Chapter 206

Two-Sample T-Test

Introduction

This procedure provides several reports for the comparison of two continuous-data distributions, including confidence intervals for the difference in means, two-sample t-tests, the z-test, the randomization test, the Mann-Whitney U (or Wilcoxon Rank-Sum) nonparametric test, and the Kolmogorov-Smirnov test. Tests of assumptions and plots are also available in this procedure.

The data for this procedure can be contained in two variables (columns) or in one variable indexed by a second (grouping) variable.

Research Questions

A common task in research is to compare two populations or groups. Researchers may wish to compare the income level of two regions, the nitrogen content of two lakes, or the effectiveness of two drugs. An initial question that arises is what aspects (parameters) of the populations should be compared. We might consider comparing the averages, the medians, the standard deviations, the distributional shapes, or maximum values. The comparison parameter is based on the particular problem.

The typical comparison of two distributions is the comparison of means. If we can safely make the assumption of the data in each group following a normal distribution, we can use a two-sample t-test to compare the means of random samples drawn from these two populations. If these assumptions are severely violated, the nonparametric Mann-Whitney U test, the randomization test, or the Kolmogorov-Smirnov test may be considered instead.

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 and Confidence Intervals of Each Group
- Confidence Interval of $\mu_1 - \mu_2$
- Bootstrap Confidence Intervals
- Equal-Variance T-Test and associated power report
- Aspin-Welch Unequal-Variance T-Test and associated power report
- Z-Test
- Randomization Test
- Mann-Whitney Test
- Kolmogorov-Smirnov Test
- Tests of Assumptions
- Graphs

Data Structure

The data may be entered in two formats, as shown in the two examples below. The examples give the yield of corn for two types of fertilizer. The first format is shown in the first table in which the responses for each group are entered in separate variables. That is, each variable contains all responses for a single group. In the second format the data are arranged so that all responses are entered in a single variable. A second variable, the Grouping Variable, contains an index that gives the group (A or B) to which the row of data belongs.

In most cases, the second format is more flexible. Unless there is some special reason to use the first format, we recommend that you use the second.

Two Response Variables

Yield A	Yield B
452	546
874	547
554	774
447	465
356	459
754	665
558	467
574	365
664	589
682	534
547	456
435	651
245	654
	665
	546
	537

Grouping and Response Variables

Fertilizer	Yield
B	546
B	547
B	774
B	465
B	459
B	665
B	456
.	.
.	.
A	452
A	874
A	554
A	447
A	356
A	754
A	558
A	574
A	664
.	.
.	.

Null and Alternative Hypotheses

For comparing two means, the basic null hypothesis is that the means are equal,

$$H_0: \mu_1 = \mu_2$$

with three common alternative hypotheses,

$$H_a: \mu_1 \neq \mu_2 ,$$

$$H_a: \mu_1 < \mu_2 , \text{ or}$$

$$H_a: \mu_1 > \mu_2 ,$$

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

A slightly different set of null and alternative hypotheses are used if the goal of the test is to determine whether μ_1 or μ_2 is greater than or less than the other by a given amount.

The null hypothesis then takes on the form

$$H_0: \mu_1 - \mu_2 = \textit{Hypothesized Difference}$$

and the alternative hypotheses,

$$H_a: \mu_1 - \mu_2 \neq \textit{Hypothesized Difference}$$

$$H_a: \mu_1 - \mu_2 < \textit{Hypothesized Difference}$$

$$H_a: \mu_1 - \mu_2 > \textit{Hypothesized Difference}$$

These hypotheses are equivalent to the previous set if the *Hypothesized Difference* is zero.

Assumptions

The following assumptions are made by the statistical tests described in this section. One of the reasons for the popularity of the t-test, particularly the Aspin-Welch Unequal-Variance t-test, is its robustness in the face of assumption violation. However, if an assumption is not met even approximately, the significance levels and the power of the t-test are invalidated. Unfortunately, in practice it sometimes happens that one or more assumption is not met. Hence, take the appropriate 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.

Two-Sample T-Test Assumptions

The assumptions of the two-sample t-test are:

1. The data are continuous (not discrete).
2. The data follow the normal probability distribution.
3. The variances of the two populations are equal. (If not, the Aspin-Welch Unequal-Variance test is used.)
4. The two samples are independent. There is no relationship between the individuals in one sample as compared to the other (as there is in the paired t-test).
5. Both samples are simple random samples from their respective populations. Each individual in the population has an equal probability of being selected in the sample.

Mann-Whitney U Test Assumptions

The assumptions of the Mann-Whitney U test are:

1. The variable of interest is continuous (not discrete). The measurement scale is at least ordinal.
2. The probability distributions of the two populations are identical, except for location.
3. The two samples are independent.
4. Both samples are simple random samples from their respective populations. Each individual in the population has an equal probability of being selected in the sample.

Kolmogorov-Smirnov Test Assumptions

The assumptions of the Kolmogorov-Smirnov test are:

1. The measurement scale is at least ordinal.
2. The probability distributions are continuous.
3. The two samples are mutually independent.
4. Both samples are simple random samples from their respective populations.

Example 1 – Comparing Two Groups

This section presents an example of how to run a comparison of two groups. We will use the corn yield data found in YldA and YldB of the Yield2 dataset.

Setup

To run this example, complete the following steps:

1 Open the Yield2 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Yield2** and click **OK**.

2 Specify the Two-Sample T-Test procedure options

- Find and open the **Two-Sample T-Test** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Data Input Type	**Two Variables with Response Data in each Variable**
Group 1 Variable	**YldA**
Group 2 Variable	**YldB**
Reports Tab	
Confidence Level	**95**
Descriptive Statistics …	**Checked**
Confidence Interval of μ1 - μ2	**Checked**
Limits	**Two-Sided**
Bootstrap Confidence Intervals	**Checked**
Confidence Levels	**90 95 99**
Samples (N)	**3000**
C.I. Method	**Reflection**
Retries	**50**
Percentile Type	**Ave X(p[n+1])**
Confidence Intervals of Each…	**Checked**
Confidence Interval of σ1/σ2	**Checked**
Alpha	**0.05**
H0 μ1 - μ2 =	**0.0**
Ha	**Two-Sided and One-Sided**
Equal-Variance T-Test	**Checked**
Unequal-Variance T-Test	**Checked**
Z-Test	**Checked**
σ1	**150**
σ2	**110**
Power Report for Equal-Variance T-Test	**Checked**
Power Report for Unequal-Variance T-Test	**Checked**
Randomization Test	**Checked**
Monte Carlo Samples	**10000**

Mann-Whitney U Test… ..... **Checked**
Exact Test ..... **Checked**
Normal Approximation Test ..... **Checked**
Normal Approximation Test with… ..... **Checked**
Kolmogorov-Smirnov Test ..... **Checked**
Tests of Assumptions ..... **Checked**
Assumptions Alpha **0.05**

Plots Tab
All Plots ..... **Checked**

Report Options (*in the Toolbar*)
Variable Labels ..... **Column Names**

3 Run the procedure

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

Descriptive Statistics Section

This section gives a descriptive summary of each group. See the Descriptive Statistics chapter for details about this section.

This report can be used to confirm that the Means and Counts are as expected.

Descriptive Statistics

Variable	Count	Mean	Standard Deviation of Data	Standard Error of Mean	T*	95% LCL of Mean	95% UCL of Mean
YldA	13	549.3846	168.7629	46.80641	2.1788	447.4022	651.367
YldB	16	557.5	104.6219	26.15546	2.1314	501.7509	613.249

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.

Descriptive Statistics for the Median Section

This section gives the medians and the confidence intervals for the medians of each of the groups.

Descriptive Statistics for the Median

Variable	Count	Median	95% LCL of Median	95% UCL of Median
YldA	13	554	435	682
YldB	16	546	465	651

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*.

Median

The median is the 50th percentile of the group data, using the AveXp(n+1) method. The details of this method are described in the Descriptive Statistics chapter under Percentile Type.

LCL and UCL

These are the lower and upper confidence limits of the median. These limits are exact and make no distributional assumptions other than a continuous distribution. No limits are reported if the algorithm for this interval is not able to find a solution. This may occur if the number of unique values is small.

Two-Sided Confidence Interval for μ1 – μ2 Section

Given that the assumptions of independent samples and normality are valid, this section provides an interval estimate (confidence limits) of the difference between the two means. Results are given for both the equal and unequal variance cases. You can use the Tests of Assumptions section to assess the assumption of equal variance.

Two-Sided Confidence Interval for μ1 - μ2

Variance Assumption	DF	Mean Difference	Standard Error	T*	95% LCL of Difference	95% UCL of Difference
Equal	27	-8.115385	51.11428	2.0518	-112.9932	96.76247
Unequal	19.17	-8.115385	53.61855	2.0918	-120.2734	104.0426

DF

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

For the equal variance case:

$$df = n_1 + n_2 - 2$$

For the unequal variance case:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Mean Difference

This is the difference between the sample means, $\overline{X}_1 - \overline{X}_2$.

Standard Error

This is the estimated standard deviation of the distribution of differences between independent sample means.

For the equal variance case:

$$SE_{\overline{X}_1 - \overline{X}_2} = \sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

For the unequal variance case:

$$SE_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

T*

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

Lower and Upper Confidence Limits

These are the confidence limits of the confidence interval for $\mu_1 - \mu_2$. The confidence interval formula is

$$\overline{X}_1 - \overline{X}_2 \pm T_{df}^* \cdot SE_{\overline{X}_1 - \overline{X}_2}$$

The equal-variance and unequal-variance assumption formulas differ by the values of T* and the standard error.

Bootstrap Section

Bootstrapping was developed to provide standard errors and confidence intervals in situations in which the standard assumptions are not valid. The method is simple in concept, but it requires extensive computation.

In bootstrapping, the sample is assumed to be the population and *B* samples of size N1 are drawn from the original group one dataset and N2 from the original group 2 dataset, with replacement. *With replacement sampling* means that each observation is placed back in the population before the next one is selected so that each observation may be selected more than once. For each bootstrap sample, the means and their difference are computed and stored.

Suppose that you want the standard error and a confidence interval of the difference. The bootstrap sampling process provides *B* estimates of the difference. The standard deviation of these *B* differences is the bootstrap estimate of the standard error of the difference. The bootstrap confidence interval is found by arranging the *B*

values in sorted order and selecting the appropriate percentiles from the list. For example, a 90% bootstrap confidence interval for the difference is given by fifth and ninety-fifth percentiles of the bootstrap difference values.

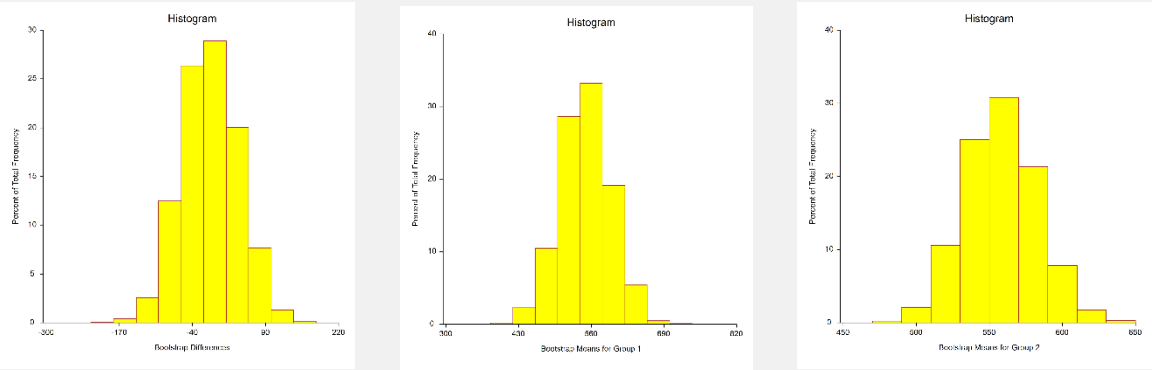
The main assumption made when using the bootstrap method is that your sample approximates the population fairly well. Because of this assumption, bootstrapping does not work well for small samples in which there is little likelihood that the sample is representative of the population. Bootstrapping should only be used in medium to large samples.

Bootstrap Confidence Intervals

Estimation Results		Bootstrap Confidence Limits		
Parameter	**Estimate**	**Conf. Level**	**Lower**	**Upper**
Difference				
Original Value	-8.1154	90.00	-96.5356	77.7144
Bootstrap Mean	-7.1700	95.00	-109.7418	93.1466
Bias (BM - OV)	0.9454	99.00	-144.8579	114.0953
Bias Corrected	-9.0608			
Standard Error	52.1560			
Mean 1				
Original Value	549.3846	90.00	472.5385	623.2269
Bootstrap Mean	550.3126	95.00	458.3096	640.3750
Bias (BM - OV)	0.9279	99.00	429.6154	661.2300
Bias Corrected	548.4567			
Standard Error	45.5424			
Mean 2				
Original Value	557.5000	90.00	515.5656	599.1250
Bootstrap Mean	557.4825	95.00	505.5656	607.4969
Bias (BM - OV)	-0.0175	99.00	489.5044	620.9359
Bias Corrected	557.5175			
Standard Error	25.3045			

Confidence Interval Method = Reflection, Number of Samples = 3000.
Computer-Generated Random Seed = 3285290

Bootstrap Histograms



This report provides bootstrap confidence intervals of the two means and their 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

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

Bias (BM - OV)

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

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.

Note that these intervals should 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.

Confidence Intervals of Standard Deviations

Confidence Intervals of Standard Deviations

					95% C. I. of σ	
Sample	N	Mean	Standard Deviation	Standard Error	Lower Limit	Upper Limit
1	13	549.3846	168.7629	46.80641	121.0175	278.5829
2	16	557.5	104.6219	26.15546	77.28468	161.9223

This report gives a confidence interval for the standard deviation in each group. Note that the usefulness of these intervals is very dependent on the assumption that the data are sampled from a normal distribution.

Using the common notation for sample statistics (see, for example, ZAR (1984) page 115), a $100(1-\alpha)\%$ confidence interval for the standard deviation is given by

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}} \le \sigma \le \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}}$$

Confidence Interval for Standard Deviation Ratio

Confidence Interval for Standard Deviation Ratio (σ1 / σ2)

					95% C. I. of σ1 / σ2	
N1	N2	SD1	SD2	SD1 / SD2	Lower Limit	Upper Limit
13	16	168.7629	104.6219	1.613075	0.9370615	2.875259

This report gives a confidence interval for the ratio of the two standard deviations. Note that the usefulness of these intervals is very dependent on the assumption that the data are sampled from a normal distribution.

Using the common notation for sample statistics (see, for example, ZAR (1984) page 125), a $100(1-\alpha)\%$ confidence interval for the ratio of two standard deviations is given by

$$\frac{s_1}{s_2\sqrt{F_{1-\alpha/2,n_1-1,n_2-1}}} \le \frac{\sigma_1}{\sigma_2} \le \frac{s_1\sqrt{F_{1-\alpha/2,n_2-1,n_1-1}}}{s_2}$$

Equal-Variance T-Test Section

This section presents the results of the traditional equal-variance 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.

Equal-Variance T-Test

Alternative Hypothesis	Mean Difference	Standard Error	T-Statistic	DF	Prob Level	Reject H0 at α = 0.050?
μ1 - μ2 ≠ 0	-8.115385	51.11428	-0.1588	27	0.87503	No
μ1 - μ2 < 0	-8.115385	51.11428	-0.1588	27	0.43752	No
μ1 - μ2 > 0	-8.115385	51.11428	-0.1588	27	0.56248	No

Alternative Hypothesis

The (unreported) null hypothesis is

$$H_0: \mu_1 - \mu_2 = \textit{Hypothesized Difference} = 0$$

and the alternative hypotheses,

$$H_a: \mu_1 - \mu_2 \neq \textit{Hypothesized Difference} = 0$$
$$H_a: \mu_1 - \mu_2 < \textit{Hypothesized Difference} = 0$$
$$H_a: \mu_1 - \mu_2 > \textit{Hypothesized Difference} = 0$$

Since the *Hypothesized Difference* is zero in this example, the null and alternative hypotheses can be simplified to

Null hypothesis:

$$H_0: \mu_1 = \mu_2$$

Alternative hypotheses:

$$H_a: \mu_1 \neq \mu_2 \text{ ,}$$
$$H_a: \mu_1 < \mu_2 \text{ , or}$$
$$H_a: \mu_1 > \mu_2 \text{ .}$$

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

Mean Difference

This is the difference between the sample means, $\overline{X}_1 - \overline{X}_2$.

Standard Error

This is the estimated standard deviation of the distribution of differences between independent sample means.

The formula for the standard error of the difference in the equal-variance T-test is:

$$SE_{\overline{X}_1-\overline{X}_2} = \sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

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:

$$T - Statistic = \frac{\overline{X}_1 - \overline{X}_2 - Hypothesized\ Difference}{SE_{\overline{X}_1 - \overline{X}_2}}$$

In this instance, the hypothesized difference is zero, so the T-Statistic formula reduces to

$$T - Statistic = \frac{\overline{X}_1 - \overline{X}_2}{SE_{\overline{X}_1 - \overline{X}_2}}$$

DF

The degrees of freedom define the T *distribution* upon which the probability values are based. The formula for the degrees of freedom in the equal-variance T-test is:

$$df = n_1 + n_2 - 2$$

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 Equal-Variance T-Test

The power report gives the power of a test where it is assumed that the population means and standard deviations are equal to the sample means and standard deviations. 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 Equal-Variance T-Test
This section assumes the population means and standard deviations are equal to the sample values.

Alternative Hypothesis	N1	N2	μ1	μ2	σ1	σ2	Power (α = 0.05)	Power (α = 0.01)
μ1 - μ2 ≠ 0	13	16	549.3846	557.5	168.7629	104.6219	0.05269	0.01084
μ1 - μ2 < 0	13	16	549.3846	557.5	168.7629	104.6219	0.06811	0.01480
μ1 - μ2 > 0	13	16	549.3846	557.5	168.7629	104.6219	0.03595	0.00662

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.

N1 and N2

N1 and N2 are the assumed sample sizes for groups 1 and 2.

μ1 and μ2

These are the assumed population means on which the power calculations are based.

σ1 and σ2

These are the assumed population standard deviations on which the power calculations are based.

Power (α = 0.05) and Power (α = 0.01)

Power is the probability of rejecting the hypothesis that the means are equal 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 magnitudes of the standard deviations, the alpha level, and the true difference between the two population means.

The power value calculated here assumes that the population standard deviations are equal to the sample standard deviations and that the difference between the population means is exactly equal to the difference between the sample means.

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 alpha = 0.05 instead of alpha = 0.01.
2. Increase the sample size.
3. Decrease the magnitude of the standard deviations. Perhaps the study can be redesigned so that measurements are more precise and extraneous sources of variation are removed.

Aspin-Welch Unequal-Variance T-Test Section

This section presents the results of the T-test where equal variance is not assumed. Reports for all three alternative hypotheses are shown here, 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.

Aspin-Welch Unequal-Variance T-Test

Alternative Hypothesis	Mean Difference	Standard Error	T-Statistic	DF	Prob Level	Reject H0 at α = 0.050?
μ1 - μ2 ≠ 0	-8.115385	53.61855	-0.1514	19.17	0.88128	No
μ1 - μ2 < 0	-8.115385	53.61855	-0.1514	19.17	0.44064	No
μ1 - μ2 > 0	-8.115385	53.61855	-0.1514	19.17	0.55936	No

Alternative Hypothesis

The (unreported) null hypothesis is

$$H_0: \mu_1 - \mu_2 = Hypothesized\ Difference = 0$$

and the alternative hypotheses,

$$H_a: \mu_1 - \mu_2 \neq Hypothesized\ Difference = 0$$

$$H_a: \mu_1 - \mu_2 < Hypothesized\ Difference = 0$$

$$H_a: \mu_1 - \mu_2 > Hypothesized\ Difference = 0$$

Since the *Hypothesized Difference* is zero in this example, the null and alternative hypotheses can be simplified to

Null hypothesis:

$$H_0: \mu_1 = \mu_2$$

Alternative hypotheses:

$$H_a: \mu_1 \neq \mu_2 ,$$
$$H_a: \mu_1 < \mu_2 , \text{ or}$$
$$H_a: \mu_1 > \mu_2 .$$

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

Mean Difference

This is the difference between the sample means, $\overline{X}_1 - \overline{X}_2$.

Standard Error

This is the estimated standard deviation of the distribution of differences between independent sample means.

The formula for the standard error of the difference in the Aspin-Welch unequal-variance T-test is:

$$SE_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

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:

$$T - Statistic = \frac{\overline{X}_1 - \overline{X}_2 - Hypothesized\ Difference}{SE_{\overline{X}_1 - \overline{X}_2}}$$

In this instance, the hypothesized difference is zero, so the T-Statistic formula reduces to

$$T - Statistic = \frac{\overline{X}_1 - \overline{X}_2}{SE_{\overline{X}_1 - \overline{X}_2}}$$

DF

The degrees of freedom define the T *distribution* upon which the probability values are based. The formula for the degrees of freedom in the Aspin-Welch unequal-variance T-test is:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

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 Aspin-Welch Unequal-Variance T-Test

The power report gives the power of a test where it is assumed that the population means and standard deviations are equal to the sample means and standard deviations. 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 Aspin-Welch Unequal-Variance T-Test
This section assumes the population means and standard deviations are equal to the sample values.

Alternative Hypothesis	N1	N2	μ1	μ2	σ1	σ2	Power (α = 0.05)	Power (α = 0.01)
μ1 - μ2 ≠ 0	13	16	549.3846	557.5	168.7629	104.6219	0.05238	0.01072
μ1 - μ2 < 0	13	16	549.3846	557.5	168.7629	104.6219	0.06697	0.01444
μ1 - μ2 > 0	13	16	549.3846	557.5	168.7629	104.6219	0.03665	0.00680

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.

N1 and N2

N1 and N2 are the assumed sample sizes for groups 1 and 2.

μ1 and μ2

These are the assumed population means on which the power calculations are based.

σ1 and σ2

These are the assumed population standard deviations on which the power calculations are based.

Power (α = 0.05) and Power (α = 0.01)

Power is the probability of rejecting the hypothesis that the means are equal 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 magnitudes of the standard deviations, the alpha level, and the true difference between the two population means.

The power value calculated here assumes that the population standard deviations are equal to the sample standard deviations and that the difference between the population means is exactly equal to the difference between the sample means.

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 alpha = 0.05 instead of alpha = 0.01.
2. Increase the sample size.
3. Decrease the magnitude of the standard deviations. Perhaps the study can be redesigned so that measurements are more precise and extraneous sources of variation are removed.

Z-Test Section

This section presents the results of the two-sample Z-test, which is used when the population standard deviations are known. Because the population standard deviations are rarely known, this test is not commonly used in practice. The Z-test is included in this procedure for the less-common case of known standard deviations, and for two-sample hypothesis test training. In this example, 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.

Z-Test

Alternative Hypothesis	Mean Difference	σ1	σ2	Standard Error	Z-Statistic	Prob Level	Reject H0 at α = 0.050?
μ1 - μ2 ≠ 0	-8.115385	150	110	49.87002	-0.1627	0.87073	No
μ1 - μ2 < 0	-8.115385	150	110	49.87002	-0.1627	0.43537	No
μ1 - μ2 > 0	-8.115385	150	110	49.87002	-0.1627	0.56464	No

Alternative Hypothesis

The (unreported) null hypothesis is

$$H_0: \mu_1 - \mu_2 = \textit{Hypothesized Difference} = 0$$

and the alternative hypotheses,

$$H_a: \mu_1 - \mu_2 \neq \textit{Hypothesized Difference} = 0$$

$$H_a: \mu_1 - \mu_2 < \textit{Hypothesized Difference} = 0$$

$$H_a: \mu_1 - \mu_2 > \textit{Hypothesized Difference} = 0$$

Since the *Hypothesized Difference* is zero in this example, the null and alternative hypotheses can be simplified to

Null hypothesis:

$$H_0: \mu_1 = \mu_2$$

Alternative hypotheses:

$$H_a: \mu_1 \neq \mu_2 \text{ ,}$$

$$H_a: \mu_1 < \mu_2 \text{ , or}$$

$$H_a: \mu_1 > \mu_2 \text{ .}$$

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

Mean Difference

This is the difference between the sample means, $\overline{X}_1 - \overline{X}_2$.

σ1 and σ2

These are the known Group 1 and Group 2 population standard deviations.

Standard Error of Difference

This is the estimated standard deviation of the distribution of differences between independent sample means.

The formula for the standard error of the difference in the Z-test is:

$$SE_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Z-Statistic

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

$$Z - Statistic = \frac{\overline{X}_1 - \overline{X}_2 - Hypothesized\ Difference}{SE_{\overline{X}_1 - \overline{X}_2}}$$

In this instance, the hypothesized difference is zero, so the Z-Statistic formula reduces to

$$Z - Statistic = \frac{\overline{X}_1 - \overline{X}_2}{SE_{\overline{X}_1 - \overline{X}_2}}$$

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*.

Randomization Tests Section

The Randomization test is a non-parametric test for comparing two distributions. A randomization test is conducted by enumerating all possible permutations of the groups while leaving the data values in the original order. The appropriate statistic (here, each of the two *t*-statistics) is calculated for each permutation and the number of permutations that result in a statistic with a magnitude greater than or equal to the statistic is counted. 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 (1987) suggests that at least 1,000 permutations be selected. We suggest that current computer speeds permit the number of permutations selected to be increased to 10,000 or more.

Randomization Tests
Alternative Hypothesis: |μ1 - μ2| ≠ 0. This is a Two-Sided Test.
Number of Monte Carlo samples: 10000
Computer-Generated Random Seed: 3285290

Variance Assumption	Prob Level	Reject H0 at α = 0.050?
Equal Variance	0.87000	No
Unequal Variance	0.87700	No

Variance Assumption

The variance assumption distinguishes which T-statistic formula is used to create the distribution of T-statistics. If the equal variance assumption is chosen the T-statistics are based on the formula:

$$T - Statistic = \frac{\overline{X}_1 - \overline{X}_2 - Hypothesized\ Difference}{\sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

If the unequal variance assumption is chosen, the T-statistics are based on the formula:

$$T - Statistic = \frac{\overline{X}_1 - \overline{X}_2 - Hypothesized\ Difference}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

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. The Prob Level for the randomization test is calculated by finding the proportion of the permutation t-statistics that are more extreme than the original data t-statistic. 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*. For a randomization test, a significant *p*-value does not necessarily indicate a difference in means, but instead an effect of some kind from a difference in group populations.

Mann-Whitney U or Wilcoxon Rank-Sum Test for Difference in Location

This test is the most common nonparametric substitute for the *t*-test when the assumption of normality is not valid. The assumptions for this test were given in the Assumptions section at the beginning of this chapter. Two key assumptions are that the distributions are at least ordinal in nature and that they are identical, except for location.

When ties are present in the data, an approximate solution for dealing with ties is available, you can use the approximation provided, but know that the exact results no longer hold.

This particular test is based on ranks and has good properties (asymptotic relative efficiency) for symmetric distributions. There are exact procedures for this test given small samples with no ties, and there are large sample approximations.

Mann-Whitney U or Wilcoxon Rank-Sum Test for Difference in Location

Variable	Mann-Whitney U	Sum of Ranks (W)	Mean of W	Std Dev of W
YldA	101.5	192.5	195	22.79508
YldB	106.5	242.5	240	22.79508

Number of Sets of Ties = 3, Multiplicity Factor = 18

Variable

This is the name for each sample, group, or treatment.

Mann-Whitney U

The Mann-Whitney test statistic, *U*, is defined as the total number of times an observation in one group is preceded by an observation in the other group in the ordered configuration of combined samples (Gibbons, 1985). It can be found by examining the observations in a group one-by-one, counting the number of observations in the other group that have smaller rank, and then finding the total. Or it can be found by the formula

$$U_1 = W_1 - \frac{n_1(n_1+1)}{2}$$

for Group 1, and

$$U_2 = W_2 - \frac{n_2(n_2+1)}{2}$$

for Group 2, where W_1 is the sum of the ranks in sample 1, and W_2 is the sum of the ranks in sample 2.

Sum of Ranks (W)

The *W* statistic is obtained by combining all observations from both groups, assigning ranks, and then summing the ranks,

$$W_1 = \sum ranks_1, W_2 = \sum ranks_2$$

Tied values are resolved by using the average rank of the tied values.

Mean of W

This is the mean of the distribution of *W*, formulated as follows:

$$\overline{W}_1 = \frac{n_1(n_1+n_2+1)}{2}$$

and

$$\overline{W}_2 = \frac{n_2(n_1+n_2+1)}{2}$$

Std Dev of W

This is the standard deviation of the *W* corrected for ties. If there are no ties, this standard deviation formula simplifies since the second term under the radical is zero.

$$\sigma_W = \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12} - \frac{n_1 n_2 \sum_{i=1}(t_i^3 - t_i)}{12(n_1+n_2)(n_1+n_2-1)}}$$

where t_1 is the number of observations tied at value one, t_2 is the number of observations tied at some value two, and so forth. Generally, this correction for ties in the standard deviation makes little difference unless there is a large number of ties.

Number Sets of Ties

This gives the number of sets of tied values. If there are no ties, this number is zero. A set of ties is two or more observations with the same value. This is used in adjusting the standard deviation for the W.

Multiplicity factor

This is the tie portion of the standard deviation of W, given by

$$\sum_{i=1}(t_i^3 - t_i)$$

Mann-Whitney U or Wilcoxon Rank-Sum Test for Difference in Location (continued)

Test Type	Alternative Hypothesis	Z-Value	Prob Level	Reject H0 at α = 0.050?
Exact*	Location Diff. ≠ 0			
Exact*	Location Diff. < 0			
Exact*	Location Diff. > 0			
Normal Approximation	Location Diff. ≠ 0	-0.1097	0.91267	No
Normal Approximation	Location Diff. < 0	-0.1097	0.45633	No
Normal Approximation	Location Diff. > 0	-0.1097	0.54367	No
Normal Approx. with C.C.	Location Diff. ≠ 0	-0.0877	0.93008	No
Normal Approx. with C.C.	Location Diff. < 0	-0.0877	0.46504	No
Normal Approx. with C.C.	Location Diff. > 0	-0.1316	0.55235	No

* The Exact Test is provided only when there are no ties and the sample size is ≤ 20 in both groups.

Alternative Hypothesis

For the Wilcoxon rank-sum test, the null and alternative hypotheses relate to the equality or non-equality of the central tendency of the two distributions. If a hypothesized difference other than zero is used, the software adds the hypothesized difference to each value of Group 2, and the test is run based on the original Group 1 values and the transformed Group 2 values.

The Alternative Hypothesis identifies the test direction of the test reported in this row. Although all three hypotheses are shown in this example, it is more common (and recommended) that the appropriate alternative hypothesis be chosen in advance.

Exact Probability: Prob Level

This is an exact *p*-value for this statistical test based on the distribution of W. This *p*-value assumes no ties (if ties are detected, this value is left blank). 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. The exact probability value is available for sample sizes up to 38.

Exact Probability: Reject H0 (α = .050)

This is the conclusion reached about the null hypothesis.

Approximation Without Correction: Z-Value

A normal approximation method can be used to obtain a *Z*-value, which can then be compared to the standard normal distribution to obtain the *p*-value. The Z-value is obtained as

$$Z = \frac{W_n - \overline{W}_n}{\sigma_W}$$

where W_n is the sum of ranks for the group with smaller sample size, and $\overline{W}_n$ is the mean corresponding to W_n. The *Z*-value, the *p*-value, and the decision at specified α-level are provided in the table. This method corrects for ties but does not have a continuity correction factor.

Approximation with correction: Z value

This is a normal approximation method that corrects for ties *and* has the correction factor for continuity. The *Z*-value is:

Left-tail test case:

$$Z = \frac{W_n - \overline{W}_n + 0.5}{\sigma_W}$$

Right-tail test case:

$$Z = \frac{W_n - \overline{W}_n - 0.5}{\sigma_W}$$

Two-tail test case:

$$Z = \frac{\left|-\left|W_n - \overline{W}_n\right| + 0.5\right|}{\sigma_W}$$

where W_n is the sum of ranks for the group with smaller sample size, and $\overline{W}_n$ is the mean corresponding to W_n. The Z-value, the p-value, and the decision at specified α-level are provided in the table.

Kolmogorov-Smirnov Test

This is a two-sample test for differences between two samples or distributions. If a statistical difference is found between the distributions of Groups 1 and 2, the test does not indicate the particular cause of the difference. The difference could be due to differences in location (mean), variation (standard deviation), presence of outliers, skewness, kurtosis, number of modes, and so on.

The assumptions for this nonparametric test are: (1) there are two independent random samples; (2) the two population distributions are continuous; and (3) the data are at least ordinal in scale. This test is sometimes preferred over the Wilcoxon sum-rank test when there are a lot of ties. The test statistic is the maximum distance between the empirical distribution functions (EDF) of the two samples.

Kolmogorov-Smirnov Test For Comparing Distributions

Alternative Hypothesis	Largest-Difference Criterion Value	Prob Level	Reject H0 at α = 0.050?
D(1) ≠ D(2)	0.322115	0.346836	No
D(1) < D(2)	0.322115	0.173418	No
D(1) > D(2)	0.177885	0.467425	No

Alternative Hypothesis

The null and alternative hypotheses relate to the equality of the two distribution functions (noted as D(1) or D(2)). This value identifies the test direction of the test reported in this row. In most cases, you would select the null and alternative hypotheses prior to your analysis.

D(1) ≠ D(2). This is the two-tail test case. The null and alternative hypotheses are

H_0: D(1) = D(2), H_a: D(1) ≠ D(2)

D(1) < D(2). This is the left-tail test case. The null and alternative hypotheses are

H_0: D(1) = D(2), H_a: D(1) < D(2)

D(1) > D(2). This is the right-tail test case. The null and alternative hypotheses are

H_0: D(1) = D(2), H_a: D(1) > D(2)

Largest-Difference Criterion Value

This is the maximum difference between the two empirical distribution functions. It is the Kolmogorov-Smirnov test statistic.

Prob Level

This is the p-value for the test. The algorithm for obtaining the two-sided exact probabilities is given in Kim and Jennrich (1973). One-sided p-values are obtained by halving the corresponding two-sided probabilities.

Reject H0 at α = 0.050?

This is the conclusion reached about the null hypothesis.

Tests of Assumptions Section

This section presents the results of tests for checking the normality and equal variance assumptions. Assumptions concerning independence and random sampling are not tested here.

When viewing this report, you can examine the decisions of the column at the right side. Rejections in this decision column indicate evidence of a departure from normality or equal variance. If none of the tests are rejected, these results do not prove that the distributions are normal, and that the variances are equal, but instead that there is not sufficient evidence to suggest otherwise. If the sample sizes are small, say less than 25 per group, the power of these normality and equal-variance tests may be low, and failure to reject may not be a strong assertion of the assumptions being met.

If the Skewness normality test is rejected, it might be possible to use the square root or logarithmic transformation to normalize your data. If normality is rejected for one of the groups, but not the other, you could look at the normal probability plot or box plot for the group that is not normally distributed to see if an outlier or two may have caused the non-normality.

There are differing opinions as to whether a preliminary test for variance equality is proper before deciding which t-test to use. If there is any doubt of the equal-variance assumption, the Aspin-Welch two-sample T-test is a robust choice that performs well in all but the cases of extreme underlying distributions. Our suggestion is to use the equal variance t-test only when the sample sizes are equal or approximately equal and the variances are very similar. In nearly all other cases the unequal variance t-test is preferred. When the sample sizes are different, the most serious situation is when the smaller sample is associated with the larger variance. Conover and others (1981) did extensive simulation involving different distributions, sample sizes, means, and variances; and they found that the modified-Levene test is one of the most robust and powerful tests for equality of variance. Thus, if a preliminary test is to be preferred, the modified-Levene test is recommended.

In the case of non-normality, two nonparametric tests may be suggested. The basic assumptions of independent samples, continuous random variables, and a measurement scale of at least ordinal scale hold for both tests. The Mann-Whitney U / Wilcoxon Rank-Sum test has the additional assumption that the distributions for the two variables are identical (although not necessary normal) in form and shape (i.e., same variance) but differ only in location (i.e., in medians). On the other hand, the Kolmogorov-Smirnov is a general test for differences between two groups. As a general test, it is somewhat sensitive to all kinds of differences between groups or populations and yet not particularly sensitive to any specific type of difference. The Kolmogorov-Smirnov test is a good choice when there are a lot of ties in your data that tends to invalidate the Wilcoxon Rank-Sum test.

Box plots, histograms, and probability plots of the two groups can also be used for examining assumptions. These plots allow you to visually determine if the assumptions of normality (probability plots) and equal variance (box plots) are justified.

Two-Sample T-Test

Tests of the Normality Assumption for YldA

Normality Test	Test Statistic	Prob Level	Reject H0 of Normality at α = 0.050?
Shapiro-Wilk	0.9843	0.99420	No
Skewness	0.2691	0.78785	No
Kurtosis	0.3081	0.75803	No
Omnibus (Skewness or Kurtosis)	0.1673	0.91974	No

Tests of the Normality Assumption for YldB

Normality Test	Test Statistic	Prob Level	Reject H0 of Normality at α = 0.050?
Shapiro-Wilk	0.9593	0.64856	No
Skewness	0.4587	0.64644	No
Kurtosis	0.1291	0.89726	No
Omnibus (Skewness or Kurtosis)	0.2271	0.89267	No

Tests of the Equal Variance Assumption

Equal-Variance Test	Test Statistic	Prob Level	Reject H0 of Equal Variances at α = 0.050?
Variance-Ratio	2.6020	0.08315	No
Modified-Levene	1.9940	0.16935	No

Shapiro-Wilk Normality

This test for normality has been found to be the most powerful test in most situations. It is the ratio of two estimates of the variance of a normal distribution based on a random sample of *n* observations. The numerator is proportional to the square of the best linear estimator of the standard deviation. The denominator is the sum of squares of the observations about the sample mean. The test statistic *W* may be written as the square of the Pearson correlation coefficient between the ordered observations and a set of weights which are used to calculate the numerator. Since these weights are asymptotically proportional to the corresponding expected normal order statistics, *W* is roughly a measure of the straightness of the normal quantile-quantile plot. Hence, the closer *W* is to one, the more normal the sample is.

The probability values for *W* are valid for sample sizes greater than 3. The test was developed by Shapiro and Wilk (1965) for sample sizes up to 20. **NCSS** uses the approximations suggested by Royston (1992) and Royston (1995) which allow unlimited sample sizes. Note that Royston only checked the results for sample sizes up to 5000 but indicated that he saw no reason larger sample sizes should not work. *W* may not be as powerful as other tests when ties occur in your data.

Skewness Normality

This is a skewness test reported by D'Agostino (1990). Skewness implies a lack of symmetry. One characteristic of the normal distribution is that it has no skewness. Hence, one type of non-normality is skewness.

The Value is the test statistic for skewness, while Prob Level is the *p*-value for a two-tailed test for a null hypothesis of normality. If this *p*-value is less than a chosen level of significance, there is evidence of non-normality. Under Decision (α = 0.050), the conclusion about skewness normality is given.

Kurtosis Normality

Kurtosis measures the heaviness of the tails of the distribution. D'Agostino (1990) reported a second normality test that examines kurtosis. The Value column gives the test statistic for kurtosis, and Prob Level is the *p*-value for a two-tail test for a null hypothesis of normality. If this *p*-value is less than a chosen level of significance, there is evidence of kurtosis non-normality. Under Decision (α = 0.050), the conclusion about normality is given.

Omnibus Normality

This third normality test, also developed by D'Agostino (1990), combines the skewness and kurtosis tests into a single measure. Here, as well, the null hypothesis is that the underlying distribution is normally distributed. The definitions for Value, Prob Level, and Decision are the same as for the previous two normality tests.

Variance-Ratio Equal-Variance Test

This equal variance test is based on the ratio of the two sample variances. This variance ratio is assumed to follow an *F* distribution with n_1 - 1 degrees of freedom for the numerator, and n_2 - 1 degrees of freedom for the denominator. The formula for the *F*-statistic is

$$F = \frac{s_1^2}{s_2^2}$$

This variance ratio is shown under the column heading Value. Because this test assumes that the two samples are drawn from normal populations, you should be confident this assumption is met before using this test. The *p*-value (Prob Level) is compared to the level of significance to produce the decision. If the *p*-value is less than the level of significance, we conclude that there is a difference in variances and the Decision is rejection of the null hypothesis of equal variances. If the p-value is greater than the level of significance, there not sufficient evidence to reject equal variances.

Modified-Levene Equal-Variance Test

The Modified-Levene test is more commonly the recommended test for testing equality of variances. Levene's procedure is outlined as follows. First, redefine the variables for each group by taking the absolute value of the difference from the sample median. For the first group, this transformation would be

$$z_{1j} = \left|x_{1j} - Med_1\right|$$

And for the second group,

$$z_{2j} = \left|x_{2j} - Med_2\right|$$

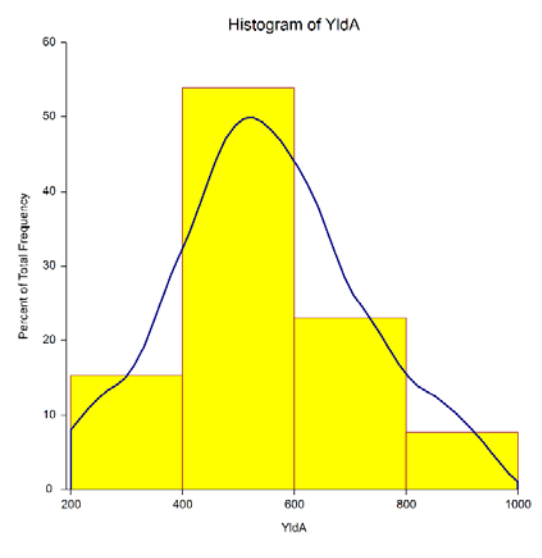
Next, a two-group one-way analysis-of-variance on this redefined variable is run. The *F*-value for this one-way analysis of variance is shown under the column heading Value and its corresponding *p*-value under Prob Level.

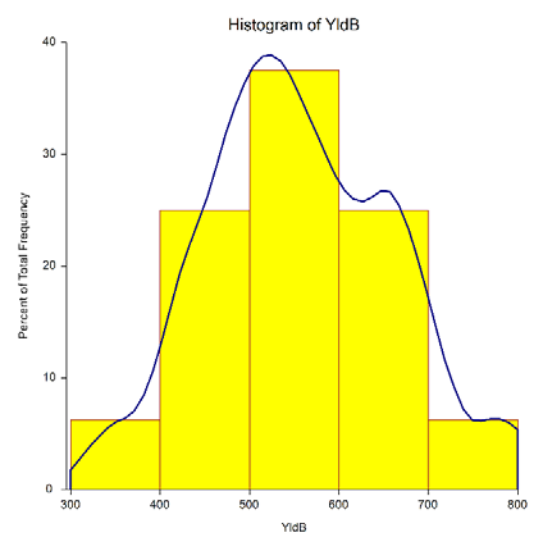
The *p*-value (Prob Level) is compared to the level of significance to give the conclusion concerning the null hypothesis of equal variances. If the *p*-value is less than the level of significance, we can conclude there is a difference in variances and the resulting decision is rejection of the null hypothesis of equal variances. Otherwise, there is not sufficient evidence to reject the null hypothesis of equal variances.

Histograms with Density Trace

The histogram with the density trace overlay (the wavy line) lets you study the distributional features of the two samples to determine if (and which) two-sample tests may be appropriate. When sample sizes are small, and since the shape of the histogram is influenced by the number of classes or bins and the width of the bins, the best choice is often to examine the density trace, which is a smoothed histogram. A complete discussion of histograms is given in the chapter on this topic.

Histograms with Density Trace

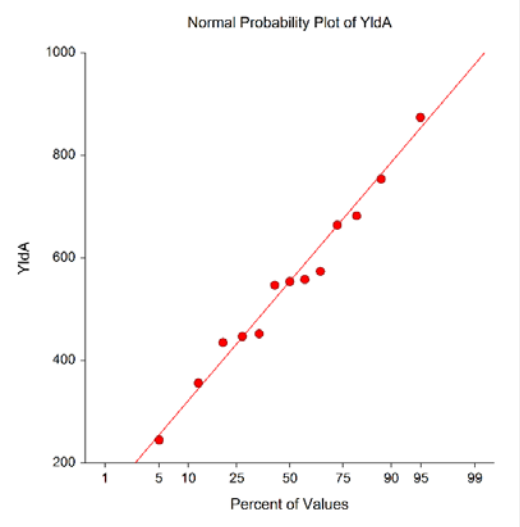


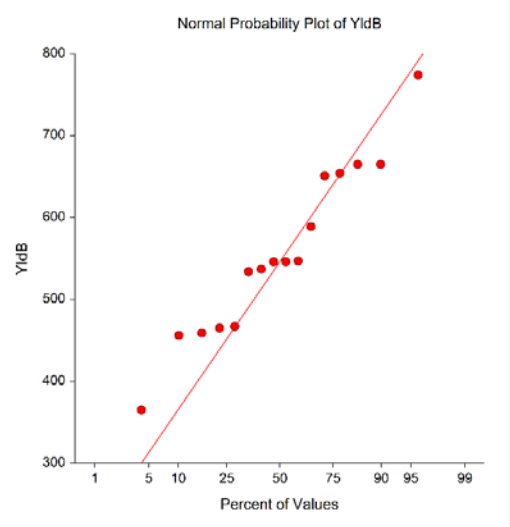


Probability Plots

These plots are used to examine the normality of the groups. The results of the goodness-of-fit tests mentioned earlier, especially the omnibus test, should be reflected in this plot. For more details, you can examine the Probability Plots chapter.

Probability Plots

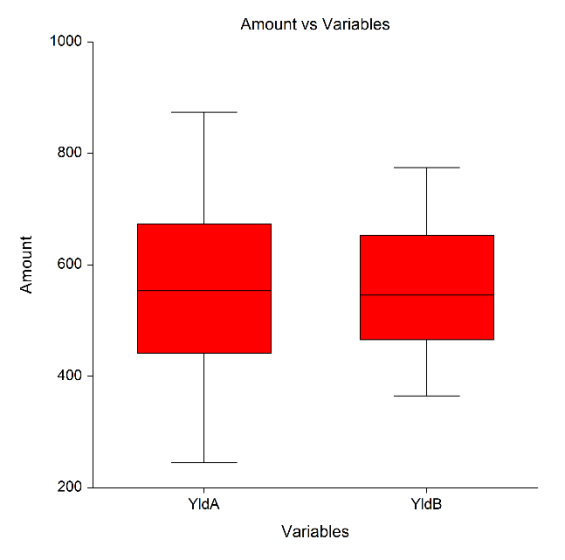




Box Plots

Box plots are useful for assessing symmetry, presence of outliers, general equality of location, and equality of variation.

Box Plots



Two-Sample T-Test Checklist

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

Step 1 – Data Preparation

Introduction

This step involves scanning your data for anomalies, keypunch errors, typos, and so on. You would be surprised how often we hear of people completing an analysis, only to find that they had mistakenly selected the wrong database.

Sample Size

The sample size (number of nonmissing rows) has a lot of ramifications. The equal-variance two-sample t-test was developed under the assumption that the sample sizes of each group would be equal. In practice, this seldom happens, but the closer you can get to equal sample sizes the better.

With regard to the combined sample size, the t-test may be performed on very small samples, say 4 or 5 observations per group. However, in order to test assumptions and obtain reliable estimates of variation, you should attempt to obtain at least 30 individuals per group.

It is possible to have a sample size that is too large. When your sample size is quite 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 are always issues 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 non-missing. If you find large differences in other variables, you should begin to worry about whether the missing values are cause for 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 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 most popular 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 T-Test 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 and familiarize yourself with the issues involved.

Enter Variables

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

Select Reports

The default reports are good initial set, but you should examine whether you want confidence intervals, or non-parametric tests.

Select All Plots

As a rule, you should select all diagnostic plots (box 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 and 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 assumptions are met by your data.

Sometimes, when the data are nonnormal for both samples, a data transformation (like square roots or logs) might normalize the data. Frequently, when only one sample is normal and the other is not, this transformation, or re-expression, approach works well.

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 equal-variance t-test on the original two samples, the equal-variance t-test on the logarithmically transformed data, the Wilcoxon rank-sum test, and the Kolmogorov-Smirnov test. An article by Gans (“The Search for Significance: Different Tests on the Same Data,” *The Journal of Statistical Computation and Simulation*,” 1984, pp. 1-21) 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 are normal. The conclusion from that investigation should direct one to the right test.

Random Sample

The validity of this assumption depends upon the method used to select the sample. If the method used assures 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 it or statistics from it.

Sample Independence

The two samples must be independent. For example, if you randomly divide a group of individuals into two groups, you have met this requirement. However, if your population consists of cars and you assign the left tire to one group and the right tire to the other, you do not have independence. Here again, you cannot tell if the samples are independent by looking at them. You must consider the sampling methodology.

Check Descriptive Statistics

You should check the Individual-Group Statistics Section first to determine if the Count and the Mean are reasonable. If you have selected the wrong variable, 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. No data will be perfectly normal. 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.

Equal Variance

The equal variance assumption is important in determining which statistical test to use. Check the box plots for boxes with about the same widths. Confirm your conclusion by looking at the Equal-Variance Test (Modified Levene) line. Note that, strictly speaking, these equal variance tests require the assumption of normality. If your data are not normal, you should use the modified Levene test. It works in many nonnormal situations.

Some researchers recommend against using a preliminary test on variances (which research and simulations do not strongly support). If you decide against these preliminary tests, base your choice of a test procedure on the sample sizes. If the two sample sizes are approximately equal, use the equal-variance t-test. If the ratio of the two sample sizes (larger sample size over the smaller sample size) is equal to or greater than 1.5, use the unequal-variance t-test. This is the recommendation of Ott (1984), page 144.

Step 4 – Choose the Appropriate Statistical Test

Introduction

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

Normal Data with Equal Variances

Use the Equal Variance T-Test Section for hypothesis testing and the Equal Variance portion of the Confidence Limits Section for interval estimation.

Normal Data with Unequal Variances

Use the Unequal Variance T-Test Section for hypothesis testing and the Unequal Variance portion of the Confidence Limits Section for interval estimation.

Non-Normal Data with Equal Variances

Use the Mann-Whitney U or Wilcoxon Rank-Sum Test for hypothesis testing.

Non-Normal Data with Unequal Variances

Use the Kolmogorov-Smirnov Test in this case or if your data have a lot of ties.

Step 5 – Interpret Findings

Introduction

You are now ready to conduct your two-sample test. Depending upon the nature of your study, you look at either of the following sections.

Hypothesis Testing

First find the appropriate Alternative Hypothesis row. Usually, you will use the first (Var1-Var2<>0) row. This two-tailed test is the standard. If the probability level is less than your chosen alpha level, you reject the null hypothesis of equal means (or medians) and conclude that the means are different. Your next task is to look at the means themselves to determine if the size of the difference is of practical interest.

Confidence Limits

The confidence limits of the difference let you put bounds on the size of the difference. If these limits are narrow and close to zero, you might determine that even though your results are statistically significant, the magnitude of their difference is not of practical interest.

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.

Example of Two-Sample T-Test Steps

This example illustrates the interpretation of two-sample tests. Of course, no example is infallible, but the intention is to highlight a number of the issues that you must consider in choosing the right two-sample test for your data as you proceed through the Two-Sample Checklist.

Two friends, who are also neighbors, love pizza, and they each usually order their pizzas from different places. Friend A orders from pizza company 1, while friend B orders from pizza company 2. The two friends got in an argument about which pizza company delivers the fastest or whether there was a difference at all in delivery times. Friend A took a random sample of 10 delivery times from pizza place 1 over the next six months. Friend B took a random sample of 8 delivery times over the same time frame. The pizza orders were not necessarily taken on the same day, but the orders were usually placed in the evening hours from 6 to 9 p.m. The data are shown below.

Pizza1	Pizza2
21	15
20	17
25	17
20	19
23	22
20	12
13	16
18	21
25	
24	

Step 1 – Data Preparation

The sample sizes here are not as large as we would like, but they are typical. There are no missing values, and the data are continuous (although the times are rounded to the closest minute). There is no way to assess outliers until Step 3.

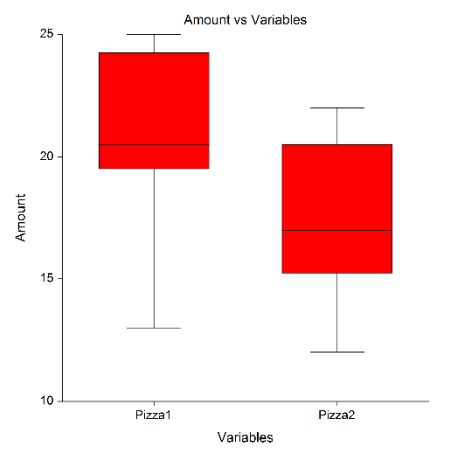
Step 2 – Setup and Run the T-Test Panel

The selection and running of the Two-Sample T-Test from the Analysis menu on the two response variables, Pizza1 and Pizza2, would produce the reports that follow. The alpha value has been set at 0.05.

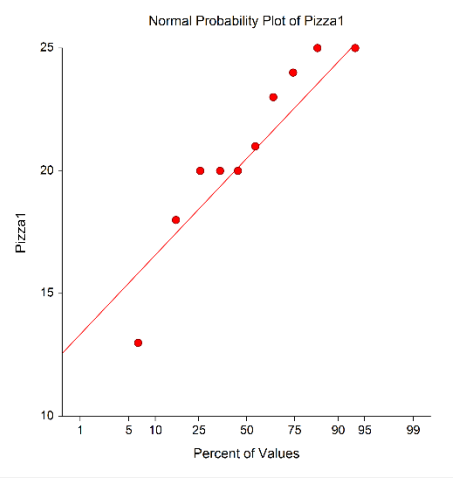
Step 3 – Check Assumptions

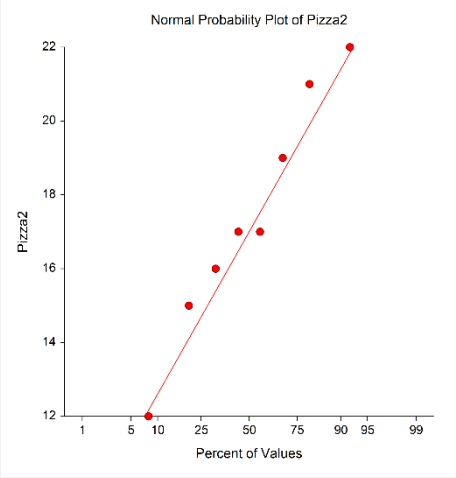
We first check for normality with the graphic perspectives: box plots, normal probability plots, histograms, and density traces.

Box Plots

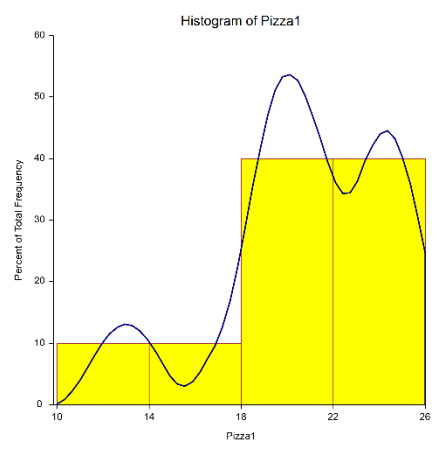


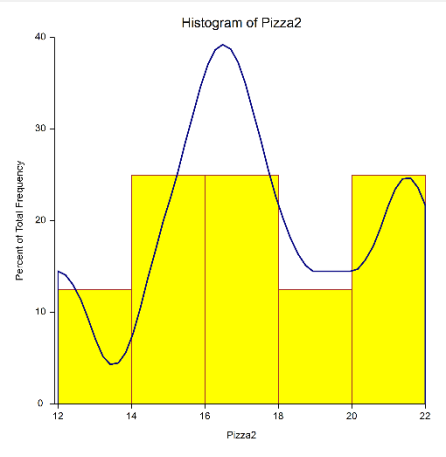
Probability Plots





Histograms with Density Trace





The tails of the box plot for Pizza1 show left skewness, and the median is not in the middle of the box itself (i.e., it is also pulled left). While the tails for Pizza2 are symmetrical, the median is also pulled left toward the short delivery times. Remember that these samples are small, and interpretation of box plots for small samples must be flexible. The interpretation from the box plots is that both groups show some non-normality.

The normal probability plots in give a similar picture. Since all of the data values for Pizza2 lie within the 95% confidence bands, delivery times seem to be normal. On the other hand, the normal probability plot for Pizza1 shows a possible outlier among the short delivery times since the observation of 13 minutes is outside the confidence bands. If it were not for this one observation, the normal probability plot for Pizza1 would be normal.

The histogram does not usually give an accurate graphic perception of normality for small samples, although the super-imposed density trace helps a lot. Examination of the histogram for Pizza1 shows that there is at least one observation that contributes to the left skewness, and the histogram for Pizza1 does not look normal. However, the histogram for Pizza2 reveals a reasonably normal distribution.

At this point of the graphic analysis of the normality assumption, you would likely say the Pizza2 delivery times are normal while Pizza1 delivery times are not. However, since these samples are small, be sure to evaluate the numerical confirmation of normality by the skewness, kurtosis, and omnibus normality tests for each pizza firm using the Tests of Assumptions Section.

Tests of the Normality Assumption for Pizza1

Normality Test	Test Statistic	Prob Level	Reject H0 of Normality at α = 0.050?
Shapiro-Wilk	0.9038	0.24108	No
Skewness	-1.4244	0.15434	No
Kurtosis	1.0525	0.29259	No
Omnibus (Skewness or Kurtosis)	3.1366	0.20840	No

Tests of the Normality Assumption for Pizza2

Normality Test	Test Statistic	Prob Level	Reject H0 of Normality at α = 0.050?
Shapiro-Wilk	0.9720	0.91303	No
Skewness	-0.1541	0.87751	No
Kurtosis			
Omnibus (Skewness or Kurtosis)			

Tests of the Equal Variance Assumption

Equal-Variance Test	Test Statistic	Prob Level	Reject H0 of Equal Variances at α = 0.050?
Variance-Ratio	1.2729	0.76728	No
Modified-Levene	0.0945	0.76249	No

When evaluating normality, focus your attention on the probability (p-value) and the decision for the given alpha of 0.05. In this case, the decision is acceptance of the hypothesis that the data for Pizza1 is normally distributed by all three normality tests. The lowest probability is 0.1543 for the skewness test, and this is greater than 0.05, the set alpha value. This same amount of skewness for a larger sample size would have rejected the normality assumption. However, for our example, it seems reasonable to assume that both Pizza1 and Pizza2 are normally distributed. We would strongly recommend that the one outlying value in Pizza1 be double-checked for validity.

We next check for equal variance. Both variance tests (variance-ratio and modified-Levene) indicate acceptance of the hypothesis of equal variances as a shown by the probability greater than 0.05 and the “cannot reject” under the decision conclusion. This equality of variances is portrayed by the box plots.

If you do not consider the preliminary test on variances appropriate, use the sample size criterion. If the sample sizes are roughly equal (no more than a 1.5 ratio of the larger sample size divided by the smaller sample size), use

the equal-variance t-test. In this case, this sample size ratio is 10/8 or 1.25. Thus, go ahead with the equal variance t-test. If you are in doubt, run both tests and compare answers.

Step 4 – Choose the Appropriate Statistical Test

In this example, the conclusions from the assumption checking have been that both samples are normally distributed and that the variances are equal or that the sample sizes are roughly equal. In light of these findings, the appropriate test is the equal-variance t-test, sometimes called the pooled t-test.

Step 5 – Interpret Findings

In order to understand the following discussion, you should run the two-sample t-test on the above data and look at the statistical reports.

The mean delivery times are 20.9 and 17.4 minutes. Note that the standard deviations are about equal at 3.665 and 3.249 minutes for Pizza1 and Pizza2, respectively.

We are interested in the difference between the means. Under the Confidence Limits Section and the Equal Variance Case, the 95% confidence limits for the difference ranges from 0.016557 to 7.033442 minutes. Since zero is not in this interval, there is a statistically significant difference between the two means.

The formal two-tail hypothesis test for this example is shown under the Equal-Variance T-Test section. The p-value or probability of accepting H_o is 0.049, which is less than the chosen alpha level at 0.05, resulting in the rejection of H_o. That is, there is a difference between the two pizza delivery times. The power of this two-tail t-test at 0.05 level of significance is 0.5166. The higher the power (i.e., closer to 1), the better the statistical test is able to detect that the alternative hypothesis is true. The power is not great here (many would find it bearable), and it could have been greatly improved by slightly larger sample sizes.

If we had been interested in checking for the average Pizza1 delivery times being greater than that of Pizza2, we would have looked at the right-tail test in the equal-variance t-test section. The decision here is definitely a rejection since the p-value or the probability of accepting H_o is significantly less than 0.05 (i.e., 0.0245). The power of this one-tail test is much better at 0.653.

This would usually finish the interpretation of this example. However, if you were having second thoughts about the normality for Pizza1 delivery times, you might check the nonparametric equivalent of the equal-variance t-test--the Mann-Whitney U Test--to see if you obtain a similar conclusion. The approximate p-value for the two-tail test is 0.044. This p-value is close to that which we had under the equal-variance t-test. Note that we still reject the null hypothesis. The right-tail test yields a p-value of 0.022, which is almost identical to the equal-variance t-test p-value for this right tail test.

Whenever the data are normal, use the appropriate t-test because the power is always better. If in doubt, cross check your t-test with the appropriate nonparametric test.

This concludes the analysis of this example.