Chapter 208

Paired T-Test

Introduction

This procedure provides several reports for making inference about the difference between two population means based on a paired sample. These reports include confidence intervals of the mean difference, the paired sample t-test, and non-parametric tests including the randomization test, the quantile (sign) test, and the Wilcoxon Signed-Rank test. Tests of assumptions and distribution plots are also available in this procedure.

Research Questions

For the paired-sample situation, the prime concern in research is examining a measure of central tendency (location) for the paired-difference population of interest. The best-known measures of location are the mean and median. In the paired case, we take two measurements on the same individual, or we have one measurement on each individual of a pair. Examples of this are two insurance-claim adjusters assessing the dollar damage for the same 15 cases, evaluation of the improvement in aerobic fitness for 15 subjects where measurements are made at the beginning of the fitness program and at the end of it, or the testing of the effectiveness of two drugs, A and B, on 20 pairs of patients who have been matched on physiological and psychological variables. One patient in the pair receives drug A, and the other patient gets drug B.

Technical Details

The technical details and formulas for the methods of this procedure are presented in line with the Example 1 output. The output and technical details are presented in the following order:

- Descriptive Statistics
- Confidence Interval of $\mu_1 - \mu_2$
- Bootstrap Confidence Intervals
- Paired-Sample T-Test and associated power report
- Randomization Test
- Quantile (Sign) Test
- Wilcoxon Signed-Rank Test
- Tests of Assumptions
- Graphs

Data Structure

In the matched-pairs case, the analysis will require two variables. This example shows matched-pairs data with tire wear for the right and left tires of the same car.

Right Tire	Left Tire
42	54
75	73
24	22
56	59
52	51
56	45
23	29
55	58
46	49
52	58
47	49
62	67
55	58
62	64

Null and Alternative Hypotheses

The basic null hypothesis is that the population mean difference is equal to a hypothesized value,

$$H_0: \mu_{diff} = Hypothesized\ Value$$

with three common alternative hypotheses,

$$H_a: \mu_{diff} \neq Hypothesized\ Value\ ,$$

$$H_a: \mu_{diff} < Hypothesized\ Value\ , \text{ or}$$

$$H_a: \mu_{diff} > Hypothesized\ Value\ ,$$

one of which is chosen according to the nature of the experiment or study.

In the most common paired t-test scenario, the hypothesized value is 0, in which the null hypothesis becomes

$$H_0: \mu_{diff} = 0$$

with alternative hypothesis options of

$$H_a: \mu_{diff} \neq 0\ ,$$

$$H_a: \mu_{diff} < 0\ , \text{ or}$$

$$H_a: \mu_{diff} > 0\ .$$

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.

Quantile Test Assumptions

The assumptions of the quantile (sign) test are:

1. A random sample has been taken resulting in observations that are independent and identically distributed.
2. The measurement scale is at least ordinal.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

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

Variables

Paired 1 Variable(s)

Enter the first column of each pair here. The second column of each pair is entered in the "Paired 2 Variable(s)" box. Paired differences are calculated as Paired 1 - Paired 2. If multiple columns are specified in both Paired 1 Variable(s) and Paired 2 Variable(s), the first columns in each box are compared, then the second columns in each box are compared, and so on.

Paired 2 Variable(s)

Enter the second column of each pair here. The first column of each pair is entered in the "Paired 1 Variable(s)" box. Paired differences are calculated as Paired 1 - Paired 2. If multiple columns are specified in both Paired 1 Variable(s) and Paired 2 Variable(s), the first columns in each box are compared, then the second columns in each box are compared, and so on.

Reports Tab

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

Descriptive Statistics and Confidence Intervals

Confidence Level

This confidence level is used for the descriptive statistics confidence intervals of each group, as well as for the confidence interval of the mean difference. Typical confidence levels are 90%, 95%, and 99%, with 95% being the most common.

Descriptive Statistics and Confidence Intervals of Each Group

This section reports the group name, count, mean, standard deviation, standard error, and confidence interval of the mean for each group.

Confidence Interval of the Mean Difference

This section reports the confidence interval for the difference between the two means based on the paired differences.

Limits

Specify whether a two-sided or one-sided confidence interval of the mean difference is to be reported.

- **Two-Sided**

  For this selection, the lower and upper limits of the mean difference are reported, giving a confidence interval of the form (Lower Limit, Upper Limit).

- **One-Sided Upper**

  For this selection, only an upper limit of the mean difference is reported, giving a confidence interval of the form (-∞, Upper Limit).

- **One-Sided Lower**
  For this selection, only a lower limit of the mean difference is reported, giving a confidence interval of the form (Lower Limit, ∞).

Bootstrap Confidence Intervals

This section provides confidence intervals of desired levels for the mean difference. A bootstrap distribution histogram may be specified on the Plots tab.

Bootstrap Confidence Levels

These are the confidence levels of the bootstrap confidence intervals. All values must be between 50 and 100. You may enter several values, separated by blanks or commas. A separate confidence interval is generated for each value entered.

Examples:

90 95 99

90:99(1)

90

Samples (N)

This is the number of bootstrap samples used. A general rule of thumb is to use at least 100 when standard errors are the focus or 1000 when confidence intervals are your focus. With current computer speeds, 10,000 or more samples can usually be run in a short time.

Retries

If the results from a bootstrap sample cannot be calculated, the sample is discarded and a new sample is drawn in its place. This parameter is the number of times that a new sample is drawn before the algorithm is terminated. It is possible that the sample cannot be bootstrapped. This parameter lets you specify how many alternative bootstrap samples are considered before the algorithm gives up. The range is 1 to 1000, with 50 as a recommended value.

Bootstrap Confidence Interval Method

This option specifies the method used to calculate the bootstrap confidence intervals.

- **Percentile**
  The confidence limits are the corresponding percentiles of the bootstrap values.

- **Reflection**
  The confidence limits are formed by reflecting the percentile limits. If $X0$ is the original value of the parameter estimate and XL and XU are the percentile confidence limits, the Reflection interval is ($2\,X0 - XU$, $2\,X0 - XL$).

Percentile Type

This option specifies which of five different methods is used to calculate the percentiles. More details of these five types are offered in the Descriptive Statistics chapter.

Confidence Interval of σ

This section provides one- or two-sided confidence limits for σ. These limits rely heavily on the assumption of normality of the population from which the sample is taken.

Tests

Alpha

Alpha is the significance level used in the hypothesis tests. A value of 0.05 is most commonly used, but 0.1, 0.025, 0.01, and other values are sometimes used. Typical values range from 0.001 to 0.20.

H0 Value

This is the hypothesized value of the population mean difference under the null hypothesis. This is the mean difference to which the sample difference will be compared.

Alternative Hypothesis Direction

Specify the direction of the alternative hypothesis. This direction applies to t-test, power report, and Wilcoxon Signed-Rank test. The Randomization test results given in this procedure are always two-sided tests. The Quantile (Sign) test results given in this procedure are always the two-sided and one-sided tests together. The selection 'Two-Sided and One-Sided' will produce all three tests for each test selected.

Tests - Parametric

T-Test

This report provides the results of the common paired-sample T-Test.

Power Report for T-Test

This report gives the power of the paired-sample T-Test when it is assumed that the population mean difference and standard deviation or differences are equal to the sample mean difference and standard deviation of differences.

Tests - Nonparametric

Randomization Test

A randomization test is conducted by first determining the signs of all the paired differences relative to the null hypothesized mean difference – that is, the signs of the paired differences after subtracting the null hypothesized mean difference. Then all possible permutations of the signs are enumerated. Each of these permutations is assigned to absolute values of the original subtracted values in their original order, and the corresponding t-statistic is calculated. The original t-statistic, based on the original data, is compared to each of the permutation t-statistics, and the total count of permutation t-statistics more extreme than the original t-statistic is determined. Dividing this count by the number of permutations tried gives the significance level of the test.

For even moderate sample sizes, the total number of permutations is in the trillions, so a Monte Carlo approach is used in which the permutations are found by random selection rather than complete enumeration. Edgington suggests that at least 1,000 permutations be selected. We suggest that this be increased to 10,000.

Monte Carlo Samples

Specify the number of Monte Carlo samples used when conducting randomization tests. You also need to check the ‘Randomization Test’ box under the Variables tab to run this test.

Somewhere between 1000 and 100000 Monte Carlo samples are usually necessary. Although the default is 1000, we suggest the use of 10000 when using this test.

Quantile (Sign) Test

The quantile (sign) test tests whether the null hypothesized quantile difference is the quantile for the quantile test proportion.

Each of the values in the sample is compared to the null hypothesized quantile difference and determined whether it is larger or smaller. Values equal to the null hypothesized difference are removed. The binomial distribution based on the Quantile Test Proportion is used to determine the probabilities in each direction of obtaining such a sample or one more extreme if the null quantile is the true one. These probabilities are the p-values for the one-sided test of each direction. The two-sided p-value is twice the value of the smaller of the two one-sided p-values.

When the quantile of interest is the median (the quantile test proportion is 0.5), this quantile test is called the sign test.

Quantile Test Proportion

This is the value of the binomial proportion used in the Quantile test. The quantile (sign) test tests whether the null hypothesized quantile difference is the quantile for this proportion. A value of 0.5 results in the Sign Test.

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 ≥ 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 ≥ 10.

Assumptions

Tests of Assumptions

This report gives a series of four tests of the normality assumptions based on the paired differences.

These tests are:

- **Shapiro-Wilk Normality**
- **Skewness Normality**
- **Kurtosis Normality**
- **Omnibus Normality**

The Correlation Coefficient is also given.

Assumptions Alpha

Assumptions Alpha is the significance level used in all the assumptions tests. A value of 0.05 is typically used for hypothesis tests in general, but values other than 0.05 are often used for the case of testing assumptions. Typical values range from 0.001 to 0.20.

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

Decimals

This option allows the user to specify the number of decimal places directly or based on the significant digits. These decimals are used for output except the nonparametric tests and bootstrap results, where the decimal places are defined internally.

If one of the significant digits 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

1.314583689 is displayed as 1.314584

The output formatting system is not designed to accommodate 'Significant Digits (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.

Bootstrap Confidence Interval Histograms

This plot requires that the bootstrap confidence intervals report has been selected. It gives the plot of the bootstrap distribution used in creating the bootstrap confidence interval.

Example 1 – Running a Paired T-Test

This section presents an example of how to run a paired t-test as well as other paired comparisons. The data for this example are the tire data shown above and are found in the Tire dataset. The two columns containing the paired data are ***RtTire*** and ***LtTire***.

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 window.

1 Open the Tire dataset.

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

2 Open the Paired T-Test window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Paired T-Test** 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 **Paired 1 Variable(s)** text box. This will bring up the variable selection window.
- Select **RtTire** from the list of variables and then click **Ok**. "RtTire" will appear in the Paired 1 Variables box.
- Double-click in the **Paired 2 Variable(s)** text box. This will bring up the variable selection window.
- Select **LtTire** from the list of variables and then click Ok. "LtTire" will appear in the Paired 2 Variables box.

4 Specify the reports.

- Select the **Reports tab**.
- Leave the **Confidence Level** at **95%**.
- Make sure the **Descriptive Statistics and Confidence Intervals of Each Group** check box is checked.
- Make sure the **Confidence Interval of the Mean Difference** check box is checked.
- Leave the Confidence Interval **Limits** at **Two-Sided**.
- Check **Bootstrap Confidence Intervals**. Leave the sub-options at the default values.
- Leave **Alpha** at **0.05**.
- Leave the **H0 Value** at **0.0**.
- For **Ha,** select **Two-Sided and One-Sided**. Usually a single alternative hypothesis is chosen, but all three alternatives are shown in this example to exhibit all the reporting options.
- Make sure the **T-Test** check box is checked.
- Check the **Power Report for T-Test** check box.
- Check the **Randomization Test** check box. Leave the Monte Carlo Samples at 10000.
- Check the **Quantile (Sign) Test** check box. Leave the Quantile Test Proportion at 0.5.
- Check the **Wilcoxon Signed-Rank Test** check box. Leave all sub-boxes checked.
- Make sure the **Test of Assumptions** check box is checked. Leave the **Assumptions Alpha** at **0.05**.

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 Section

Descriptive Statistics

Variable	Count	Mean	Standard Deviation of Data	Standard Error of Mean	T*	95.0% LCL of Mean	95.0% UCL of Mean
RtTire	14	50.5	13.96011	3.730996	2.1604	42.43967	58.56033
LtTire	14	52.57143	13.7657	3.679038	2.1604	44.62335	60.51951

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 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.

UCL of the Mean

This is the upper limit of the interval estimate for the mean based on a t distribution with n - 1 degrees of freedom.

Two-Sided Confidence Interval of the Mean Difference

Two-Sided Confidence Interval of the Mean Difference

Statistic	Count	Mean Difference	Standard Deviation	Standard Error	T*	DF	95.0% C. I. of Mean Diff Lower Limit	95.0% C. I. of Mean Diff Upper Limit
Mean Difference	14	-2.071429	5.225151	1.39648	2.1604	13	-5.088341	0.9454835

Statistic

This shows the statistic displayed in the report.

Count

This is the number of non-missing values.

Mean Difference

This is the average of the paired differences.

Standard Deviation

This is the sample standard deviation of the paired differences.

Standard Error

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

T*

This is the t-value used to construct the confidence limits. It is based on the degrees of freedom and the confidence level.

DF

The degrees of freedom are used to determine the T distribution from which T* is generated:

$$df = n - 1$$

Lower and Upper Confidence Limits

These are the confidence limits of the confidence interval for the mean difference. The confidence interval formula is

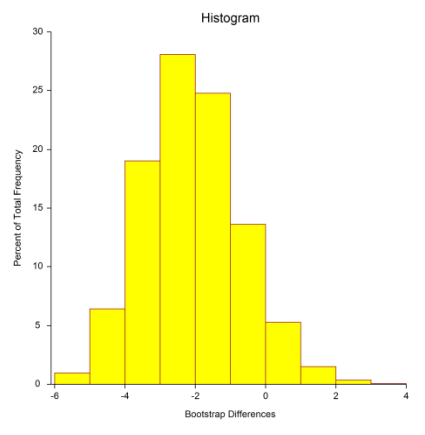
$$\overline{x} \pm T^*_{df} \cdot SE_{\overline{x}}$$

Bootstrap Section

Estimation Results		Bootstrap Confidence Limits		
Parameter	**Estimate**	**Conf. Level**	**Lower**	**Upper**
Mean				
Original Value	-2.0714	90.00	-4.4286	0.0714
Bootstrap Mean Difference	-2.0754	95.00	-5.0000	0.5000
Bias (BMD - OV)	-0.0040	99.00	-5.8571	1.1429
Bias Corrected	-2.0675			
Standard Error	1.3590			

Sampling Method = Observation, Confidence Limit Type = Reflection, Number of Samples = 3000.

Bootstrap Histograms Section



This report provides bootstrap confidence intervals of the mean difference. Note that since these results are based on 3000 random bootstrap samples, they will differ slightly from the results you obtain when you run this report.

Original Value

This is the parameter estimate obtained from the complete sample without bootstrapping.

Bootstrap Mean Difference

This is the average of the parameter estimates of the bootstrap samples.

Bias (BMD - OV)

This is an estimate of the bias in the original estimate. It is computed by subtracting the original value from the bootstrap mean difference.

Bias Corrected

This is an estimated of the parameter that has been corrected for its bias. The correction is made by subtracting the estimated bias from the original parameter estimate.

Standard Error

This is the bootstrap method's estimate of the standard error of the parameter estimate. It is simply the standard deviation of the parameter estimate computed from the bootstrap estimates.

Conf. Level

This is the confidence coefficient of the bootstrap confidence interval given to the right.

Bootstrap Confidence Limits – Lower and Upper

These are the limits of the bootstrap confidence interval with the confidence coefficient given to the left. These limits are computed using the confidence interval method (percentile or reflection) designated on the Bootstrap panel.

Note that to be accurate, these intervals must be based on over a thousand bootstrap samples and the original sample must be representative of the population.

Bootstrap Histogram

The histogram shows the distribution of the bootstrap parameter estimates.

T-Test Section

This section presents the results of the traditional paired-sample T-test. Here, reports for all three alternative hypotheses are shown, but a researcher would typically choose one of the three before generating the output. All three tests are shown here for the purpose of exhibiting all the output options available.

Paired-Sample T-Test

Alternative Hypothesis	Mean Difference	Standard Error	T-Statistic	DF	Prob Level	Reject H0 at α = 0.050?
Mean Diff. ≠ 0	-2.071429	1.39648	-1.4833	13	0.16182	No
Mean Diff. < 0	-2.071429	1.39648	-1.4833	13	0.08091	No
Mean Diff. > 0	-2.071429	1.39648	-1.4833	13	0.91909	No

Alternative Hypothesis

The (unreported) null hypothesis is

$$H_0: \mu_{diff} = 0$$

and the alternative hypotheses,

$$H_a: \mu_{diff} \neq 0 ,$$
$$H_a: \mu_{diff} < 0 , \text{ or}$$
$$H_a: \mu_{diff} > 0 .$$

In practice, the alternative hypothesis should be chosen in advance.

Mean Difference

This is the average of the paired differences.

Standard Error

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

$$SE_{\overline{x}_{diff}} = \frac{s_{diff}}{\sqrt{n}}$$

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:

$$T - Statistic = \frac{\overline{x}_{diff} - Hypothesized\ Value}{SE_{\overline{x}_{diff}}}$$

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 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.

Reject H0 at α = 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 α. A test in which the null hypothesis is rejected is sometimes called *significant*.

Power for the Paired T-Test

The power report gives the power of a test where it is assumed that the population mean difference and standard deviation is equal to the sample mean difference and standard deviation. Powers are given for alpha values of 0.05 and 0.01. For a much more comprehensive and flexible investigation of power or sample size, we recommend you use the PASS software program.

Power for the Paired-Sample T-Test
This section assumes the population mean of paired differences and standard deviation of paired differences are equal to the sample values.

Alternative Hypothesis	N	μ	σ	Power (α = 0.05)	Power (α = 0.01)
Mean Diff. ≠ 0	14	-2.071429	5.225151	0.27964	0.10154
Mean Diff. < 0	14	-2.071429	5.225151	0.40555	0.16041
Mean Diff. > 0	14	-2.071429	5.225151	0.00112	0.00012

Alternative Hypothesis

This value identifies the test direction of the test reported in this row. In practice, you would select the alternative hypothesis prior to your analysis and have only one row showing here.

N

N is the assumed sample size.

μ

This is the assumed population mean difference on which the power calculation is based.

σ

This is the assumed population standard deviation on which the power calculation is based.

Power (α = 0.05) and Power (α = 0.01)

Power is the probability of rejecting the hypothesis that the mean difference is equal to the hypothesized value when they are in fact not equal. Power is one minus the probability of a type II error (β). The power of the test depends on the sample size, the magnitude of the standard deviation, the alpha level, and the true difference between the population mean difference and the hypothesized value.

The power value calculated here assumes that the population standard deviation is equal to the sample standard deviation and that the difference between the population mean difference and the hypothesized value is exactly equal to the difference between the sample mean difference and the hypothesized value.

High power is desirable. High power means that there is a high probability of rejecting the null hypothesis when the null hypothesis is false.

Some ways to increase the power of a test include the following:

1. Increase the alpha level. Perhaps you could test at α = 0.05 instead of α = 0.01.
2. Increase the sample size.
3. Decrease the magnitude of the standard deviation. Perhaps the study can be redesigned so that measurements are more precise and extraneous sources of variation are removed.

Randomization Test Section

A randomization test is conducted by first determining the signs of all the mean differences relative to the null hypothesized mean difference – that is, the signs of the differences after subtracting the null hypothesized mean difference, which is usually zero. Then all possible permutations of the signs are enumerated. Each of these permutations is assigned to absolute values of the original subtracted differences in their original order, and the corresponding t-statistic is calculated. The original t-statistic, based on the original differences, is compared to each of the permutation t-statistics, and the total count of permutation t-statistics more extreme than the original t-statistic is determined. Dividing this count by the number of permutations tried gives the significance level of the test.

For even moderate sample sizes, the total number of permutations is in the trillions, so a Monte Carlo approach is used in which the permutations are found by random selection rather than complete enumeration. Edgington suggests that at least 1,000 permutations by selected. We suggest that this be increased to 10,000.

Randomization Test (Two-Sided)
Alternative Hypothesis: The distribution center of paired differences is not 0.
Number of Monte Carlo samples: 10000

Test	Prob Level	Reject H0 at α = 0.050?
Randomization Test	0.16360	No

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 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.

Reject H0 at α = (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 α. A test in which the null hypothesis is rejected is sometimes called *significant*.

based upon the sampling distribution of the statistic being normal under the alternative hypothesis.

Quantile (Sign) Test Section

The quantile (sign) test is one of the older nonparametric procedures. Each of the paired differences in the sample is compared to the null hypothesized quantile value and determined whether it is larger or smaller. Paired differences equal to the null hypothesized value are removed. The binomial distribution based on the Quantile Test Proportion is used to determine the probabilities in each direction of obtaining such a sample or one more extreme if the null quantile is the true one. These probabilities are the *p*-values for the one-sided test of each direction. The two-sided *p*-value is twice the value of the smaller of the two one-sided *p*-values.

When the quantile of interest is the median (the quantile test proportion is 0.5), a quantile test is called the *sign test*.

While the quantile (sign) test is simple, there are more powerful nonparametric alternatives, such as the Wilcoxon signed-rank test. However, if the shape of the underlying distribution of a variable is the double exponential distribution, the sign test may be the better choice.

Quantile (Sign) Test
This Quantile test is equivalent to the Sign test if the Quantile Proportion is 0.5.

Null Quantile (Q0)	Quantile Proportion	Number Lower	Number Higher	H1: Q ≠ Q0 Prob Level	H1: Q < Q0 Prob Level	H1: Q > Q0 Prob Level
0	0.5	10	4	0.179565	0.089783	0.971313

Null Quantile (Q0)

Under the null hypothesis, the proportion of all paired differences below the null quantile is the quantile proportion. For the sign test, the null quantile is the null median.

Quantile Proportion

Under the null hypothesis, the quantile proportion is the proportion of all paired differences below the null quantile. For the sign test, this proportion is 0.5.

Number Lower

This is the actual number of paired differences that are below the null quantile.

Number Higher

This is the actual number of paired differences that are above the null quantile.

H1:Q ≠ Q0 Prob Level

This is the two-sided probability that the true quantile is equal to the stated null quantile (Q0), for the quantile proportion stated and given the observed values. A small *p*-value indicates that the true quantile for the stated quantile proportion is different from the null quantile.

H1:Q < Q0 Prob Level

This is the one-sided probability that the true quantile is greater than or equal to the stated null quantile (Q0), for the quantile proportion stated and given the observed values. A small *p*-value indicates that the true quantile for the stated quantile proportion is less than the null quantile.

H1:Q > Q0 Prob Level

This is the one-sided probability that the true quantile is less than or equal to the stated null quantile (Q0), for the quantile proportion stated and given the observed values. A small *p*-value indicates that the true quantile for the stated quantile proportion is greater than the null quantile.

Wilcoxon Signed-Rank Test Section

This nonparametric test makes use of the sign and the magnitude of the rank of the differences (original paired differences minus the hypothesized value). It is typically the best nonparametric alternative to the paired-sample t-test.

Wilcoxon Signed-Rank Test

Sum of Ranks (W)	Mean of W	Std Dev of W	Number of Zeros	Number Sets of Ties	Multiplicity Factor
21	52.5	15.84692	0	3	126

Test Type	Alternative Hypothesis	Z-Value	Prob Level	Reject H0 at α = 0.050?
Exact*	Median Diff. ≠ 0			
Exact*	Median Diff. < 0			
Exact*	Median Diff. > 0			
Normal Approximation	Median Diff. ≠ 0	1.9878	0.04684	Yes
Normal Approximation	Median Diff. < 0	-1.9878	0.02342	Yes
Normal Approximation	Median Diff. > 0	-1.9878	0.97658	No
Normal Approx. with C.C.	Median Diff. ≠ 0	1.9562	0.05044	No
Normal Approx. with C.C.	Median Diff. < 0	-1.9562	0.02522	Yes
Normal Approx. with C.C.	Median Diff. > 0	-2.0193	0.97827	No

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

Sum of Ranks (W)

The basic statistic for this test is the sum of the positive ranks, ΣR_+ (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.

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^{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 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. In the two-tail test for the median difference (assuming a hypothesized value of 0), the null hypothesis would be H_0: Median = 0 with the alternative being H_a: Median ≠ 0.

The left-tail alternative is represented by Median < 0 (i.e., H_a: median < 0) while the right-tail alternative is depicted by Median > 0.

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 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 (α = 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 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 (α = 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.

Tests of Assumptions Section

Tests of Assumptions Section

Assumption	Value	Probability	Decision (α = 0.050)
Shapiro-Wilk Normality	0.9103	0.159217	Cannot reject normality
Skewness Normality	1.3651	0.172212	Cannot reject normality
Kurtosis Normality	1.9065	0.056589	Cannot reject normality
Omnibus Normality	5.4982	0.063985	Cannot reject normality
Correlation Coefficient	0.929062		

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, skewness, kurtosis, or omnibus normality tests and visually by the histogram or normal probability plot.

In the case of non-normality, one of the nonparametric tests may be considered. Sometimes a transformation, such as the natural logarithm or the square root of the original data, can change the underlying distribution from skewed to normal. To evaluate whether the underlying distribution of the variable is normal after the transformation, rerun the normal probability plot on the transformed variable. If some of the data values are negative or zero, it may be necessary to add a constant to the original data prior to the transformation. If the transformation produces a suitable result, then the paired-sample t-test could be performed on the transformed data.

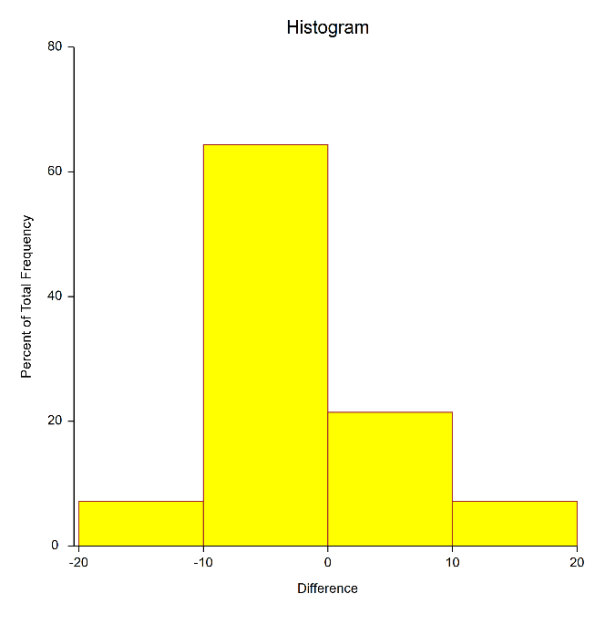
Normality (Shapiro-Wilk, Skewness, Kurtosis, and Omnibus)

These four tests allow you to test the skewness, kurtosis, and overall normality of the data. If any of them reject the hypothesis of normality, the data should not be considered normal. These tests are discussed in more detail in the Descriptive Statistics chapter.

Correlation Coefficient

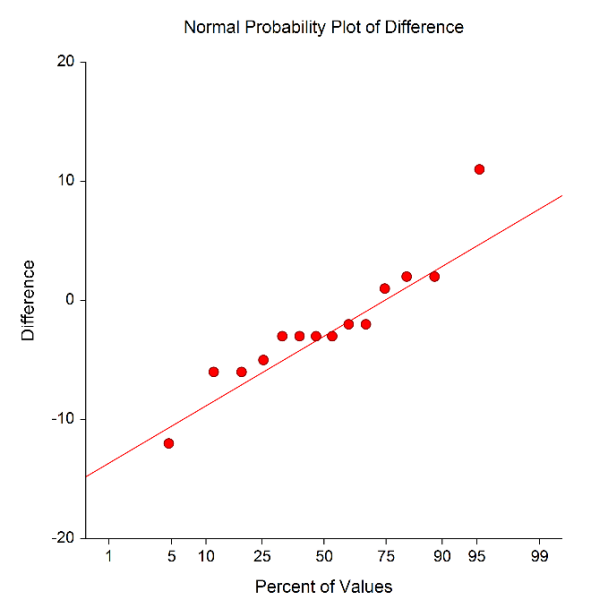
This is the correlation of the two paired variables.

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.

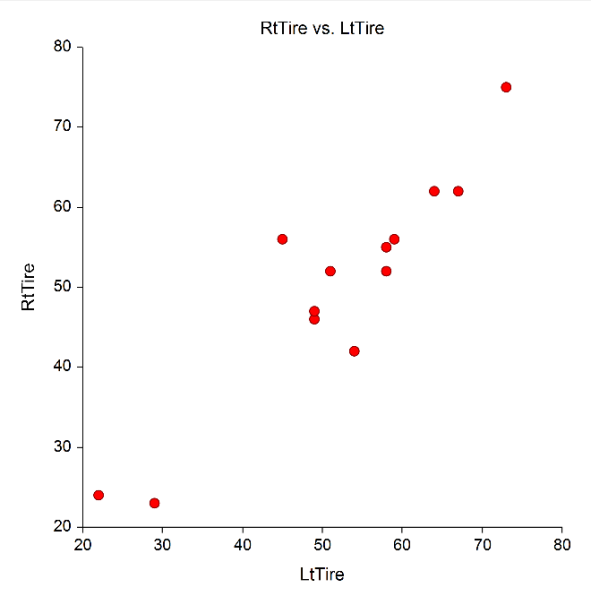
Probability Plot of Differences



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.

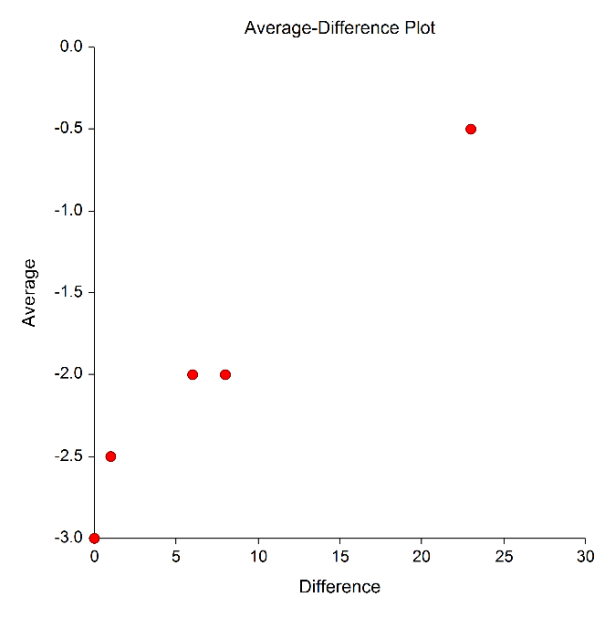
Note that the confidence bands are based on large-sample formulas. They may not be accurate for small samples.

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 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 paired 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.

Paired T-Test Checklist

This checklist, prepared by a professional statistician, is a flowchart of the steps you should complete to conduct a valid paired-sample t-test (or one of its nonparametric counterparts). You should complete these tasks in order.

Step 1 – Data Preparation

Introduction

This step involves scanning your data for anomalies, data entry errors, typos, and so on. Frequently we hear of people who completed an analysis with the right techniques but obtained strange conclusions because they had mistakenly selected the data.

Sample Size

The sample size (number of nonmissing rows) has a lot of ramifications. The larger the sample size the better. Of course, the t-test may be performed on very small samples, say 4 or 5 observations, but it is impossible to assess the validity of assumptions with such small samples. It is our statistical experience that at least 20 observations

are necessary to evaluate normality properly. On the other hand, since skewness can have unpleasant effects on t-tests with small samples, particularly for one-tailed tests, larger sample sizes (30 to 50) may be necessary.

It is possible to have a sample size that is too large for a statistical significance test. When your sample size is very large, you are almost guaranteed to find statistical significance. However, the question that then arises is whether the magnitude of the difference is of practical importance.

Missing Values

The number and pattern of missing values is always an issue to consider. Usually, we assume that missing values occur at random throughout your data. If this is not true, your results will be biased since a particular segment of the population is underrepresented. If you have a lot of missing values, some researchers recommend comparing other variables with respect to missing versus nonmissing. If you find large differences in other variables, you should begin to worry about whether the missing values might cause a systematic bias in your results.

Type of Data

The mathematical basis of the t-test assumes that the data are continuous. Because of the rounding that occurs when data are recorded, all data are technically discrete. The validity of assuming the continuity of the data then comes down to determining when we have too much rounding. For example, most statisticians would not worry about human-age data that was rounded to the nearest year. However, if these data were rounded to the nearest ten years or further to only three groups (young, adolescent, and adult), most statisticians would question the validity of the probability statements. Some studies have shown that the t-test is reasonably accurate when the data has only five possible values (most would call this discrete data). If your data contains less than five unique values, any probability statements made are tenuous.

Outliers

Generally, outliers cause distortion in statistical tests. You must scan your data for outliers (the box plot is an excellent tool for doing this). If you have outliers, you have to decide if they are one-time occurrences or if they would occur in another sample. If they are one-time occurrences, you can remove them and proceed. If you know they represent a certain segment of the population, you have to decide between biasing your results (by removing them) or using a nonparametric test that can deal with them. Most would choose the nonparametric test.

Step 2 – Setup and Run the Panel

Introduction

NCSS is designed to be simple to operate, but it requires some learning. When you go to run a procedure such as this for the first time, take a few minutes to read through the chapter again and familiarize yourself with the issues involved.

Enter Variables

The NCSS panels are set with ready-to-run defaults. About all you have to do is select the appropriate variables.

Select All Plots

As a rule, you should select all diagnostic plots (histograms, etc.). They add a great deal to your analysis of the data.

Specify Alpha

Most beginners in statistics forget this important step and let the alpha value default to the standard 0.05. You should make a conscious decision as to what value of alpha is appropriate for your study. The 0.05 default came about when people had to rely on printed probability tables in which there were only two values available: 0.05 or 0.01. Now you can set the value to whatever is appropriate.

Step 3 – Check Assumptions

Introduction

Once the program output is displayed, you will be tempted to go directly to the probability of the t-test, determine if you have a significant result, and proceed to something else. However, it is very important that you proceed through the output in an orderly fashion. The first task is to determine which of the assumptions are met by your data.

Sometimes, when the data are nonnormal, a data transformation (like square roots or logs) might normalize the data. Frequently, this kind of transformation or re-expression approach works very well. However, always check the transformed variable to see if it is normally distributed.

It is not unusual in practice to find a variety of tests being run on the same basic null hypothesis. That is, the researcher who fails to reject the null hypothesis with the first test will sometimes try several others and stop when the hoped-for significance is obtained. For instance, a statistician might run the one-sample t-test on the original data, the one-sample t-test on the logarithmically transformed data, the Wilcoxon signed-rank test, and the Quantile test. An article by Gans (1984) suggests that there is no harm on the true significance level if no more than two tests are run. This is not a bad option in the case of questionable outliers. However, as a rule of thumb, it seems more honest to investigate whether the data is normal. The conclusion from that investigation should direct you to the right test.

Random Sample

The validity of this assumption depends on the method used to select the sample. If the method used ensures that each individual in the population of interest has an equal probability of being selected for this sample, you have a random sample. Unfortunately, you cannot tell if a sample is random by looking at either it or statistics from it.

Check Descriptive Statistics

You should check the Descriptive Statistics Section first to determine if the Count and the Mean are reasonable. If you have selected the wrong variables, these values will alert you.

Normality

To validate this assumption, you would first look at the plots. Outliers will show up on the box plots and the probability plots. Skewness, kurtosis, more than one mode, and a host of other problems will be obvious from the density trace on the histogram. After considering the plots, look at the Tests of Assumptions Section to get numerical confirmation of what you see in the plots. Remember that the power of these normality tests is directly related to the sample size, so when the normality assumption is accepted, double-check that your sample is large enough to give conclusive results (at least 20).

Symmetry

The nonparametric tests need the assumption of symmetry. The easiest ways to evaluate this assumption are from the density trace on the histogram or from the average-difference plot.

Step 4 – Choose the Appropriate Statistical Test

Introduction

After understanding how your data fit the assumptions of the various one-sample tests, you are ready to determine which statistical procedures will be valid. You should select one of the following three situations based on the status of the normality.

Normal Data

Use the T-Test Section for hypothesis testing and the Descriptive Statistics Section for interval estimation.

Nonnormal and Asymmetrical Data

Try a transformation, such as the natural logarithm or the square root, on the original data since these transformations frequently change the underlying distribution from skewed to normal. If some of the data values are negative or zero, add a constant to the original data prior to the transformation. If the transformed data is now normal, use the T-Test Section for hypothesis testing and the Descriptive Statistics Section for interval estimation.

Nonnormal and Symmetrical Data

Use the Wilcoxon Signed-Rank Test or the Quantile Test for hypothesis testing.

Step 5 – Interpret Findings

Introduction

You are now ready to conduct your test. Depending on the nature of your study, you should look at either of the following sections.

Hypothesis Testing

Here you decide whether to use a two-tailed or one-tailed test. The two-tailed test is the standard. If the probability level is less than your chosen alpha level, reject the null hypothesis of equality to a specified mean (or median) and conclude that the mean is different. Your next task is to look at the mean itself to determine if the size of the difference is of practical interest.

Confidence Limits

The confidence limits let you put bounds on the size of the mean (for one independent sample) or mean difference (for dependent samples). If these limits are narrow and close to your hypothesized value, you might determine that even though your results are statistically significant, there is no practical significance.

Step 6 – Record Your Results

Finally, as you finish a test, take a moment to jot down your impressions. Explain what you did, why you did it, what conclusions you reached, which outliers you deleted, areas for further investigation, and so on. Since this is a technical process, your short-term memory will not retain these details for long. These notes will be worth their weight in gold when you come back to this study a few days later!

Example of Paired T-Test Steps

This example will illustrate the use of one-sample tests for paired data. A new four-lane road is going through the west end of a major metropolitan area. About 150 residential properties will be affected by the road. A random sample of 15 properties was selected. These properties were evaluated by two different property assessors. We are interested in determining whether there is any difference in their assessment. The assessments are recorded in thousands of dollars and are shown in the table. The assessment values are represented by Value1 and Value2 for the two property assessors (Property Value dataset).

Value1	Value2
118.5	117.1
154.2	159.6
130.8	136.5
154.8	146.9
131.4	136.0
104.1	99.7
154.9	157.8
97.6	96.1
140.0	144.8
116.9	112.4
129.6	129.1
108.2	114.5
108.6	113.7
178.3	194.3
92.9	87.6

Step 1 – Data Preparation

These data are paired measurements. The sample size is smaller than you would like, but it is 10% of the current population. There are no missing values, and the use of the dollar value makes the data continuous.

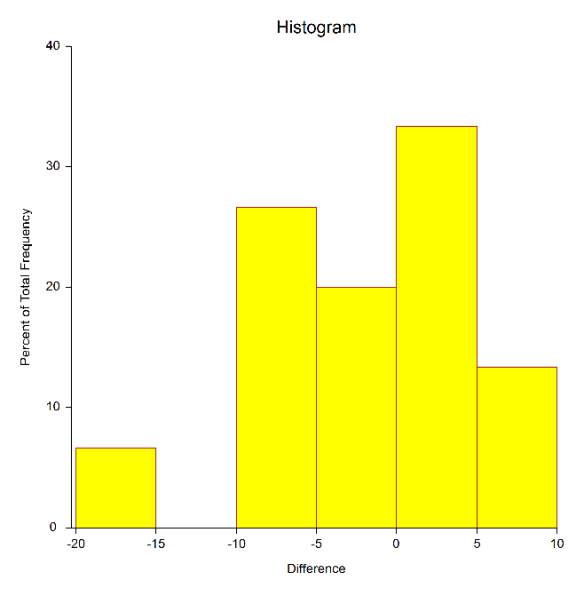
Step 2 – Setup and Run the Paired T-Test Panel

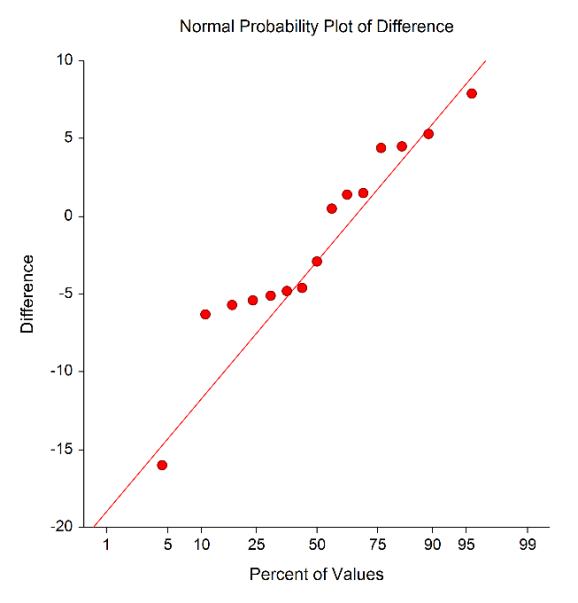
The selection and running of the Paired T-Test from the Analysis menu on the pairs of assessments, Value1 and Value2, would produce the output that follows. The alpha value has been set at 0.05. Interpretation of the results will come in the steps to follow.

Step 3 – Check Assumptions

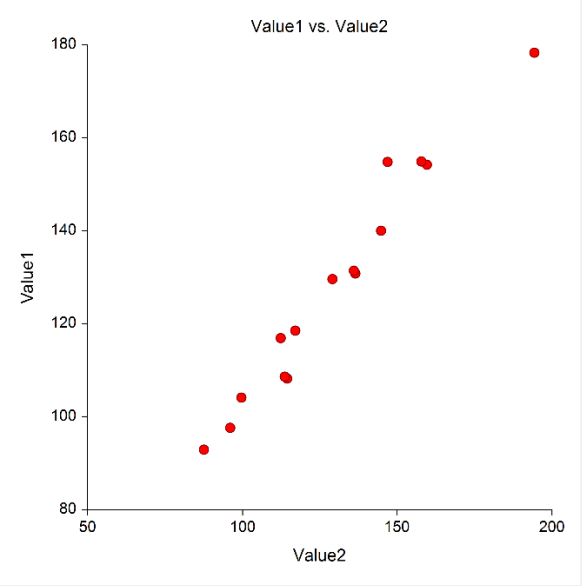
The major assumption to check for is normality. We begin with the graphic perspectives: normal probability plots, histograms, density traces, and box plots. Since this is paired data, we look at the normality of the differences.

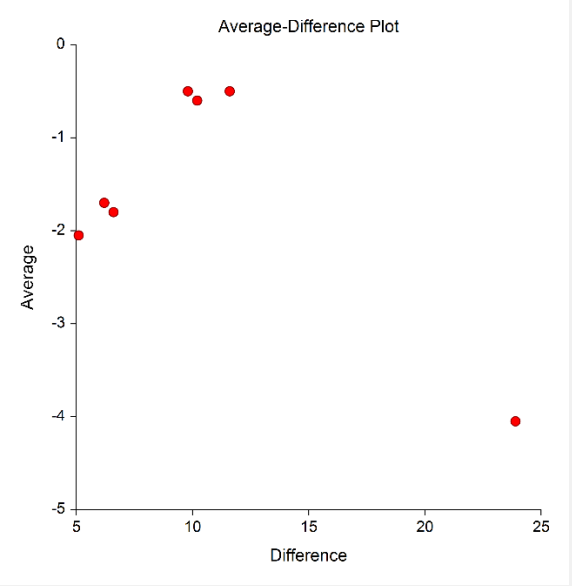
Histogram and Normal Probability Plot





Scatter Plot and Average-Difference Plot





The normal probability plot on the differences indicates normality, except for an outlier on the low side. While the histogram and density trace are not good tools for evaluating normality on small samples, they do show the left skewness created by this one observation. This observation could be an outlier. Of course, a larger sample size would have been a definite advantage for the histogram and density trace, but normality seems to be valid (we make ourselves a note to check up on this outlier).

In evaluating normality by numerical measures, look at the Probability (*p*-value) and the Decision for the given alpha of 0.05. Investigation of the Tests of Assumptions Section confirms that the differences in assessment are normal by all three normality tests since the p-values are greater than 0.05. In fact, the p-values are much greater than 0.05. The “Cannot reject normality” under Decision ($\alpha = 0.05$) is the formal conclusion of the normality tests.

Tests of Assumptions

Assumption	Value	Prob Level	Decision (α = 0.050)
Shapiro-Wilk Normality	0.9328	0.300128	Cannot reject normality
Skewness Normality	-0.9490	0.342635	Cannot reject normality
Kurtosis Normality	0.7722	0.440019	Cannot reject normality
Omnibus Normality	1.4968	0.473127	Cannot reject normality
Correlation Coefficient	0.982357		

From the scatter plot above, it is evident that there is a strong positive linear relationship between the two assessments, as also confirmed by the Pearson correlation of 0.9824.

Step 4 – Choose the Appropriate Statistical Test

In Step 3, the conclusions from checking the assumptions were three-fold: (1) the data are continuous, (2) the differences are normally distributed, and (3) there is a strong positive relationship between the two assessments. As a result of these findings, the appropriate statistical test is the paired t-test, which is shown next.

Descriptive Statistics

Variable	Count	Mean	Standard Deviation	Standard Error	95% LCL of Mean	95% UCL of Mean
Value1	15	128.0533	24.68883	6.374629	114.3811	141.7256
Value2	15	129.74	28.30113	7.307321	114.0674	145.4126

T* for Confidence Limits: T* (Value1) = 2.1448, T* (Value2) = 2.1448

Paired-Sample T-Test

Alternative Hypothesis	Mean Difference	Standard Error	T-Statistic	DF	Prob Level	Reject H0 at α = 0.050?
Mean Diff. ≠ 0	-1.686667	1.585436	-1.0639	14	0.30540	No

Step 5 – Interpret Findings

In the Descriptive Statistics Section, the mean difference is -$1.687 thousand with the standard deviation of differences being $6.140 thousand. The 95% interval estimate for the mean difference ranges from -$5.087 thousand to $1.714 thousand.

The formal two-tail hypothesis test for this example is shown under the T-Test Section. The p-value for this two-tail test is 0.30540, which is much greater than 0.05. Thus, the conclusion of this hypothesis test is acceptance, i.e., there is no difference in the assessments.

Remember when checking the assumption of normality, we noted that there was one possible outlier in the normal probability plot in the output. If we had run the Wilcoxon Signed-Rank test instead of the paired t-test, the p-value would be 0.30280. Hence, the conclusion is the same--- there is no difference between assessments. This kind of decision confirmation does not always happen, but it is a simple option on questionable assumption situations. However, since the data are normally distributed, the paired t-test was the correct statistical test to choose.

Wilcoxon Signed-Rank Test

Sum of Ranks (W)	Mean of W	Std Dev of W	Number of Zeros	Number Sets of Ties	Multiplicity Factor
41	60	17.60682	0	0	0

Test Type	Alternative Hypothesis	Z-Value	Prob Level	Reject H0 at 0.050?
Exact*	Median Diff. ≠ 0		0.30280	No
Normal Approximation	Median Diff. ≠ 0	1.0791	0.28053	No
Normal Approx. with C.C.	Median Diff. ≠ 0	1.0507	0.29338	No

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

Step 6 – Record Your Results

The conclusions for this example are that there is no difference between assessors for residential properties evaluated in this area, according to the paired t-test. The Wilcoxon Signed-Rank gave the same conclusion. If you were troubled by the one outlier, you could use a transformation on the differences plus a constant and rerun the paired t-test. Or, further examination of the one outlier might reveal extenuating circumstances that confirm that this is a one-time anomaly. If that were the case, the observation could be omitted and the analysis redone.