence      | -1 < Diff < 1           |              | 0.00033         | Yes                             |                     |



## Paired T-Test for Equivalence

† "Diff" refers to the Median Difference.

\* The Exact Test is provided only when there are no ties.

This report shows the nonparametric Wilcoxon Signed-Rank Test for equivalence. Since the Prob Level for the equivalence test (the maximum of the two one-sided tests) is less than the designated value of alpha (0.05), we can reject the null hypothesis and conclude that the medians are equivalent for an equivalence margin of 1. This test required no assumption that the differences are normal.

### Sum of Ranks (W)

The basic statistic for this test is the sum of the positive ranks,  $\Sigma R_+$  (The sum of the positive ranks is chosen arbitrarily. The sum of the negative ranks could equally be used). This statistic is called  $W$ .

$$W = \Sigma R_+$$

### Mean of W

This is the mean of the sampling distribution of the sum of ranks for a sample of  $n$  items.

$$\mu_W = \frac{n(n+1) - d_0(d_0+1)}{4}$$

where  $d_0$  is the number of zero differences.

### Std Dev of W

This is the standard deviation of the sampling distribution of the sum of ranks. Here  $t_i$  represents the number of times the  $i^{\text{th}}$  value occurs.

$$s_W = \sqrt{\frac{n(n+1)(2n+1) - d_0(d_0+1)(2d_0+1)}{24} - \frac{\Sigma t_i^3 - \Sigma t_i}{48}}$$

where  $d_0$  is the number zero differences,  $t_i$  is the number of absolute differences that are tied for a given non-zero rank, and the sum is over all sets of tied ranks.

### Number of Zeros

This is the number of times that the difference between the observed paired difference and the hypothesized value is zero. The zeros are used in computing ranks, but are not considered positive ranks or negative ranks.

### Number Sets of Ties

The treatment of ties is to assign an average rank for the particular set of ties. This is the number of sets of ties that occur in the data, including ties at zero.

### Multiplicity Factor

This is the correction factor that appeared in the standard deviation of the sum of ranks when there were ties.

### Test Type

This is the type of test that is being reported on the current row. The Exact Test is provided only when there are no ties.

### Alternative Hypothesis

For the Wilcoxon signed-rank test, the null and alternative hypotheses relate to the median. The left-tail alternative is represented by Median < M (i.e.,  $H_a$ : median < M) while the right-tail alternative is depicted by Median > -M.

## Paired T-Test for Equivalence

### Exact Probability: Prob Level

This is an exact  $p$ -value for this statistical test, assuming no ties. The  $p$ -value is the probability that the test statistic will take on a value at least as extreme as the actually observed value, assuming that the null hypothesis is true. If the  $p$ -value is less than  $\alpha$ , say 5%, the null hypothesis is rejected. If the  $p$ -value is greater than  $\alpha$ , the null hypothesis is accepted.

### Exact Probability: Reject H0 ( $\alpha = 0.050$ )

This is the conclusion reached about the null hypothesis. It will be to either fail to reject  $H_0$  or reject  $H_0$  at the assigned level of significance.

### Approximations with (and without) Continuity Correction: Z-Value

Given the sample size is at least ten, a normal approximation method may be used to approximate the distribution of the sum of ranks. Although this method does correct for ties, it does not have the continuity correction factor. The  $z$  value is as follows:

$$z = \frac{W - \mu_W}{\sigma_W}$$

If the correction factor for continuity is used, the formula becomes:

$$z = \frac{W - \mu_W \pm \frac{1}{2}}{\sigma_W}$$

### Approximations with (and without) Continuity Correction: Prob Level

This is the  $p$ -value for the normal approximation approach for the Wilcoxon signed-rank test. The  $p$ -value is the probability that the test statistic will take a value at least as extreme as the actually observed value, assuming that the null hypothesis is true. If the  $p$ -value is less than  $\alpha$ , say 5%, the null hypothesis is rejected. If the  $p$ -value is greater than  $\alpha$ , the null hypothesis is accepted.

### Approximations with (and without) Continuity Correction: Reject H0 ( $\alpha = 0.050$ )

This is the conclusion reached about the whether to reject null hypothesis. It will be either Yes or No at the given level of significance.

---

## Test of Assumptions

### Test of the Normality of Differences Assumption

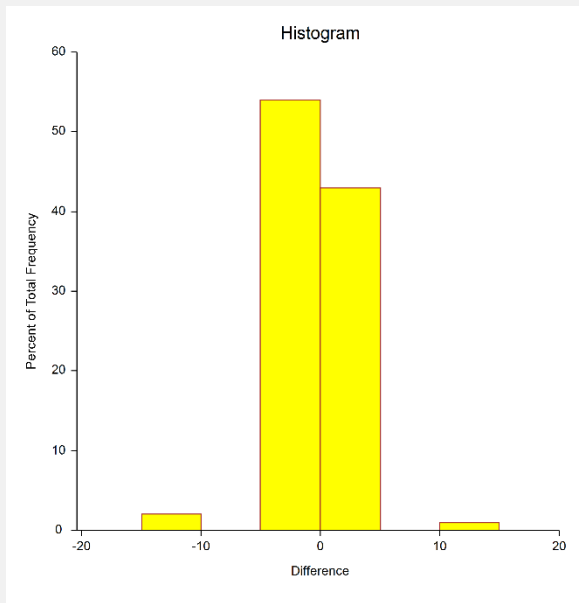
| Assumption   | Value  | Prob Level | Decision ( $\alpha = 0.050$ ) |
|--------------|--------|------------|-------------------------------|
| Shapiro-Wilk | 0.8926 | 0.00000    | Reject normality              |

The main assumption when using the t-test is that the paired-difference data come from a normal distribution. The normality assumption can be checked statistically by the Shapiro-Wilk normality test and visually by the histogram or normal probability plot.

This section reports the results of a diagnostic test to determine if the differences are normal. In this case, they are not, probably because of the outliers that were present. This would indicate that the Wilcoxon Signed-Rank Test would be the better test to use.

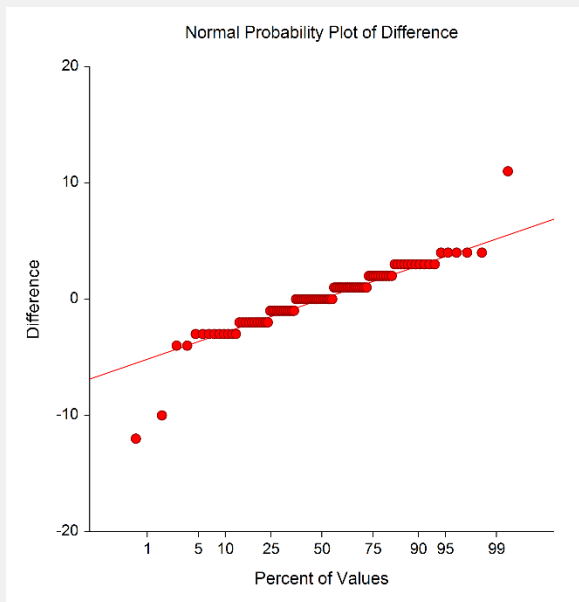
## Paired T-Test for Equivalence

## Histogram of Differences



The nonparametric tests need the assumption of symmetry, and these two graphic tools can provide that information. If the distribution of differences is symmetrical but not normal, proceed with the nonparametric test.

## Probability Plot of Differences

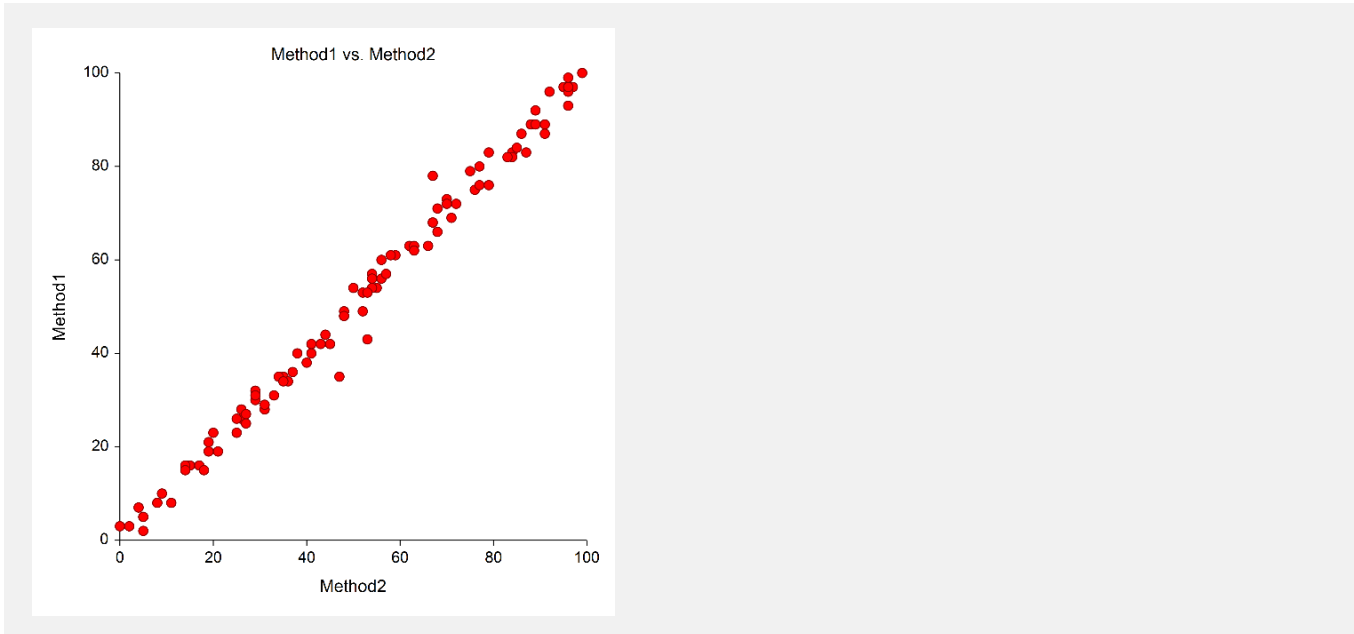


If any of the observations fall outside the confidence bands (if shown), the data are not normal. The goodness-of-fit tests mentioned earlier, especially the omnibus test, should confirm this fact statistically. If only one observation falls outside the confidence bands and the remaining observations hug the straight line, there may be an outlier. If the data were normal, we would see the points falling along a straight line.

Note that the confidence bands are based on large-sample formulas. They may not be accurate for small samples.

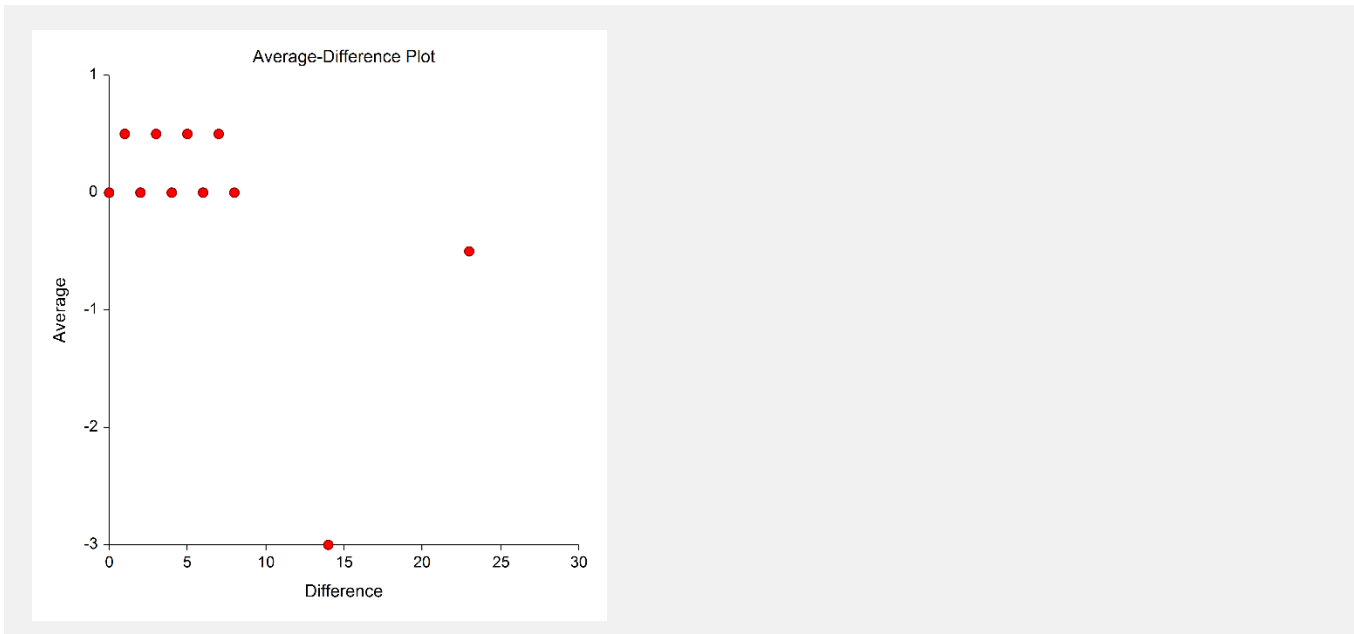
## Paired T-Test for Equivalence

## Scatter Plot of Pairs



This plot allows you to look for patterns between the pairs. Preferably, you would like to see either no correlation or a positive linear correlation between Y and X. If there is a curvilinear relationship between Y and X, the paired t-test is not appropriate. If there is a negative relationship between the observations in the pairs, the paired t-test is not appropriate. If there are outliers, a nonparametric approach might be safer.

## Average vs Difference Plot



This average-difference plot is designed to detect a lack of symmetry in the data. This plot is constructed from the paired differences, not the original data. Here's how. Let  $D(i)$  represent the  $i$ th ordered difference. Pairs of these sorted differences are considered, with the pairing being done as you move toward the middle from either end. That is, consider the pairs  $D(1)$  and  $D(n)$ ,  $D(2)$  and  $D(n-1)$ ,  $D(3)$  and  $D(n-2)$ , etc. Plot the average versus the difference of each of these pairs. Your plot will have about  $n/2$  points, depending on whether  $n$  is odd or even. If

## Paired T-Test for Equivalence

the data are symmetric, the average of each pair will be the median and the difference between each pair will be zero.

Symmetry is an important assumption for the t-test. A perfectly symmetric set of data should show a vertical line of points hitting the horizontal axis at the value of the median. Departures from symmetry would deviate from this standard.