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, ..., 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

$$H_0: \mu_D < -M \quad \text{or} \quad \mu_D > M$$
$$H_1: -M < \mu_D < M$$

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

$$H_{01}: \mu_D < -M$$
$$H_{02}: \mu_D > M$$

If both of these one-sided tests are rejected, we conclude H1 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 α, then each of the one-sided tests should be α as well (not α/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.

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

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.

Setup

To run this example, complete the following steps:

1 Open the Bland-Altman example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Bland-Altman** and click **OK**.

2 Specify the Paired T-Test for Equivalence procedure options

- Find and open the **Paired T-Test for Equivalence** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab	
Treatment Variable	**Method1**
Control Variable	**Method2**
Equivalence Bounds	**Symmetric**
Equivalence Margin	**1**

Reports Tab	
Wilcoxon Signed-Rank Test	**Checked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Descriptive Statistics and Confidence Intervals for the Mean and the Mean Difference

Descriptive Statistics and Confidence Intervals for the Mean and the Mean Difference

Variable	N	Mean	Standard Deviation of the Data	Standard Error of the Mean	T*	95% Confidence Interval Limits for the Mean	
						Lower	Upper
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

Correlation Coefficient = 0.995147

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

Variable

These are the names of the variables or groups.

N

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

Mean

This is the average for each group.

Standard Deviation of the Data

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

Standard Error of the Mean

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.

95% Confidence Interval Limits for the Mean (Lower and Upper)

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

$$\overline{x} \pm T_{df} SE_{\overline{x}}$$

Correlation Coefficient

This is the correlation of the two paired variables.

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

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

Null Hypothesis (H0): [Mean of (Method1) - (Method2)] ≤ -1 or [Mean of (Method1) - (Method2)] ≥ 1
Equivalence Hypothesis (H1): -1 < [Mean of (Method1) - (Method2)] < 1

Test	Alternative Hypothesis†	Mean Difference	Standard Error	T-Statistic	DF	P-Value	Reject H0 at α = 0.05?
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 of the Paired Differences.

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

$$H_0: \mu_D < -M \quad \text{or} \quad \mu_D > M$$
$$H_1: -M < \mu_D < M$$

Mean Difference

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

Standard Error

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

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

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:

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

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:

$$df = n - 1$$

P-Value

The *p*-value, also known as the probability level 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 *α*, 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 α = 0.05?

This column indicates whether or not the null hypothesis is rejected, in favor of the alternative hypothesis, based on the *p*-value and chosen *α*. 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)

Null Hypothesis (H0): [Median of (Method1) - (Method2)] ≤ -1 or [Median of (Method1) - (Method2)] ≥ 1
Equivalence Hypothesis (H1): -1 < [Median of (Method1) - (Method2)] < 1

Test Details

Test	Sum of Ranks (W)	Mean of W	Standard Deviation of W	Number of Zeros	Number of 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 Results

Test Type	Test	Alternative Hypothesis†	Z-Value	P-Value	Reject H0 at α = 0.05?
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

† "Diff" refers to the Median of the Paired Differences.
* 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, $\sum R_+$ (The sum of the positive ranks is chosen arbitrarily. The sum of the negative ranks could equally be used). This statistic is called W.

$$W = \sum R_+$$

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.

Standard Deviation 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^{th} value occurs.

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

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 of 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.

Exact Probability: P-Value

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 *α*, say 5%, the null hypothesis is rejected. If the *p*-value is greater than *α*, the null hypothesis is accepted.

Exact Probability: Reject H0 at α = 0.05?

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: P-Value

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 *α*, say 5%, the null hypothesis is rejected. If the *p*-value is greater than *α*, the null hypothesis is accepted.

Approximations with (and without) Continuity Correction: Reject H0 at α = 0.05?

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.

Shapiro-Wilk Test for Normality of the Paired Differences

Shapiro-Wilk Test for Normality of the Paired Differences

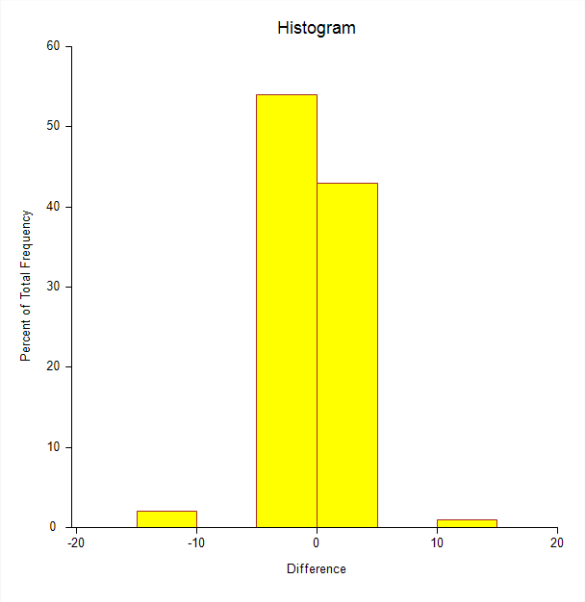
Normality Test	Test Statistic	P-Value	Reject the Assumption of Normality at α = 0.05?
Shapiro-Wilk	0.8926	0.00000	Yes

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.

Histogram of Differences

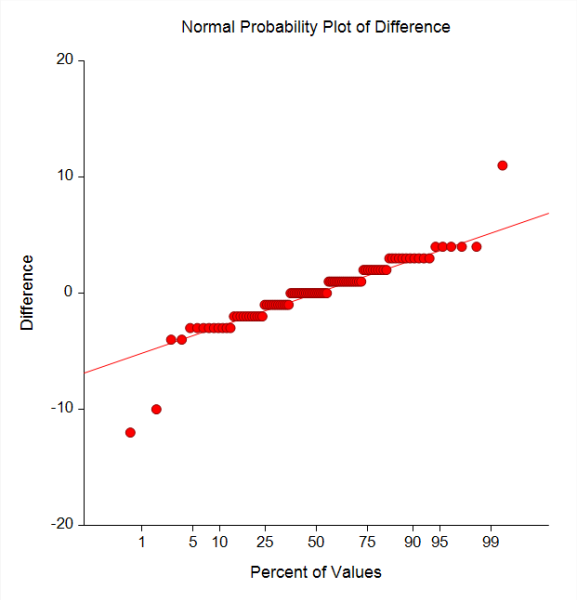
Histogram of Differences



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

Probability Plot of Differences

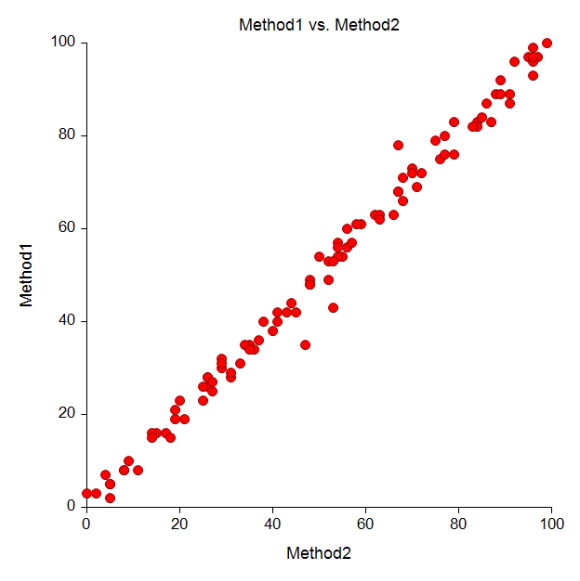


If any of the observations fall outside the confidence bands (if shown), the data are not normal. The goodness-of-fit tests mentioned earlier, especially the omnibus test, should confirm this fact statistically. If only one observation falls outside the confidence bands and the remaining observations 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.

Scatter Plot of Pairs

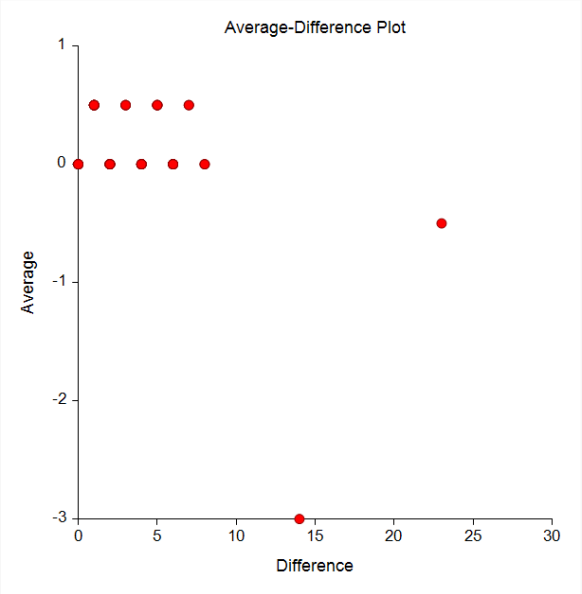
Scatter Plot of Pairs



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

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 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.