Statistics

Table Of Contents

Basic Statistics

Overview

Basic Statistic

Use Minitab's basic statistics capabilities for calculating basic statistics and for simple estimation and hypothesis testing with one or two samples. The basic statistics capabilities include procedures for:

- Calculating or storing descriptive statistics
- Hypothesis tests and confidence intervals of the mean or difference in means
- Hypothesis tests and confidence intervals for a proportion or the difference in proportions
- Hypothesis test for equality of variance
- Measuring association
- Testing for normality of a distribution

Calculating and storing descriptive statistics

- Display Descriptive Statistics produces descriptive statistics for each column or subset within a column. You can display the statistics in the Session window and/or display them in a graph.
- Store Descriptive Statistics stores descriptive statistics for each column or subset within a column.
- Graphical Summary produces four graphs and an output table in one graph window.

For a list of descriptive statistics available for display or storage, see Descriptive Statistics Available for Display or Storage. To calculate descriptive statistics individually and store them as constants, see Column Statistics.

Confidence intervals and hypothesis tests of means

The four procedures for hypothesis tests and confidence intervals for population means or the difference between means are based upon the distribution of the sample mean following a normal distribution. According to the Central Limit Theorem, the normal distribution becomes an increasingly better approximation for the distribution of the sample mean drawn from any distribution as the sample size increases.

- 1-Sample Z computes a confidence interval or performs a hypothesis test of the mean when the population standard deviation, σ, is known. This procedure is based upon the normal distribution, so for small samples, this procedure works best if your data were drawn from a normal distribution or one that is close to normal. From the Central Limit Theorem, you may use this procedure if you have a large sample, substituting the sample standard deviation for σ. A common rule of thumb is to consider samples of size 30 or higher to be large samples. Many analysts choose the t-procedure over the Z-procedure whenever σ is unknown.
- 1-Sample t computes a confidence interval or performs a hypothesis test of the mean when σ is unknown. This procedure is based upon the t-distribution, which is derived from a normal distribution with unknown σ. For small samples, this procedure works best if your data were drawn from a distribution that is normal or close to normal. This procedure is more conservative than the Z-procedure and should always be chosen over the Z-procedure with small sample sizes and an unknown σ. Many analysts choose the t-procedure over the Z-procedure anytime σ is unknown. According to the Central Limit Theorem, you can have increasing confidence in the results of this procedure as sample size increases, because the distribution of the sample mean becomes more like a normal distribution.
- 2-Sample t computes a confidence interval and performs a hypothesis test of the difference between two population means when σ's are unknown and samples are drawn independently from each other. This procedure is based upon the t-distribution, and for small samples it works best if data were drawn from distributions that are normal or close to normal. You can have increasing confidence in the results as the sample sizes increase.
- Paired t computes a confidence interval and performs a hypothesis test of the difference between two population means when observations are paired (matched). When data are paired, as with before-and-after measurements, the paired t-procedure results in a smaller variance and greater power of detecting differences than would the above 2-sample t-procedure, which assumes that the samples were independently drawn.

Confidence intervals and hypothesis tests of proportions

- 1 Proportion computes a confidence interval and performs a hypothesis test of a population proportion.
- 2 Proportions computes a confidence interval and performs a hypothesis test of the difference between two population proportions.

Confidence intervals and hypothesis tests of equality of variance

- 2 Variances computes a confidence interval and performs a hypothesis test for the equality, or homogeneity, of variance of two samples.

Measures of association

- Correlation calculates the Pearson product moment coefficient of correlation (also called the correlation coefficient or correlation) for pairs of variables. The correlation coefficient is a measure of the degree of linear relationship between two variables. You can obtain a p-value to test if there is sufficient evidence that the correlation coefficient is not zero.

  By using a combination of Minitab commands, you can also compute Spearman's correlation and a partial correlation coefficient. Spearman's correlation is simply the correlation computed on the ranks of the two samples. A partial correlation coefficient is the correlation coefficient between two variables while adjusting for the effects of other variables.
- Covariance calculates the covariance for pairs of variables. The covariance is a measure of the relationship between two variables but it has not been standardized, as is done with the correlation coefficient, by dividing by the standard deviation of both variables.

Distribution test

Normality Test generates a normal probability plot and performs a hypothesis test to examine whether or not the observations follow a normal distribution. Some statistical procedures, such as a Z- or t-test, assume that the samples were drawn from a normal distribution. Use this procedure to test the normality assumption.

Stat Menu

Choose an item below:

Basic Statistics

Regression

ANOVA (Analysis of Variance)

DOE (Design of Experiments)

Control Charts

Quality Tools

Reliability/Survival

Multivariate

Time Series

Tables

Nonparametrics

EDA (Exploratory Data Analysis)

Power and Sample Size

Basic Statistics

Stat > Basic Statistics

Select one of the following commands:

Display Descriptive Statistics

Store Descriptive Statistics

Graphical Summary

1-Sample Z

1-Sample t

2-Sample t

Paired t

1 Proportion

2 Proportions

2 Variances

Correlation

Covariance

Normality Test

Descriptive Statistics Available for Display or Storage

This table shows the statistics displayed in the tabular output and the graphical summary:

Statistic	Session window	Graphical summary	Store
Number of nonmissing values	x	x	x
Number of missing values	x		x
Total number	x		x
Cumulative number	x		x
Percent	x		x
Cumulative percent	x		x
Mean	x	x	x
Trimmed mean	x		x
Confidence interval for mean		x	
Standard error of mean	x		x
Standard deviation	x	x	x
Confidence interval for standard deviation		x	
Variance	x	x	x
Coefficient of variation	x		x
Sum	x		x
Minimum	x	x	x
Maximum	x	x	x
Range	x		x
Median	x	x	x
Confidence interval for median		x	
First and third quartiles	x	x	x
Interquartile range	x		x
Sums of squares	x		x
Skewness	x	x	x
Kurtosis	x	x	x
MSSD	x		x
Normality test statistic, p-value		x	

Statistics examples

Minitab Help offers examples in the following categories:

Basic Statistics
Regression
Analysis of Variance
Multivariate Analysis
Nonparametrics
Tables
Time Series
Exploratory Data Analysis
Power and Sample Size

Examples of Basic Statistics

The following examples illustrate how you can calculate statistics with Minitab. Choose an example below:

Display Descriptive Statistics

Graphical Summary

1-Sample Z: Test and Confidence Interval

1-Sample t: Test and Confidence Interval

2-Sample t: Test and Confidence Interval with the Data in One Column

Paired t: Test and Confidence Interval

1 Proportion: Test and Confidence Interval

2 Proportions: Test and Confidence Interval

2 Variances

Correlation

Partial Correlation

Normality Test

References – Basic Statistics

[1] S.F. Arnold (1990). *Mathematical Statistics.* Prentice-Hall.

[2] M.B. Brown and A.B. Forsythe (1974). "Robust Tests for the Equality of Variances," *Journal of the American Statistical Association*, 69, 364-367.

[3] R.B. D'Agostino and M.A. Stephens, Eds. (1986). *Goodness-of-Fit Techniques*, Marcel Dekker.

[4] J.J. Filliben (1975). "The Probability Plot Correlation Coefficient Test for Normality," *Technometrics*, 17, 111.

[5] T.P. Hettmansperger and S.J. Sheather (1986). "Confidence Intervals Based on Interpolated Order Statistics," Statistics and Probability Letters, 4, 75-79.

[6] N.L. Johnson and S. Kotz (1969). *Discrete Distributions*, John Wiley & Sons.

[7] H. Levene (1960). *Contributions to Probability and Statistics*, Stanford University Press.

[8] H.W. Lilliefore (1967). "On the Kolmogorov–Smirnov Test for Normality with Mean and Variance Unknown," *Journal of the American Statistical Association*, 62, 399-402.

[9] T.A. Ryan, Jr. and B.L. Joiner (1976). "Normal Probability Plots and Tests for Normality," Technical Report, Statistics Department, The Pennsylvania State University. (Available from Minitab Inc.)

[10] S.S. Shapiro and R.S. Francia (1972). "An Approximate Analysis of Variance Test for Normality," *Journal of the American Statistical Association*, 67, 215-216.

[11] S.S. Shapiro and M.B. Wilk. (1965). "An Analysis of Variance Test for Normality (Complete Samples)," Biometrika, 52, 591.

Display Descriptive Statistics

Display Descriptive Statistics

Stat > Basic Statistics > Display Descriptive Statistics

Produces descriptive statistics for each column, or for each level of a By variable.

To calculate descriptive statistics individually and store them as constants, see Column Statistics. To store many different statistics, use Store Descriptive Statistics.

Use Display Descriptive Statistics to produce statistics for each column or for subsets within a column. You can display these statistics in the Session window and optionally in a graph (see Descriptive Statistics Available for Display or Storage).

Dialog box items

Variables: Choose the columns you want to describe.

> **By variables (optional):** Enter the column containing the by variables to display descriptive statistics separately for each value of the specified variable. Column lengths must be equal. See Graph Limits for limitations associated with using a By variable.

<Statistics>

<Graphs>

Data – Display Descriptive Statistics

The data columns must be numeric. The optional grouping column (also called a By column) can be numeric, text, or date/time and must be the same length as the data columns. If you wish to change the order in which text categories are processed from their default alphabetical order, you can define your own order (see Ordering Text Categories).

Minitab automatically omits missing data from the calculations.

To display descriptive statistics

1 Choose **Stat > Basic Statistics > Display Descriptive Statistics**.

2 In **Variables** enter the columns containing the data you want to describe.

3 If you like, use any dialog box options, then click **OK**.

Graph Limits

You can display your data in a histogram, a histogram with normal curve, a dotplot, or a boxplot, or display a graphical summary. The displayed statistics are listed in Descriptive Statistics Available for Display or Storage.

The graphical summary includes a table of descriptive statistics, a histogram with normal curve, a boxplot, a confidence interval for the population mean, and a confidence interval for the population median. Minitab can display a maximum of 100 graphs at a time. Therefore, the graphical summary will not work when there are more than 100 columns, 100 distinct levels or groups in a By column, or the combination of columns and By levels is more than 100.

There is no restriction on the number of columns or levels when producing output in the Session window.

Tip If you exceed the maximum number of graphs because of the number of levels of your By variable, you can decrease the number of graphs by unstacking your data and displaying descriptive statistics for data subsets. See Unstack Columns for more information.

Display Descriptive Statistics – Statistics

Stat > Basic Statistics > Display Descriptive Statistics > Statistics

Allows you to choose the statistics that you wish to display.

Dialog box items

Mean: Choose to display the arithmetic mean.

SE of mean: Choose to display the standard error of the mean.

Standard deviation: Choose to display the standard deviation of the data.

Variance: Choose to display the variance of the data.

Coefficient of variation: Choose to display the coefficient of variation.

First quartile: Choose to display the first quartile.

Median: Choose to display the median.

Third quartile: Choose to display the third quartile.

Interquartile range: Choose to display the difference between the first and third quartiles.

Trimmed mean: Choose to display the trimmed mean.

Sum: Choose to display the data sum.

Minimum: Choose to display the data minimum.

Maximum: Choose to display the data maximum.

Range: Choose to display the data range.

N nonmissing: Choose to display the number of nonmissing column entries.

N missing: Choose to display the number of missing column entries.

N total: Choose to display the total (nonmissing and missing) number of column entries.

Cumulative N: Choose to display the cumulative number of entries.

Percent: Choose to display the percent of observations that a group constitutes. The percent will be 100 unless you use a By variable.

Cumulative percent: Choose to display the cumulative percent.

Sum of squares: Choose to display the sum of the squared data values. This is the uncorrected sums of squares, without first subtracting the mean.

Skewness: Choose to display the skewness value.

Kurtosis: Choose to display the kurtosis value

MSSD: Choose to display half the Mean of Successive Squared Differences.

Display Descriptive Statistics – Graphs

Stat > Basic Statistics > Display Descriptive Statistics > Graphs

Displays a histogram, a histogram with a normal curve, an individual value plot, and a boxplot.

Dialog box items

Histogram of data: Choose to display a histogram for each variable.

Histogram of data, with normal curve: Choose to display a histogram with a normal curve for each variable.

Individual value plot: Choose to display an individual value plot for each variable.

Boxplot of data: Choose to display a boxplot for each variable.

Example of Displaying Descriptive Statistics

You want to compare the height (in inches) of male (Sex=1) and female (Sex=2) students who participated in the pulse study. You choose to display a boxplot of the data.

1 Open the worksheet PULSE.MTW.

2 Choose **Stat > Basic Statistics > Display Descriptive Statistics**.

3 In **Variables**, enter *Height*.

4 In **By variable**, enter *Sex*.

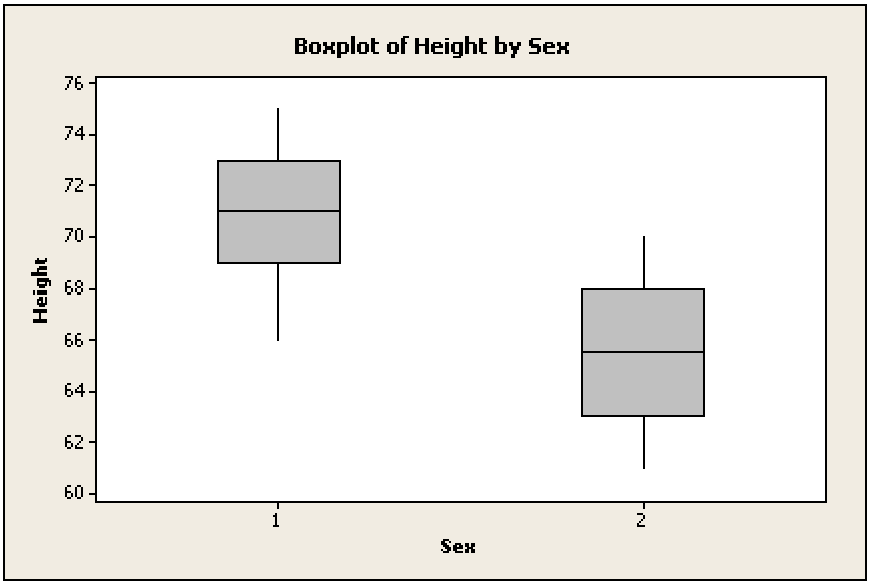
5 Click **Graphs** and check **Boxplot of data**. Click **OK** in each dialog box.

Session window output

Descriptive Statistics: Height

```
Variable  Sex   N  N*    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3
Height    1    57   0  70.754    0.342  2.583   66.000  69.000  71.000  73.000
          2    35   0  65.400    0.433  2.563   61.000  63.000  65.500  68.000

Variable  Sex  Maximum
Height    1     75.000
          2     70.000
```

Graph window output



Interpreting the results

The means shown in the Session window and the boxplots indicate that males are approximately 5.3 inches taller than females, and the spread of the data is about the same.

Store Descriptive Statistics

Store Descriptive Statistics

Stat > Basic Statistics > Store Descriptive Statistics

Stores descriptive statistics for each column, or for each level of one or more By variables.

To calculate descriptive statistics individually and store them as constants, see Column Statistics. To display many different statistics in the Session window, see Display Descriptive Statistics. For a list of statistics available for display or storage, see Descriptive Statistics Available for Display or Storage.

Dialog box items

Variables: Enter the column(s) containing the data you want to describe.

By variables (optional): Enter the column(s) containing the By variable.

<Statistics>

<Options>

Data – Store Descriptive Statistics

The data columns must be numeric. The optional grouping column (also called a By column) can be numeric, text, or date/time and must be the same length as the data columns. If you wish to change the order in which text categories are processed from their default alphabetical order, you can define your own order (see Ordering Text Categories).

Minitab automatically omits missing data from the calculations.

To store descriptive statistics

1 Choose **Stat > Basic Statistics > Store Descriptive Statistics**.

2 In **Variables**, enter the columns containing the data you want to store.

3 If you like, any dialog box options, then click **OK**.

Naming stored columns

Minitab automatically names the storage columns with the name of the stored statistic and a sequential integer starting at 1. For example, suppose you enter two columns in **Variables** and choose to store the default mean and sample size. Minitab will name the storage columns Mean1 and N1 for the first variable and Mean2 and N2 for the second variable. If you use two By variables, Minitab will store the distinct levels (subscripts) of the By variables in columns named ByVar1 and ByVar2, with the appended integer cycling as with the stored statistics.

If you erase the storage columns or rename them, the integers will start over at 1. If you store statistics for many columns, you may want to rename the corresponding stored columns so that you can keep track of their origin.

Store Descriptive Statistics – Statistics

Stat > Basic Statistics > Store Descriptive Statistics > Statistics

Allows you to choose the statistics that you wish to store.

Dialog box items

Mean: Choose to store the arithmetic mean.

SE of mean: Choose to store the standard error of the mean.

Standard deviation: Choose to store the standard deviation of the data.

Variance: Choose to store the variance of the data.

Coefficient of variation: Choose to store the coefficient of variation.

First quartile: Choose to store the first quartile.

Median: Choose to store the median.

Third quartile: Choose to store the third quartile.

Interquartile range: Choose to store the difference between the first and third quartiles.

Trimmed mean: Choose to store the trimmed mean.

Sum: Choose to store the data sum.

Minimum: Choose to store the data minimum.

Maximum: Choose to store the data maximum.

Range: Choose to store the data range.

N nonmissing: Choose to store the number of nonmissing column entries.

N missing: Choose to store the number of missing column entries.

N total: Choose to store the total (nonmissing and missing) number of column entries.

Cumulative N: Choose to store the cumulative number of entries.

Percent: Choose to store the percent of observations that a group constitutes. The percent will be 100 unless you use a By variable.

Cumulative percent: Choose to store the cumulative percent.

Sum of squares: Choose to store the sum of the squared data values. This is the uncorrected sums of squares, without first subtracting the mean.

Skewness: Choose to store the skewness value.

Kurtosis: Choose to store the kurtosis value

MSSD: Choose to store half the Mean of Successive Squared Differences.

Store Descriptive Statistics – Options

Stat > Basic Statistics > Store Descriptive Statistics > Options

Select among options for calculating and storing when you use a By variable.

Dialog box items

Store a row of output for each row of input: Check this option if you want Minitab to append the appropriate statistics to each row of input data. (By default, Minitab stores the requested statistics at the top of the worksheet only.)

Include empty cells: If you choose more than one By variable, Minitab stores summary statistics for all combinations of the By variable levels, including combinations for which there are no data (called empty cells). If you do not want to store empty cells, uncheck this option.

Include missing as a By level: Check this option to include missing values as a distinct level of the By variable. (By default, Minitab ignores data from rows with missing values in a By column.)

Store distinct values of By variables: Uncheck this option if you do not want Minitab to include columns in the summary data that indicate the levels of the By variables.

Graphical Summary

Graphical Summary

Stat > Basic Statistics > Graphical Summary

Use to produce a graphical summary for each column, or for each level of a By variable.

The graphical summary includes four graphs: histogram of data with an overlaid normal curve, boxplot, 95% confidence intervals for μ, and 95% confidence intervals for the median.

The graphical summary also displays a table of:

- Anderson-Darling Normality Test statistics
- Descriptive statistics
- Confidence intervals for μ, σ, and the median

Dialog box items

Variables: Enter the columns for which you want to create a graphical summary.

> **By variables (optional):** Enter the columns containing the by variables to create separate graphical summaries for each level of a grouping variable. Column lengths must be equal. See Graph Limits for limitations associated with using a By variable.

Confidence level: Enter a value for the confidence level for the confidence intervals. The default level is 95%. Specify any number between 0 and 100. For example, entering 90 generates 90% confidence intervals for the mean, median, and standard deviation.

Data – Graphical Summary

The data columns must be numeric. The optional grouping column (also called a By column) can be numeric, text, or date/time and must be the same length as the data columns. If you wish to change the order in which text categories are processed from their default alphabetical order, you can define your own order (see Ordering Text Categories).

Minitab automatically omits missing data from the calculations.

To make a graphical summary

1 Choose **Stat > Basic Statistics > Graphical Summary**.

2 In **Variables**, enter the columns you want to describe.

3 If you like, use any dialog box options, then click **OK**.

Example of a graphical summary

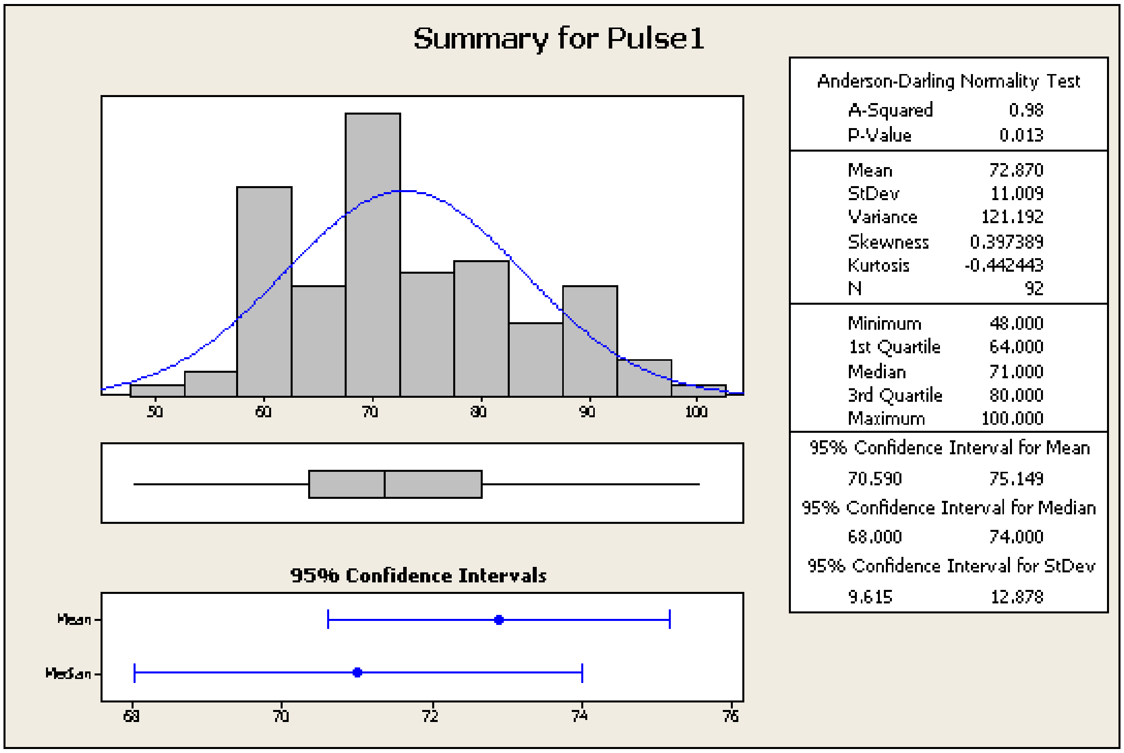
Students in an introductory statistics course participated in a simple experiment. Each student recorded his or her resting pulse. Then they all flipped coins, and those whose coins came up heads ran in place for one minute. Then the entire class recorded their pulses. You want to examine the students' resting pulse rates.

1 Open the worksheet PULSE.MTW.

2 Choose **Stat > Basic Statistics > Graphical Summary**.

3 In **Variables**, enter *Pulse1*. Click **OK**.

Graph window output



```
Anderson-Darling Normality Test
   A-Squared          0.98
   P-Value           0.013

   Mean             72.870
   StDev            11.009
   Variance        121.192
   Skewness       0.397389
   Kurtosis      -0.442443
   N                    92

   Minimum          48.000
   1st Quartile     64.000
   Median           71.000
   3rd Quartile     80.000
   Maximum         100.000
 95% Confidence Interval for Mean
   70.590           75.149
95% Confidence Interval for Median
   68.000           74.000
 95% Confidence Interval for StDev
   9.615            12.878
```

Interpreting the results

The mean of the students' resting pulse is 72.870 (95% confidence intervals of 70.590 and 75.149). The standard deviation is 11.009 (95% confidence intervals of 9.615 and 12.878).

Using a significance level of 0.05, the Anderson-Darling Normality Test (A-Squared = 0.98, P-Value = 0.013) indicates that the resting pulse data do not follow a normal distribution.

1 Sample Z

1-Sample Z

Stat > Basic Statistics > 1-Sample Z

Use 1-Sample Z to compute a confidence interval or perform a hypothesis test of the mean when σ is known. For a two-tailed one-sample Z

$H_0 : \mu = \mu_0$ versus $H_1: \mu \neq \mu_0$

where μ is the population mean and μ_0 is the hypothesized population mean.

Dialog box items

Samples in columns: Choose if you have entered raw data in columns. Enter the columns containing the sample data.

Summarized data: Choose if you have summary values for the sample size and mean.

Sample size: Enter the value for the sample size

Mean: Enter the value for the sample mean.

Standard deviation: Enter the value for the population standard deviation.

Test mean: Enter the test mean μ_0.

<Graphs>

<Options>

Data – 1-Sample Z

Enter each sample in a single numeric column. You can generate a hypothesis test or confidence interval for more than one column at a time.

Minitab automatically omits missing data from the calculations.

To do a Z-test and confidence interval of the mean

1. Choose **Stat > Basic Statistics > 1-Sample Z**.
2. In **Variables**, enter the columns containing the samples.
3. In **Sigma**, enter a value for σ.
4. If you like, use any dialog box options, then click **OK**.

1-Sample Z – Graphs

Stat > Basic Statistics > 1-Sample Z > Graphs

Displays a histogram, an individual data plot, and a boxplot of the variables. The graphs show the sample mean and a confidence interval (or bound) for the mean. When you do a hypothesis test, the graphs also show the null hypothesis test value.

Dialog box items

Histogram of data: Choose to display a histogram for each variable.

Individual data plot: Choose to display an individual value plot for each variable.

Boxplot of data: Choose to display a boxplot for each variable

1-Sample Z – Options

Stat > Basic Statistics > 1-Sample Z > Options

Specify the confidence level for the confidence interval, or define the alternative hypothesis.

Dialog box items

Confidence level: Enter the level of confidence desired. Enter any number between 0 and 100. Entering 90 will result in a 90% confidence interval. The default is 95%.

Alternative: Choose less than (lower-tailed), not equal (two-tailed), or greater than (upper-tailed). If you choose a lower-tailed or an upper-tailed hypothesis test, an upper or lower confidence bound will be constructed, respectively, rather than a confidence interval.

Example of a 1-Sample Z-Test and Z-Confidence Interval

Measurements were made on nine widgets. You know that the distribution of measurements has historically been close to normal with σ = 0.2. Because you know σ, and you wish to test if the population mean is 5 and obtain a 90% confidence interval for the mean, you use the Z-procedure.

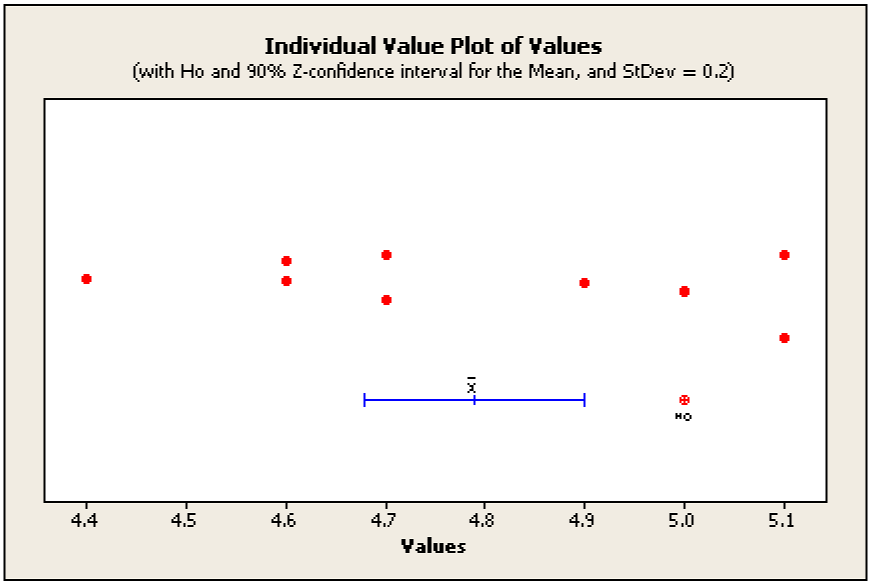
1. Open the worksheet EXH_STAT.MTW.
2. Choose **Stat > Basic Statistics > 1-Sample Z**.
3. In **Samples in Columns**, enter *Values*.
4. In **Standard deviation**, enter *0.2*.
5. In **Test mean**, enter *5*.
6. Click **Options**. In **Confidence level**, enter *90*. Click **OK**.
7. Click **Graphs**. Check **Individual value plot**. Click **OK** in each dialog box.

Session window output

One-Sample Z: Values

```
Test of mu = 5 vs not = 5
The assumed standard deviation = 0.2


Variable  N     Mean    StDev  SE Mean          90% CI              Z      P
Values    9  4.78889  0.24721  0.06667  (4.67923, 4.89855)  -3.17  0.002
```

Graph window output



Interpreting the results

The test statistic, Z, for testing if the population mean equals 5 is –3.17. The p-value, or the probability of rejecting the null hypothesis when it is true, is 0.002. This is called the attained significance level, p-value, or attained α of the test. Because the p-value of 0.002 is smaller than commonly chosen α-levels, there is significant evidence that μ is not equal to 5, so you can reject H0 in favor of μ not being 5.

A hypothesis test at α = 0.1 could also be performed by viewing the individual value plot. The hypothesized value falls outside the 90% confidence interval for the population mean (4.67923, 4.89855), and so you can reject the null hypothesis.

1 Sample t

1-Sample t

Stat > Basic Statistics > 1-Sample t

Performs a one sample t-test or t-confidence interval for the mean.

Use 1-Sample t to compute a confidence interval and perform a hypothesis test of the mean when the population standard deviation, σ, is unknown. For a two-tailed one-sample t,

$H_0: \mu = \mu_0$ versus $H_1: \mu \neq \mu_0$

where μ is the population mean and μ_0 is the hypothesized population mean.

Dialog box items

Samples in columns: Choose if you have entered raw data in columns. Enter the columns containing the sample data.

Summarized data: Choose if you have summary values for the sample size, mean, and standard deviation.

Sample size: Enter the value for the sample size

Mean: Enter the value for the sample mean.

Standard deviation: Enter the value for the sample standard deviation.

Test mean: Enter the test mean μ_0 .

<Options>

<Graphs>

Data – 1-Sample t

Enter each sample in a single numeric column. You can generate a hypothesis test or confidence interval for more than one column at a time.

Minitab automatically omits missing data from the calculations.

To compute a t-test and confidence interval of the mean

1 Choose **Stat > Basic Statistics > 1-Sample t**.

2 In **Variables**, enter the columns containing the samples.

3 If you like, use any dialog box options, then click **OK**.

1-Sample t – Graphs

Stat > Basic Statistics > 1-Sample t > Graphs

Displays a histogram, an individual value plot, and a boxplot of the variables. The graphs show the sample mean and a confidence interval (or bound) for the mean. In addition, the null hypothesis test value is displayed when you do a hypothesis test.

Dialog box items

Histogram of data: Choose to display a histogram for each variable.

Individual value plot : Choose to display an individual value plot for each variable.

Boxplot of data: Choose to display a boxplot for each variable

1-Sample t – Options

Stat > Basic Statistics > 1-Sample t > Options

Specify the confidence level for the confidence interval, or define the alternative hypothesis.

Dialog box items

Confidence level: Enter the level of confidence desired. Enter any number between 0 and 100. Entering 90 will result in a 90% confidence interval. The default is 95%.

Alternative: Enter less than (lower-tailed), not equal (two-tailed), or greater than (upper-tailed). If you choose a lower-tailed or an upper-tailed hypothesis test, an upper or lower confidence bound will be constructed, respectively, rather than a confidence interval.

Example of a 1-Sample t-Test and t-Confidence Interval

Measurements were made on nine widgets. You know that the distribution of widget measurements has historically been close to normal, but suppose that you do not know σ. To test if the population mean is 5 and to obtain a 90% confidence interval for the mean, you use a t-procedure.

1 Open the worksheet EXH_STAT.MTW.

2 Choose **Stat > Basic Statistics > 1-Sample t**.

3 In **Samples in columns**, enter *Values*.

4 In **Test mean**, enter *5*.

5 Click **Options**. In **Confidence level** enter *90*. Click **OK** in each dialog box.

Session window output

One-Sample T: Values

```
Test of mu = 5 vs not = 5


Variable  N     Mean    StDev  SE Mean         90% CI             T      P
Values    9  4.78889  0.24721  0.08240  (4.63566, 4.94212)  -2.56  0.034
```

Interpreting the results

The test statistic, T, for H0: $\mu = 5$ is calculated as –2.56.

The p-value of this test, or the probability of obtaining more extreme value of the test statistic by chance if the null hypothesis was true, is 0.034. This is called the attained significance level, or p-value. Therefore, reject H0 if your acceptable α level is greater than the p-value, or 0.034.

A 90% confidence interval for the population mean, μ, is (4.63566,4.94212). This interval is slightly wider than the corresponding Z-interval shown in Example of 1-Sample Z.

2 Sample t

2-Sample t

Stat > Basic Statistics > 2-Sample t

Performs an independent two-sample t-test and generates a confidence interval.

When you have dependent samples, use Stat > Basic Statistics > Paired t.

Use 2-Sample t to perform a hypothesis test and compute a confidence interval of the difference between two population means when the population standard deviations, σ's, are unknown. For a two-tailed two-sample t

H_0: $\mu_1 - \mu_2 = \delta_0$ versus H_1: $\mu_1 - \mu_2 \neq \delta_0$

where μ_1 and μ_2 are the population means and δ_0 is the hypothesized difference between the two population means.

Dialog box items

Samples in one column: Choose if the sample data are in a single column, differentiated by subscript values (group codes) in a second column.

Samples: Enter the column containing the data.

Subscripts: Enter the column containing the sample subscripts.

Samples in different columns: Choose if the data of the two samples are in separate columns.

First: Enter the column containing one sample.

Second: Enter the column containing the other sample.

Summarized data: Choose if you have summary values for the sample size, mean, and standard deviation for each sample.

First

Sample size: Enter the value for the sample size.

Mean: Enter the value for the mean.

Standard deviation: Enter the value for the standard deviation.

Second

Sample size: Enter the value for the sample size.

Mean: Enter the value for the mean.

Standard deviation: Enter the value for the standard deviation.

Assume equal variances: Check to assume that the populations have equal variances. The default is to assume unequal variances. See Equal or unequal variances.

<Graphs>

<Options>

Data – 2-Sample t

Data can be entered in one of two ways:

- Both samples in a single numeric column with another grouping column (called subscripts) to identify the population. The grouping column may be numeric, text, or date/time.
- Each sample in a separate numeric column.

The sample sizes do not need to be equal. Minitab automatically omits missing data from the calculations.

To do a 2-sample t-confidence interval and test

1 Choose **Stat > Basic Statistics > 2-Sample t**.

2 Choose one of the following:

- If your data are stacked in a single column:
  - Choose **Samples in one column**.
  - In **Samples**, enter the column containing the numeric data.
  - In **Subscripts**, enter the column containing the group or population codes.
- If your data are unstacked, that is each sample is in a separate column:
  - Choose **Samples in different columns**.
  - In **First**, enter the column containing the first sample.
  - In **Second**, enter the column containing the other sample.

3 If you like, use any dialog box options, and click **OK**.

Equal or unequal variances

If you check **Assume equal variances**, the sample standard deviations are pooled to obtain a single estimate of σ.

The two-sample t-test with a pooled variances is slightly more powerful than the two-sample t-test with unequal variances, but serious error can result if the variances are not equal. Therefore, the pooled variance estimate should not be used in many cases. Use Test for Equal Variances to test the equal variance assumption.

2-Sample t – Graphs

Stat > Basic Statistics > 2-Sample t > Graphs

Displays an individual value plot and a boxplot of the variables. The graphs also display the sample means.

Dialog box items

Individual value plot: Choose to display an individual value plot for each sample.

Boxplots of data: Choose to display a boxplot for each sample.

2-Sample t – Options

Stat > Basic Statistics > 2-Sample t > Options

Dialog box items

Confidence level: Enter the level of confidence desired. Enter any number between 0 and 100. The default is 95.

Test difference: Enter the null hypothesis value, which is the hypothesized difference in population means δ_0. The default is zero, or that the two population means are equal.

Alternative: Choose less than (lower-tailed), not equal (two-tailed), or greater than (upper-tailed) depending on the kind of test that you want.

Example of a 2-Sample t with the Samples in one Column

A study was performed in order to evaluate the effectiveness of two devices for improving the efficiency of gas home-heating systems. Energy consumption in houses was measured after one of the two devices was installed. The two devices were an electric vent damper (Damper=1) and a thermally activated vent damper (Damper=2). The energy consumption data (BTU.In) are stacked in one column with a grouping column (Damper) containing identifiers or subscripts to denote the population. Suppose that you performed a variance test and found no evidence for variances being unequal (see Example of 2 Variances). Now you want to compare the effectiveness of these two devices by determining whether or not there is any evidence that the difference between the devices is different from zero.

1 Open the worksheet FURNACE.MTW.

2 Choose **Stat > Basic Statistics > 2-Sample T**.

3 Choose **Samples in one column**.

4 In **Samples**, enter *'BTU.In'*.

5 In **Subscripts**, enter *Damper*.

6 Check **Assume equal variances**. Click **OK**.

Session window output

Two-Sample T-Test and CI: BTU.In, Damper

```
Two-sample T for BTU.In

Damper   N   Mean  StDev  SE Mean
1       40   9.91   3.02     0.48
2       50  10.14   2.77     0.39


Difference = mu (1) - mu (2)
Estimate for difference:  -0.235250
95% CI for difference:  (-1.450131, 0.979631)
T-Test of difference = 0 (vs not =): T-Value = -0.38  P-Value = 0.701  DF = 88
Both use Pooled StDev = 2.8818
```

Interpreting the results

Minitab displays a table of the sample sizes, sample means, standard deviations, and standard errors for the two samples.

Since we previously found no evidence for variances being unequal, we chose to use the pooled standard deviation by choosing **Assume equal variances**. The pooled standard deviation, 2.8818, is used to calculate the test statistic and the confidence intervals.

A second table gives a confidence interval for the difference in population means. For this example, a 95% confidence interval is (–1.45, 0.98) which includes zero, thus suggesting that there is no difference. Next is the hypothesis test result. The test statistic is –0.38, with p-value of 0.701, and 88 degrees of freedom.

Since the p-value is greater than commonly chosen α-levels, there is no evidence for a difference in energy use when using an electric vent damper versus a thermally activated vent damper.

Paired t

Paired t

Stat > Basic Statistics > Paired t

Performs a paired t-test. This is appropriate for testing the mean difference between paired observations when the paired differences follow a normal distribution.

Use the Paired t command to compute a confidence interval and perform a hypothesis test of the mean difference between paired observations in the population. A paired t-test matches responses that are dependent or related in a pairwise manner. This matching allows you to account for variability between the pairs usually resulting in a smaller error term, thus increasing the sensitivity of the hypothesis test or confidence interval.

Typical examples of paired data include measurements on twins or before-and-after measurements. For a paired t-test:

$H_0: \mu_d = \mu_0$ versus $H_1: \mu_d \neq \mu_0$

where μ_d is the population mean of the differences and μ_0 is the hypothesized mean of the differences.

When the samples are drawn independently from two populations, use Stat > Basic Statistics > 2-sample t.

Dialog box items

Sample in columns: Choose if you have entered raw data in two columns.

First sample: Enter the column containing the first sample

Second sample: Enter the column containing the second sample

Summarized data: Choose if you have summary values for the sample size, mean, and standard deviation of the difference.

Sample size: Enter the value for the sample size.

Mean: Enter the value for the mean.

Standard deviation: Enter the value for the standard deviation.

Paired t evaluates the first sample minus the second sample.

<Graphs>

<Options>

Data – Paired t

The data from each sample must be in separate numeric columns of equal length. Each row contains the paired measurements for an observation. If either measurement in a row is missing, Minitab automatically omits that row from the calculations.

To compute a paired t-test and confidence interval

1 Choose **Stat > Basic Statistics > Paired t**.

2 In **First Sample**, enter the column containing the first sample.

3 In **Second Sample**, enter the column containing the second sample.

4 If you like, use any dialog box options, and click **OK**.

Paired t – Graphs

Stat > Basic Statistics > Paired t > Graphs

Displays a histogram, an individual value plot, and a boxplot of the paired differences. The graphs show the sample mean of the differences and a confidence interval (or bound) for the mean of the differences. In addition, the null hypothesis test value is displayed when you do a hypothesis test.

Dialog box items

Histogram of differences: Choose to display a histogram of paired differences

Individual value plot: Choose to display an individual value plot.

Boxplot of differences: Choose to display a boxplot of paired differences

Paired t – Options

Stat > Basic Statistics > Paired t > Options

Select among options for the confidence interval and hypothesis test.

Dialog box items

Confidence level: Enter the level of confidence desired. Enter any number between 0 and 100. The default is 95%.

Test mean: Enter the null hypothesis value, which is the hypothesized population mean of the paired differences. The default is 0.

Alternative: Choose the form of the alternative hypothesis: less than (lower-tailed), not equal (two-tailed), or greater than (upper-tailed). The default is a two-tailed test. If you choose a lower-tailed or an upper-tailed hypothesis test, an upper or lower confidence bound will be constructed, respectively, rather than a confidence interval.

Example of Paired t

A shoe company wants to compare two materials, A and B, for use on the soles of boys' shoes. In this example, each of ten boys in a study wore a special pair of shoes with the sole of one shoe made from Material A and the sole on the other shoe made from Material B. The sole types were randomly assigned to account for systematic differences in wear between the left and right foot. After three months, the shoes are measured for wear.

For these data, you would use a paired design rather than an unpaired design. A paired t-procedure would probably have a smaller error term than the corresponding unpaired procedure because it removes variability that is due to differences between the pairs. For example, one boy may live in the city and walk on pavement most of the day, while another boy may live in the country and spend much of his day on unpaved surfaces.

1 Open the worksheet EXH_STAT.MTW.

2 Choose **Stat > Basic Statistics > Paired t**.

3 Choose **Samples in columns**.

4 In **First sample**, enter *Mat-A*. In **Second sample**, enter *Mat-B*. Click **OK**.

Session window output

Paired T-Test and CI: Mat-A, Mat-B

```
Paired T for Mat-A - Mat-B

             N       Mean     StDev   SE Mean
Mat-A       10    10.6300    2.4513    0.7752
Mat-B       10    11.0400    2.5185    0.7964
Difference  10  -0.410000  0.387155  0.122429


95% CI for mean difference: (-0.686954, -0.133046)
T-Test of mean difference = 0 (vs not = 0): T-Value = -3.35  P-Value = 0.009
```

Interpreting the results

The confidence interval for the mean difference between the two materials does not include zero, which suggests a difference between them. The small p-value (p = 0.009) further suggests that the data are inconsistent with H0: µ d = 0, that is, the two materials do not perform equally. Specifically, Material B (mean = 11.04) performed better than Material A (mean = 10.63) in terms of wear over the three month test period.

Compare the results from the paired procedure with those from an unpaired, two-sample t-test (**Stat > Basic Statistics > 2-Sample t**). The results of the paired procedure led us to believe that the data are not consistent with H0 (t = –3.35; p = 0.009). The results of the unpaired procedure (not shown) are quite different, however. An unpaired t-test results in a t-value of –0.37, and a p-value of 0.72. Based on such results, we would fail to reject the null hypothesis and would conclude that there is no difference in the performance of the two materials.

In the unpaired procedure, the large amount of variance in shoe wear between boys (average wear for one boy was 6.50 and for another 14.25) obscures the somewhat less dramatic difference in wear between the left and right shoes (the largest difference between shoes was 1.10). This is why a paired experimental design and subsequent analysis with a paired t-test, where appropriate, is often much more powerful than an unpaired approach.

1 Proportion

1 Proportion

Stat > Basic Statistics > 1 Proportion

Performs a test of one binomial proportion.

Use 1 Proportion to compute a confidence interval and perform a hypothesis test of the proportion. For example, an automotive parts manufacturer claims that his spark plugs are less than 2% defective. You could take a random sample of spark plugs and determine whether or not the actual proportion defective is consistent with the claim. For a two-tailed test of a proportion:

H_0: $p = p_0$ versus H_1: $p \neq p_0$ where p is the population proportion and p_0 is the hypothesized value.

To compare two proportions, use Stat > Basic Statistics > 2 Proportions.

Dialog box items

Samples in columns: Choose if you have data in columns, then enter the columns containing the sample data. Each cell of these columns must be one of two possible values and correspond to one item or subject. The possible values in the columns must be identical if you enter multiple columns.

Summarized data: Choose if you have summary values for the number of trials and successes.

Number of trials: Enter a single value for the number of trials. Often the number of trials will be your sample size.

Number of events: Enter the number of observed events. You may enter more than one value.

<Options>

Data – 1 Proportion

You can have data in two forms: raw or summarized.

Raw data

Enter each sample in a numeric, text, or date/time column in your worksheet. Columns must be all of the same type. Each column contains both the success and failure data for that sample. Successes and failures are determined by numeric or alphabetical order. Minitab defines the lowest value as the failure; the highest value as the success. For example:

- For the numeric column entries of "20" and "40," observations of 20 are considered failures; observations of 40 are considered successes.
- For the text column entries of "alpha" and "omega," observations of alpha are considered failures; observations of omega are considered successes. If the data entries are "red" and "yellow," observations of red are considered failures; observations of yellow are considered successes.

  You can reverse the definition of success and failure in a text column by applying a value order (see Ordering Text Categories)

With raw data, you can generate a hypothesis test or confidence interval for more than one column at a time. When you enter more than one column, Minitab performs a separate analysis for each column.

Minitab automatically omits missing data from the calculations.

Summarized data

Enter the number of trials and one or more values for the number of successful events directly in the 1 Proportion dialog box. When you enter more than one successful event value, Minitab performs a separate analysis for each one.

To calculate a test and confidence interval for a proportion

1. Choose **Stat > Basic Statistics > 1 Proportion**.
2. Do one of the following:
   - If you have raw data, choose **Samples in columns**, and enter the columns containing the raw data.
   - If you have summarized data:
     1. Choose **Summarized data**.
     2. In **Number of trials**, enter a single numeric integer for the number of trials. Often the number of trials will be your sample size..
     3. In **Number of events**, enter one or more numeric integers as the observed number of events.
3. If you like, use any dialog box options, and click **OK**.

1 Proportion – Options

Stat > Basic Statistics > 1 Proportion > Options

Select among options for the confidence interval and hypothesis test.

Dialog box items

Confidence level: Enter the level of confidence desired. Enter any number between 0 and 100. The default is 95.

Test proportion: Enter the null hypothesis value, which is the hypothesized population proportion p_0.

Alternative: Choose the form of the alternative hypothesis: less than (lower-tailed), not equal (two-tailed), or greater than (upper-tailed). Note that if you choose a lower-tailed or an upper-tailed hypothesis test, an upper or lower confidence bound will be constructed, respectively, rather than a confidence interval.

Use test and interval based on normal distribution: Check to use the normal approximation to the binomial distribution for calculating the hypothesis test and confidence interval. Most textbooks use the normal approximation method because it is easy for students to calculate by hand. By default, Minitab uses the exact method.

Example of 1 Proportion

A county district attorney would like to run for the office of state district attorney. She has decided that she will give up her county office and run for state office if more than 65% of her party constituents support her. You need to test H_0: p = .65 versus H_1: p > .65.

As her campaign manager, you collected data on 950 randomly selected party members and find that 560 party members support the candidate. A test of proportion was performed to determine whether or not the proportion of supporters was greater than the required proportion of 0.65. In addition, a 95% confidence bound was constructed to determine the lower bound for the proportion of supporters.

1. Choose **Stat > Basic Statistics > 1 Proportion**.
2. Choose **Summarized data**.
3. In **Number of trials**, enter *950*. In **Number of events**, enter *560*.

4 Click **Options**. In **Test proportion**, enter *0.65*.

5 From **Alternative**, choose **greater than**. Click **OK** in each dialog box.

Session window output

Test and CI for One Proportion

```
Test of p = 0.65 vs p > 0.65


                                     95%
                                   Lower    Exact
Sample     X    N  Sample p        Bound  P-Value
1        560  950  0.589474     0.562515    1.000
```

Interpreting the results

The p-value of 1.0 suggests that the data are consistent with the null hypothesis (H0: p = 0.65), that is, the proportion of party members that support the candidate is not greater than the required proportion of 0.65. As her campaign manager, you would advise her not to run for the office of state district attorney.

2 Proportions

2 Proportions

Stat > Basic Statistics > 2 Proportions

Performs a test of two binomial proportions.

Use the 2 Proportions command to compute a confidence interval and perform a hypothesis test of the difference between two proportions. For example, suppose you wanted to know whether the proportion of consumers who return a survey could be increased by providing an incentive such as a product sample. You might include the product sample with half of your mailings and see if you have more responses from the group that received the sample than from those who did not. For a two-tailed test of two proportions:

H_0: $p_1 - p_2 = p_0$ versus H_1: $p_1 - p_2 \neq p_0$

where p_1 and p_2 are the proportions of success in populations 1 and 2, respectively, and p_0 is the hypothesized difference between the two proportions.

To test one proportion, use Stat > Basic Statistics > 1 Proportion.

Dialog box items

Samples in one column: Choose if you have entered raw data into a single column with a second column of subscripts identifying the sample.

Samples: Enter the column containing the raw data.

Subscripts: Enter the column containing the sample subscripts.

Samples in different columns: Choose if you have entered raw data into single columns for each sample.

First: Enter the column containing the raw data for the first sample.

Second: Enter the column containing the raw data for the second sample.

Summarized data: Choose if you have summary values for the number of trials and successes.

First

Trials: Enter the number of trials

Events: Enter the number of events.

Second

Trials: Enter the number of trials

Events: Enter the number of events.

<Options>

Data – 2 Proportions

You can have data in two forms: raw or summarized.

Raw data

Raw data can be entered in two ways: stacked and unstacked.

- Enter both samples in a single column (stacked) with a group column to identify the population. Columns may be numeric, text, or date/time. Successes and failures are determined by numeric or alphabetical order. Minitab defines the lowest value as the failure; the highest value as the success. For example:
  - For the numeric column entries of "5" and "10," observations of 5 are considered failures; observations of 10 are considered successes.
  - For the text column entries of "agree" and "disagree," observations of agree are considered failures; observations of disagree are considered successes. If the data entries are "yes" and "no," observations of no are considered failures; observations of yes are considered successes.
- Enter each sample (unstacked) in separate numeric or text columns. Both columns must be the same type – numeric or text. Successes and failures are defined as above for stacked data.

You can reverse the definition of success and failure in a text column by applying a value order (see Ordering Text Categories).

The sample sizes do not need to be equal. Minitab automatically omits missing data from the calculations.

Summarized data

Enter the number of trials and the number of successful events for each sample directly in the 2 Proportions dialog box.

To calculate a test confidence interval for the difference in proportions

1 Choose **Stat > Basic Statistics > 2 Proportions**.

2 Do one of the following:

- If your raw data are stacked in a single column:

  1 Choose **Samples in one column**.

  2 In **Samples**, enter the column containing the raw data.

  3 In **Subscripts**, enter the column containing the group or population codes.

- If your raw data are unstacked, that is, each sample is in a separate column:

  1 Choose **Samples in different columns**.

  2 In **First**, enter the column containing the first sample.

  3 In **Second**, enter the column containing the other sample.

- If you have summarized data:

  1 Choose **Summarized data**.

  2 In **First sample**, enter numeric values under **Trials** and under **Events**.

  3 In **Second sample**, enter numeric values under **Trials** and under **Events**.

3 If you like, use any dialog box options, and click **OK**.

2 Proportions – Options

Stat > Basic Statistics > 2 Proportions > Options

Select among options for the confidence interval and hypothesis test.

Dialog box items

Confidence level: Enter the level of confidence desired. Enter any number between 0 and 100. The default is 95.

Test difference: Enter the null hypothesis value, which is the hypothesized difference in population proportions value p_0.

Alternative: Choose the form of the alternative hypothesis: less than (lower-tailed), not equal (two-tailed), or greater than (upper-tailed). Note that if you choose a lower-tailed or an upper-tailed hypothesis test, an upper or lower confidence bound will be constructed, respectively, rather than a confidence interval.

Use pooled estimate of p for test: Check to use a pooled estimate of p for the hypothesis test.

Example of 2 Proportions

As your corporation's purchasing manager, you need to authorize the purchase of twenty new photocopy machines. After comparing many brands in terms of price, copy quality, warranty, and features, you have narrowed the choice to two:

Brand X and Brand Y. You decide that the determining factor will be the reliability of the brands as defined by the proportion requiring service within one year of purchase.

Because your corporation already uses both of these brands, you were able to obtain information on the service history of 50 randomly selected machines of each brand. Records indicate that six Brand X machines and eight Brand Y machines needed service. Use this information to guide your choice of brand for purchase.

1 Choose **Stat > Basic Statistics > 2 Proportions**.

2 Choose **Summarized data**.

3 In **First sample**, under **Trials**, enter *50*. Under **Events**, enter *44*.

4 In **Second sample**, under **Trials**, enter *50*. Under **Events**, enter *42*. Click **OK**.

Session window output

Test and CI for Two Proportions

```
Sample   X   N  Sample p
1       44  50  0.880000
2       42  50  0.840000


Difference = p (1) - p (2)
Estimate for difference:  0.04
95% CI for difference:  (-0.0957903, 0.175790)
Test for difference = 0 (vs not = 0):  Z = 0.58  P-Value = 0.564
```

Interpreting the results

Since the p-value of 0.564 is larger than commonly chosen α levels, the data are consistent with the null hypothesis (H_0: $p_1 - p_2 = 0$). That is, the proportion of photocopy machines that needed service in the first year did not differ depending on brand. As the purchasing manager, you need to find a different criterion to guide your decision on which brand to purchase.

You can make the same decision using the 95% confidence interval. Because zero falls in the confidence interval of (–0.096 to 0.176) you can conclude that the data are consistent with the null hypothesis. If you think that the confidence interval is too wide and does not provide precise information as to the value of $p_1 - p_2$, you may want to collect more data in order to obtain a better estimate of the difference.

2 Variances

2 Variances

Stat > Basic Statistics > 2 Variances

Use to perform hypothesis tests for equality, or homogeneity, of variance among two populations using an F-test and Levene's test. Many statistical procedures, including the two sample t-test procedures, assume that the two samples are from populations with equal variance. The variance test procedure will test the validity of this assumption.

Dialog box items

Samples in one column: Choose if you have entered data into a single column with a second column of subscripts identifying the samples.

Samples: Enter the column containing the data.

Subscripts: Enter the column containing the sample subscripts.

Samples in different columns: Choose if you have entered the data for each sample into separate columns.

First: Enter the column containing the data for the first sample.

Second: Enter the column containing the data for the second sample.

Summarized data: Choose if you have summary values for the sample size and variance.

First

Sample size: Enter the value for the sample size for the first sample.

Variance: Enter the value for the variance for the second sample.

Second

Sample size: Enter the values for the sample size for the second sample.

Variance: Enter the value for the variance for the second sample.

<Options>

<Storage>

Data – 2 Variances

Data can be entered in one of two ways:

- Both samples in a single numeric column with another grouping column (called subscripts) to identify the population. The grouping column may be numeric, text, or date/time.
- Each sample in a separate numeric column.

The sample sizes do not need to be equal.

When a subscript column includes missing data, Minitab automatically omits the corresponding value from the calculations.

Data limitations include the following:

1. The F-test requires both groups to have multiple observations.
2. Levene's test requires both groups to have multiple observations, but one group must have three or more.

To perform a variance test

1. Choose **Stat > Basic Statistics > 2 Variances**.
2. Choose one of the following:
   - If your data are stacked in a single column:
     1. Choose **Samples in one column**.
     2. In **Samples**, enter the column containing the numeric data.
     3. In **Subscripts**, enter the column containing the group or population codes.
   - If your data are unstacked, that is each sample is in a separate column:
     1. Choose **Samples in different columns**.
     2. In **First**, enter the column containing the first sample.
     3. In **Second**, enter the column containing the other sample.
3. If you like, use any dialog box options, and click **OK**.

F-test versus Levene's test

Minitab calculates and displays a test statistic and p-value for both an F-test and Levene's test where the null hypothesis is of equal variances versus the alternative of unequal variances. Use the F-test when the data come from a normal distribution and Levene's test when the data come from a continuous, but not necessarily normal, distribution.

The computational method for Levene's Test is a modification of Levene's procedure [2, 7]. This method considers the distances of the observations from their sample median rather than their sample mean. Using the sample median rather than the sample mean makes the test more robust for smaller samples.

2 Variances – Options

Stat > Basic Statistics > 2 Variances > Options

Choose options for the test.

Confidence level: Enter a confidence level for the confidence interval. Enter any number between 0 and 100. The default is 95..

Title: Replace the default graph title with your own title.

2 Variances – Storage

Stat > Basic Statistics > 2 Variances > Storage

Store data from the test.

Storage

Standard deviations: Store the standard deviations.

Variances: Store the variances.

Upper confidence limits for sigmas: Store the upper confidence limits for sigmas.

Lower confidence limits for sigmas: Store the lower confidence limits for sigmas.

Example of 2 Variances

A study was performed in order to evaluate the effectiveness of two devices for improving the efficiency of gas home-heating systems. Energy consumption in houses was measured after one of the two devices was installed. The two devices were an electric vent damper (Damper = 1) and a thermally activated vent damper (Damper = 2). The energy consumption data (BTU.In) are stacked in one column with a grouping column (Damper) containing identifiers or subscripts to denote the population. You are interested in comparing the variances of the two populations so that you can construct a two-sample t-test and confidence interval to compare the two dampers.

1 Open the worksheet FURNACE.MTW.

2 Choose **Stat > Basic Statistics > 2 Variances**.

3 Choose **Samples in one column**.

4 In **Samples**, enter 'BTU.In'.

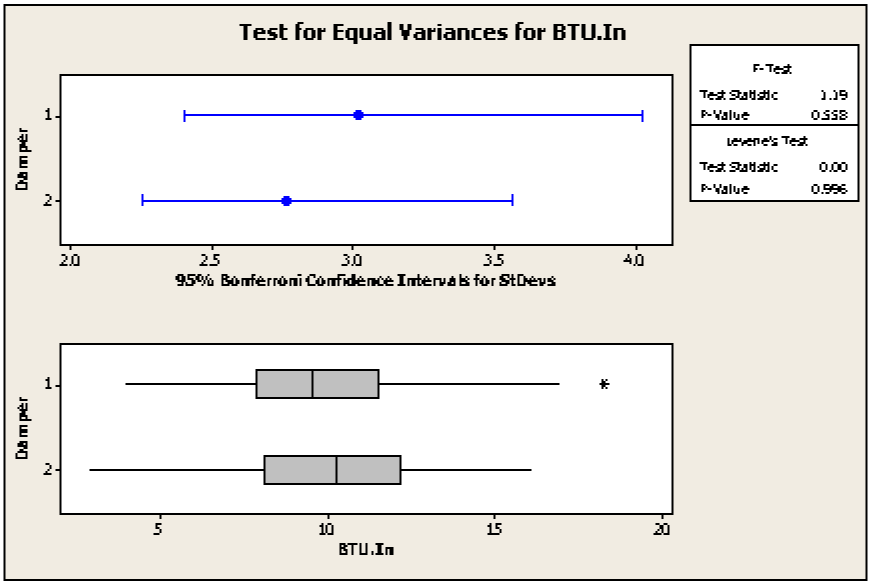
5 In **Subscripts**, enter *Damper*. Click **OK**.

Session window output

Test for Equal Variances: BTU.In versus Damper

```
95% Bonferroni confidence intervals for standard deviations

Damper   N    Lower    StDev    Upper
     1  40  2.40655  3.01987  4.02726
     2  50  2.25447  2.76702  3.56416


F-Test (normal distribution)
Test statistic = 1.19, p-value = 0.558


Levene's Test (any continuous distribution)
Test statistic = 0.00, p-value = 0.996


Test for Equal Variances for BTU.In
```

Graph window output



Interpreting the results

The variance test generates a plot that displays Bonferroni 95% confidence intervals for the population standard deviation at both factor levels. The graph also displays the side-by-side boxplots of the raw data for the two samples. Finally, the results of the F-test and Levene's test are given in both the Session window and the graph. Note that the 95% confidence level applies to the family of intervals and the asymmetry of the intervals is due to the skewness of the chi-square distribution.

For the energy consumption example, the p-values of 0.558 and 0.996 are greater than reasonable choices of α, so you fail to reject the null hypothesis of the variances being equal. That is, these data do not provide enough evidence to claim that the two populations have unequal variances. Thus, it is reasonable to assume equal variances when using a two-sample t-procedure.

Correlation

Correlation

Stat > Basic Statistics > Correlation

Calculates the Pearson product moment correlation coefficient between each pair of variables you list.

You can use the Pearson product moment correlation coefficient to measure the degree of linear relationship between two variables. The correlation coefficient assumes a value between –1 and +1. If one variable tends to increase as the other decreases, the correlation coefficient is negative. Conversely, if the two variables tend to increase together the correlation coefficient is positive. For a two-tailed test of the correlation:

H_0: $\rho = 0$ versus H_1: $\rho \neq 0$ where ρ is the correlation between a pair of variables.

Dialog box items

Variables: Choose the columns containing the variables you want to correlate. When you list two columns, Minitab calculates the correlation coefficient for the pair. When you list more than two columns, Minitab calculates the correlation for every possible pair, and displays the lower triangle of the correlation matrix (in blocks if there is insufficient room to fit across a page).

Display p-values: Check to display p-values for the hypothesis test of the correlation coefficient being zero. This is the default.

Store matrix (display nothing): Check to store the correlation matrix. Minitab does not display the correlation matrix when you choose this option to store the matrix. To display the matrix, choose **Data > Display Data**.

Data – Correlation

Data must be in numeric columns of equal length.

Minitab omits missing data from calculations using a method that is often called pairwise deletion. Minitab omits from the calculations for each column pair only those rows that contain a missing value for that pair.

If you are calculating correlations between multiple columns at the same time, pairwise deletion may result in different observations being included in the various correlations. Although this method is the best for each individual correlation, the correlation matrix as a whole may not be well behaved (for example, it may not be positive definite).

To calculate the Pearson product moment correlation

1 Choose **Stat > Basic Statistics > Correlation**.

2 In **Variables**, enter the columns containing the measurement data.

3 If you like, use any dialog box options, then click **OK**.

Example of Correlation

We have verbal and math SAT scores and first-year college grade-point averages for 200 students and we wish to investigate the relatedness of these variables. We use correlation with the default choice for displaying p-values.

1 Open the worksheet GRADES.MTW.

2 Choose **Stat > Basic Statistics > Correlation**.

3 In **Variables**, enter *Verbal Math GPA*. Click **OK**.

Session window output

Correlations: Verbal, Math, GPA

```
        Verbal    Math
Math     0.275
         0.000

GPA      0.322   0.194
         0.000   0.006


Cell Contents: Pearson correlation
               P-Value
```

Interpreting the results

Minitab displays the correlation for the lower triangle of the correlation matrix when there are more than two variables. The Pearson correlation between Math and Verbal is 0.275, between GPA and Verbal is 0.322, and between GPA and Math is 0.194. Minitab prints the p-values for the individual hypothesis tests of the correlations being zero below the correlations. Since all the p-values are smaller than 0.01, there is sufficient evidence at $\alpha = 0.01$ that the correlations are not zero, in part reflecting the large sample size of 200.

To calculate a partial correlation coefficient between two variables

1 Regress the first variable on the other variables and store the residuals – see Regression.

2 Regress the second variable on the other variables and store the residuals.

3 Calculate the correlation between the two columns of residuals.

Partial correlation coefficients

By using a combination of Minitab commands, you can also compute a partial correlation coefficient. This is the correlation coefficient between two variables while adjusting for the effects of other variables. Partial correlation coefficients can be used when you have multiple potential predictors, and you wish to examine the individual effect of predictors upon the response variable after taking into account the other predictors.

Example of a Partial Correlation

A survey was conducted in restaurants in 19 Wisconsin counties. Variables measured include: Sales (gross sales), Newcap (new capital invested), and Value (estimated market value of the business). All variables are measured in thousands of dollars.

We want to look at the relationship between sales and new capital invested removing the influence of market value of the business. First we calculate the regular Pearson correlation coefficient for comparison. Then we demonstrate calculating the partial correlation coefficient between sales and new capital.

Step 1: Calculate unadjusted correlation coefficients

1 Open the worksheet RESTRNT.MTW.

2 Choose **Stat > Basic Statistics > Correlation**.

3 In **Variables**, enter *Sales Newcap Value*. Click **OK**.

The remaining steps calculate partial correlation between Sales and Newcap.

Step 2: Regress Sales on Value and store the residuals (Resi1)

1 Choose **Stat > Regression > Regression**.

2 In **Response**, enter *Sales*. In **Predictors**, enter *Value*.

3 Click **Storage**, and check **Residuals**. Click **OK** in each dialog box.

Step 3: Regress Newcap on Value and store the residuals (Resi2)

1 Choose **Stat > Regression > Regression**.

2 In **Response**, enter *Newcap*. In **Predictors**, enter *Value*.

3 Click **OK**.

Step 4: Calculate correlations of the residual columns

1 Choose **Stat > Basic Statistics > Correlation**.

2 In **Variables**, enter *Resi1 Resi2*. Click **OK**.

Session window output

Correlations: Sales, Newcap, Value

```
         Sales  Newcap
Newcap   0.615
         0.000

Value    0.803   0.734
         0.000   0.000


Cell Contents: Pearson correlation
               P-Value
```

Session window output

Correlations: RESI1, RESI2

```
Pearson correlation of RESI1 and RESI2 = 0.078
P-Value = 0.261
```

Interpreting the results

The correlation between the residual columns is 0.078. In other words, after adjusting for the linear effect of Value, the correlation between Sales and Newcap is 0.078 – a value that is quite different from the uncorrected 0.615 value. In addition, the p-value of 0.261 indicates that there is no evidence that the correlation between Sales and Newcap – after accounting for the Value effect –i s different from zero.

You can repeat this example to obtain the partial correlation coefficients between other variables. The partial correlation between Sales and Value is 0.654; the partial correlation between Newcap and Value is 0.502.

Spearman's Rank Correlation

You can also use Correlation to obtain Spearman's ρ (rank correlation coefficient). Like the Pearson product moment correlation coefficient, Spearman's ρ is a measure of the relationship between two variables. However, Spearman's ρ is calculated on ranked data.

To calculate Spearman's ρ

1 Delete any rows that contain missing values.

2 If the data are not already ranked, use Data > Rank to rank them.

3 Compute the Pearson's correlation on the columns of ranked data using Stat > Basic Statistics > Correlation.

4 Uncheck **Display p-values**.

Caution The p-value Minitab gives is not accurate for Spearman's ρ. Do not use p-values to interpret Spearman's ρ.

Covariance

Covariance

Stat > Basic Statistics > Covariance

Calculates the covariance between each pair of columns.

You can calculate the covariance for all pairs of columns. Like the Pearson correlation coefficient, the covariance is a measure of the relationship between two variables. However, the covariance has not been standardized, as is done with the correlation coefficient. The correlation coefficient is standardized by dividing by the standard deviation of both variables.

Dialog box items

Variables: Enter the columns containing the variables you want to calculate the covariance for. When you list two columns, Minitab calculates the covariance for the pair. When you list more than two columns, Minitab calculates the covariance for every possible pair, and displays the lower triangle of the covariance matrix (in blocks if there is insufficient room to fit across a page).

Store matrix (display nothing): Check to store the covariance matrix. Minitab does not display the covariance matrix when you choose this option. To display the matrix, choose **Data > Display Data**.

Data – Covariance

Data must be in numeric columns of equal length.

Minitab omits missing data from calculations using a method that is often called pairwise deletion. Minitab omits from the calculations for each column pair only those rows that contain a missing value for that pair.

If you are calculating covariances between multiple columns at the same time, pairwise deletion may result in different observations being included in the various covariances. Although this method is the best for each individual covariance, the covariance matrix as a whole may not be well behaved (for example, it may not be positive definite).

To calculate the covariance

1 Choose **Stat > Basic Statistics > Covariance**.

2 In **Variables**, enter the columns containing the measurement data.

3 If you like, use the dialog box option, then click **OK**.

Normality Test

Normality Test

Stat > Basic Statistics > Normality Test

Generates a normal probability plot and performs a hypothesis test to examine whether or not the observations follow a normal distribution. For the normality test, the hypotheses are,

H_0: data follow a normal distribution vs. H_1: data do not follow a normal distribution

The vertical scale on the graph resembles the vertical scale found on normal probability paper. The horizontal axis is a linear scale. The line forms an estimate of the cumulative distribution function for the population from which data are drawn. Numerical estimates of the population parameters, μ and σ, the normality test value, and the associated p-value are displayed with the plot.

Dialog box items

Variable: Enter the column to use for the x-axis. Minitab calculates the probability of occurrence for each observation in the column (assuming a normal distribution) and uses the log of the calculated probabilities as y-values.

Percentile lines: Minitab marks each percent in the column with a horizontal reference line on the plot, and marks each line with the percent value. Minitab draws a vertical reference line where the horizontal reference line intersects the line fit to the data, and marks this line with the estimated data value.

> **None:** Choose to display no percentile line.
>
> **At Y values:** Choose to enter y-scale values for placing percentile lines. Enter values between 0 and 100 when percents are used as the y-scale type or 0 to 1 when probability is the y-scale type.
>
> **At data values:** Choose to enter data values for placing percentile lines.

Tests for Normality: See [3] and [10] for discussions of tests for normality.

> **Anderson-Darling:** Choose to perform an Anderson-Darling test for normality, an ECDF (empirical cumulative distribution function) based test.
>
> **Ryan-Joiner:** Choose to perform a Ryan-Joiner test, similar to the Shapiro-Wilk test. The Ryan-Joiner test is a correlation based test.
>
> **Kolmogorov-Smirnov:** Choose to perform a Kolmogorov-Smirnov test for normality, an ECDF based test.

Title: To replace the default title with your own custom title, type the desired text in this box.

Data – Normality Test

You need one numeric column. Minitab automatically omits missing data from the calculations.

To perform a normality test

1 Choose **Stat > Basic Statistics > Normality Test**.

2 In **Variables**, enter the columns containing the measurement data.

3 If you like, use any dialog box options, then click **OK**.

Example of Normality Test

In an operating engine, parts of the crankshaft move up and down. AtoBDist is the distance (in mm) from the actual (A) position of a point on the crankshaft to a baseline (B) position. To ensure production quality, a manager took five measurements each working day in a car assembly plant, from September 28 through October 15, and then ten per day from the 18th through the 25th.

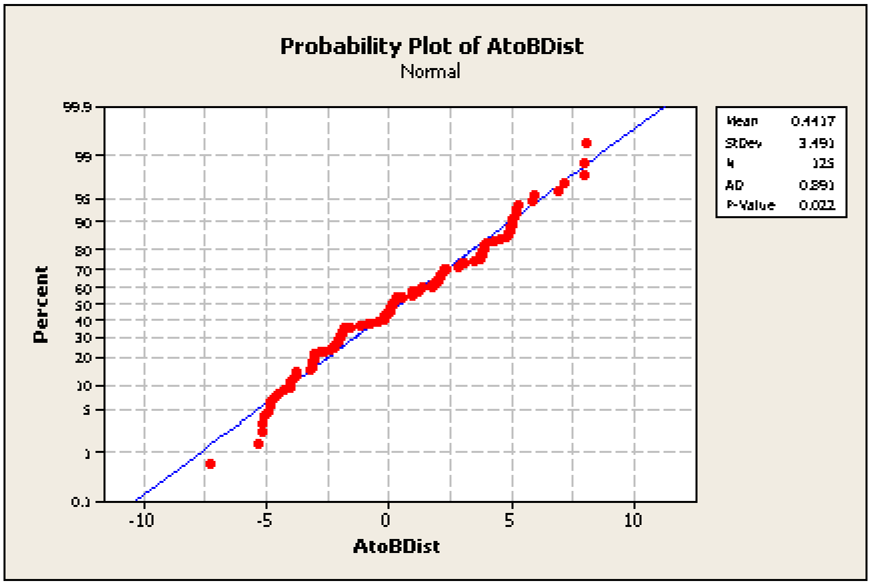
You wish to see if these data follow a normal distribution, so you use Normality test.

1 Open the worksheet CRANKSH.MTW.

2 Choose **Stat > Basic Statistics > Normality Test**.

3 In **Variable**, enter *AtoBDist*. Click **OK**.

Graph window output



Interpreting the results

The graphical output is a plot of normal probabilities versus the data. The data depart from the fitted line most evidently in the extremes, or distribution tails. The Anderson-Darling test's p-value indicates that, at α levels greater than 0.022, there is evidence that the data do not follow a normal distribution. There is a slight tendency for these data to be lighter in the tails than a normal distribution because the smallest points are below the line and the largest point is just above the line. A distribution with heavy tails would show the opposite pattern at the extremes.

Choosing a normality test

You have a choice of hypothesis tests for testing normality:

- Anderson-Darling test (the default), which is an ECDF (empirical cumulative distribution function) based test
- Ryan-Joiner test [4], [9] (similar to the Shapiro-Wilk test [10], [11]) which is a correlation based test
- Kolmogorov-Smirnov test [8], an ECDF based test

The Anderson-Darling and Ryan-Joiner tests have similar power for detecting non-normality. The Kolmogorov-Smirnov test has lesser power–see [3], [8]and [9] for discussions of these tests for normality.

The common null hypothesis for these three tests is H_0: data follow a normal distribution. If the p-value of the test is less than your α level, reject H_0.

Regression

Overview

Regression Overview

Regression analysis is used to investigate and model the relationship between a response variable and one or more predictors. Minitab provides least squares, partial least squares, and logistic regression procedures:

- Use least squares procedures when your response variable is continuous.
- Use partial least squares regression when your predictors are highly correlated or outnumber your observations.
- Use logistic regression when your response variable is categorical.

Both least squares and logistic regression methods estimate parameters in the model so that the fit of the model is optimized. Least squares regression minimizes the sum of squared errors to obtain parameter estimates, whereas Minitab's logistic regression obtains maximum likelihood estimates of the parameters. See Logistic Regression Overview for more information. Partial least squares (PLS) extracts linear combinations of the predictors to minimize prediction error. See Partial Least Squares Overview for more information.

Use the table below to select a procedure:

Use...	To...	Response type	Estimation method
Regression	perform simple, multiple regression or polynomial least squares regression	continuous	least squares
Stepwise	perform stepwise, forward selection, or backward elimination to identify a useful subset of predictors	continuous	least squares
Best Subsets	identify subsets of the predictors based on the maximum R^2 criterion	continuous	least squares
Fitted Line Plot	perform linear and polynomial regression with a single predictor and plot a regression line through the data	continuous	least squares
PLS	perform regression with ill-conditioned data	continuous	biased, non-least squares
Binary Logistic	perform logistic regression on a response with only two possible values, such as presence or absence	categorical	maximum likelihood
Ordinal Logistic	perform logistic regression on a response with three or more possible values that have a natural order, such as none, mild, or severe	categorical	maximum likelihood
Nominal Logistic	perform logistic regression on a response with three or more possible values that have no natural order, such as sweet, salty, or sour	categorical	maximum likelihood

Regression commands

Stat > Regression

Select one of the following commands to fit a model relating a response to one or more predictors :

Regression – does simple, multiple and polynomial regression

Stepwise – does stepwise regression, forward selection, and backward elimination

Best Subsets – does best subsets regression

Fitted Line Plot – fits a simple linear or polynomial regression model and plots the regression line through the actual data or the $\log_{10}$ of the data

Partial Least Squares – does partial least squares regression

Binary Logistic Regression – does logistic regression for a binary response variable

Ordinal Logistic Regression – does logistic regression for an ordinal response variable

Nominal Logistic Regression – does logistic regression for a nominal response variable

Regression Examples

The following examples illustrate how to use the various regression techniques available. Choose an example below:

Simple Linear Regression
Multiple Regression
Stepwise Regression
Best Subsets Regression
Fitted Line Plot
Partial Least Squares Regression
Binary Logistic Regression
Ordinal Logistic Regression
Nominal Logistic Regression

References – Regression

[1] A. Agresti (1984). *Analysis of Ordinal Categorical Data.* John Wiley & Sons, Inc.

[2] A. Agresti (1990). *Categorical Data Analysis.* John Wiley & Sons, Inc.

[3] D.A. Belsley, E. Kuh, and R.E. Welsch (1980). *Regression Diagnostics.* John Wiley & Sons, Inc.

[4] A. Bhargava (1989). "Missing Observations and the Use of the Durbin-Watson Statistic," *Biometrik,* 76, 828–831.

[5] C.C. Brown (1982). "On a Goodness of Fit Test for the Logistic Model Based on Score Statistics," *Communications in Statistics*, 11, 1087–1105.

[6] D.A. Burn and T.A. Ryan, Jr. (1983). "A Diagnostic Test for Lack of Fit in Regression Models," *ASA 1983 Proceedings of the Statistical Computing Section*, 286–290.

[7] R.D. Cook (1977). "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15–18.

[8] R.D. Cook and S. Weisberg (1982). *Residuals and Influence in Regression.* Chapman and Hall.

[9] N.R. Draper and H. Smith (1981). *Applied Regression Analysis*, Second Edition. John Wiley & Sons, Inc.

[10] S.E. Fienberg (1987). *The Analysis of Cross-Classified Categorical Data*. The MIT Press.

[11] I.E. Frank and J.H. Friedman (1993). "A Statistical View of Some Chemometrics Regression Tool," *Technometrics*, 35, 109–135.

[12] I.E. Frank and B.R. Kowalski (1984). "Prediction of Wine Quality and Geographic Origin from Chemical Measurements by Partial Least-Squares Regression Modeling," *Analytica Chimica Acta*, 162, 241–251.

[13] M.J. Garside (1971). "Some Computational Procedures for the Best Subset Problem," *Applied Statistics*, 20, 8–15.

[14] P. Geladi and B. Kowalski (1986). "Partial Least-Squares Regression: A Tutorial," *Analytica Chimica Acta,* 185, 1–17.

[15] P. Geladi and B. Kowalski (1986). "An Example of 2-Block Predictive Partial Least-Squares Regression with Simulated Data," *Analytica Chimica Acta,* 185, 19-32.

[16] James H. Goodnight (1979). "A Tutorial on the Sweep Operator," *The American Statistician,* 33, 149–158.

[17] W.W. Hauck and A. Donner (1977). "Wald's test as applied to hypotheses in logit analysis," *Journal of the American Statistical Association*, 72, 851-853.

[18] D.C. Hoaglin and R.E. Welsch (1978). "The Hat Matrix in Regression and ANOVA," *The American Statistician*, 32, 17–22.

[19] R.R. Hocking (1976). "A Biometrics Invited Paper: The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1–49.

[20] A. Hoskuldsson (1988). "PLS Regression Methods," *Journal of Chemometrics*, 2, 211–228.

[21] D.W. Hosmer and S. Lemeshow (2000). *Applied Logistic Regression.* 2nd ed. John Wiley & Sons, Inc.

[22] LINPACK (1979). Linpack User's Guide by J.J. Dongarra, J.R. Bunch, C.B. Moler, and G.W. Stewart, Society for Industrial and Applied Mathematics, Philadelphia, PA.

[23] A. Lorber, L. Wangen, and B. Kowalski (1987). "A Theoretical Foundation for the PLS Algorithm," *Journal of Chemometrics*, 1, 19–31.

[24] J.H. Maindonald (1984). *Statistical Computation.* John Wiley & Sons, Inc.

[25] P. McCullagh and J.A. Nelder (1992). *Generalized Linear Model*. Chapman & Hall.

[26] W. Miller (1978). "Performing Armchair Roundoff Analysis of Statistical Algorithms," *Communications in Statistics*, 243–255.

[27] D.C. Montgomery and E.A. Peck (1982). *Introduction to Linear Regression Analysis*. John Wiley & Sons.

[28] J. Neter, W. Wasserman, and M. Kutner (1985). *Applied Linear Statistical Models*. Richard D. Irwin, Inc.

[29] S.J. Press and S. Wilson (1978). "Choosing Between Logistic Regression and Discriminant Analysis," *Journal of the American Statistical Association,* 73, 699-705.

[30] M. Schatzoff, R. Tsao, and S. Fienberg (1968). "Efficient Calculation of All Possible Regressions," *Technometrics*, 10, 769–779.

[31] G.W. Stewart (1973). *Introduction to Matrix Computations*. Academic Press.

[32] R.A. Thisted (1988). *Elements of Statistical Computing: Numerical Computation*. Chapman & Hall.

[33] P. Velleman and R. Welsch (1981). "Efficient Computation of Regression Diagnostics," *The American Statistician*, 35, 234–242.

[34] P.F. Velleman, J. Seaman, and I.E. Allen (1977). "Evaluating Package Regression Routines," *ASA 1977 Proceedings of the Statistical Computing Section*.

[35] S. Weisberg (1980). *Applied Linear Regression*. John Wiley & Sons, Inc.

[36] H. Wold (1975). "Soft Modeling by Latent Variables; the Nonlinear Iterative Partial Least Squares Approach," in *Perspectives in Probability and Statistics, Papers in Honour of M.S. Bartlett*, ed. J. Gani, Academic Press.

Acknowledgments

We are very grateful for help in the design of the regression algorithm from W. Miller and from G.W. Stewart and for useful suggestions from. P.F. Velleman and S. Weisberg and many others.

Regression

Regression

Stat > Regression > Regression

You can use Regression to perform simple and multiple regression using least squares. Use this procedure for fitting general least squares models, storing regression statistics, examining residual diagnostics, generating point estimates, generating prediction and confidence intervals, and performing lack-of-fit tests.

You can also use this command to fit polynomial regression models. However, if you want to fit a polynomial regression model with a single predictor, you may find it simpler to use Fitted Line Plot.

Dialog box items

Response: Select the column containing the Y, or response variable.

Predictors: Select the column(s) containing the X, or predictor variable(s).

<Graphs>

<Options>

<Results>

<Storage>

Data – Regression

Enter response and predictor variables in numeric columns of equal length so that each row in your worksheet contains measurements on one observation or subject.

Minitab omits all observations that contain missing values in the response or in the predictors, from calculations of the regression equation and the ANOVA table items.

If you have variables which are either closely related or are nearly constant, Minitab may judge them as ill-conditioned and omit some or all of them from the model. If your data are ill-conditioned, you can modify how Minitab handles these variables by using the TOLERANCE command in the Session window. See Ill-Conditioned Data for details.

To do a linear regression

1 Choose **Stat > Regression > Regression**.

2 In **Response**, enter the column containing the response (Y) variable.

3 In **Predictors**, enter the columns containing the predictor (X) variables.

4 If you like, use one or more of the dialog box options, then click **OK**.

Regression – Graphs

Stat > Regression > Regression > Graphs

Display residual plots for diagnosis of the regression model fit.

Dialog box items

Residuals for Plots: You can specify the type of residual to display on the residual plots.

Regular: Choose to plot the regular or raw residuals.

Standardized: Choose to plot the standardized residuals.

Deleted: Choose to plot the Studentized deleted residuals.

Residual Plots

Individual plots: Choose to display one or more plots.

Histogram of residuals: Check to display a histogram of the residuals.

Normal plot of residuals: Check to display a normal probability plot of the residuals.

Residuals versus fits: Check to plot the residuals versus the fitted values.

Residuals versus order: Check to plot the residuals versus the order of the data. The row number for each data point is shown on the x-axis–for example, 1 2 3 4... n.

Four in one: Choose to display a layout of a histogram of residuals, a normal plot of residuals, a plot of residuals versus fits, and a plot of residuals versus order.

Residuals versus the variables: Enter one or more columns containing the variables against which you want to plot the residuals. Minitab displays a separate graph for each column.

Choosing a Residual Type

You can calculate three types of residuals. Use the table below to help you choose which type you would like to plot:

Residual type...	Choose when you want to...	Calculation
regular	examine residuals in the original scale of the data	response - fit
standardized	use a rule of thumb for identifying observations that are not fit well by the model. A standardized residual greater than 2, in absolute value, might be considered to be large. Minitab displays these observations in a table of unusual observations, labeled with an R.	(residual) / (standard deviation of the residual)
deleted Studentized	identify observations that are not fit well by the model. Removing observations can affect the variance estimate and also can affect parameter estimates. A large absolute Studentized residual may indicate that including the observation in the model increases the error variance or that it has a large affect upon the parameter estimates, or both.	(residual) / (standard deviation of the residual). The ith Studentized residual is computed with the ith observation removed.

Residual Plot Choices

Minitab generates residual plots that you can use to examine the goodness of model fit. You can choose the following residual plots:

- **Histogram of residuals.** An exploratory tool to show general characteristics of the data, including:
  - Typical values, spread or variation, and shape
  - Unusual values in the data

  Long tails in the plot may indicate skewness in the data. If one or two bars are far from the others, those points may be outliers. Because the appearance of the histogram changes depending on the number of intervals used to group the data, use the normal probability plot and goodness-of-fit tests to assess the normality of the residuals.

- **Normal plot of residuals.** The points in this plot should generally form a straight line if the residuals are normally distributed. If the points on the plot depart from a straight line, the normality assumption may be invalid. If your data have fewer than 50 observations, the plot may display curvature in the tails even if the residuals are normally distributed. As the number of observations decreases, the probability plot may show substantial variation and nonlinearity even if the residuals are normally distributed. Use the probability plot and goodness-of-fit tests, such as the Anderson-Darling statistic, to assess whether the residuals are normally distributed.

You can display the Anderson-Darling statistic (AD) on the plot, which can indicate whether the data are normal. If the p-value is lower than the chosen α-level, the data do not follow a normal distribution. To display the Anderson-Darling statistic, choose Tools > Options > Individual Graphs > Residual Plots. For additional tests of normality, see Stat > Basic Statistics > Normality Test.

- **Residuals versus fits.** This plot should show a random pattern of residuals on both sides of 0. If a point lies far from the majority of points, it may be an outlier. Also, there should not be any recognizable patterns in the residual plot. The following may indicate error that is not random:
  - a series of increasing or decreasing points
  - a predominance of positive residuals, or a predominance of negative residuals
  - patterns, such as increasing residuals with increasing fits
- **Residuals versus order.** This is a plot of all residuals in the order that the data was collected and can be used to find non-random error, especially of time-related effects. A positive correlation is indicated by a clustering of residuals with the same sign. A negative correlation is indicated by rapid changes in the signs of consecutive residuals.
- **Four in one.** Select this option to produce a normal plot of residuals, a histogram of residuals, a plot of residuals versus fits, and a plot of residuals versus order in one graph window.
- **Residuals versus other variables.** This is a plot of all residuals versus another variable. Plot the residuals against:
  - Each predictor to look for curvature or differences in the magnitude of the residuals
  - Important variables left out of the model to see if they have critical additional effects on the response.

If certain residual values are of concern, you can brush your graph to identify them. See graph brushing.

Checking Your Model

Regression analysis does not end once the regression model is fit. You should examine residual plots and other diagnostic statistics to determine whether your model is adequate and the assumptions of regression have been met. If your model is inadequate, it will not correctly represent your data. For example:

- The standard errors of the coefficients may be biased, leading to incorrect t- and p-values.
- Coefficients may have the wrong sign.
- The model may be overly influenced by one or two points.

Use the table below to determine whether your model is adequate.

Characteristics of an adequate regression model	Check using...	Possible solutions
Linear relationship between response and predictors.	Lack-of-fit-tests Residuals vs variables plot	• Add higher-order term to model. • Transform variables.
Residuals have constant variance.	Residuals vs fits plot	• Transform variables. • Weighted least squares.
Residuals are independent of (not correlated with) one another.	Durbin-Watson statistic Residuals vs order plot	• Add new predictor. • Use time series analysis. • Add lag variable.
Residuals are normally distributed.	Histogram of residuals Normal plot of residuals Residuals vs fit plot Normality test	• Transform variables. • Check for outliers.
No unusual observations or outliers.	Residual plots Leverages Cook's distance DFITS	• Transform variables. • Remove outlying observation.
Data are not ill-conditioned.	Variance inflation factor (VIF) Correlation matrix of predictors	• Remove predictor. • Partial least squares regression. • Transform variables.

If you determine that your model does not meet the criteria listed above, you should :

1 Check to see whether your data are entered correctly, especially observations identified as unusual.

2 Try to determine the cause of the problem. You may want to see how sensitive your model is to the issue. For example, if you have an outlier, run the regression without that observation and see how the results differ.

3 Consider using one of the possible solutions listed above. See [9], [28] for more information.

Regression – Options

Stat > Regression > Regression > Options

Use to perform weighted regression, fit the model with or without an intercept, calculate variance inflation factors and the Durbin-Watson statistic, and calculate and store prediction intervals for new observations.

Dialog box items

Weights: Enter a column of weights to perform weighted regression.

Fit Intercept: Check to fit a constant term (the y-intercept of the regression line). Uncheck to fit the model without a constant term. Minitab does not display R^2 for this model.

Display

Variance inflation factors: Check to display variance inflation factors (VIF) to check for multicollinearity effects associated with each predictor.

Durbin-Watson statistic: Check to display the Durbin-Watson statistic to detect autocorrelation in the residuals.

PRESS and predicted R-square: Check to display the PRESS statistic and predicted R-square.

Lack of Fit Tests

Pure error: Check to perform a pure error lack-of-fit test for testing model adequacy if your data contain replicates.

Data subsetting: Check to perform a data subsetting lack-of-fit test to test the model adequacy.

Prediction intervals for new observations: Type the numeric predictor values, or enter the columns or constants in which they are stored. The number of predictors must equal the number of predictors in the model. See To predict responses for new observations.

Confidence level: Type the desired confidence level (for example, type 90 for 90%). The default is 95%.

Storage

Fits: Check to store the fitted values for new observations.

SEs of fits: Check to store the estimated standard errors of the fits.

Confidence limits: Check to store the lower and upper limits of the confidence interval.

Prediction limits: Check to store the lower and upper limits of the prediction interval.

To perform weighted regression

1 Choose **Stat > Regression > Regression > Options**.

2 In **Weights**, enter the column containing the weights, which must be greater than or equal to zero. Click **OK** in each dialog box.

Note If you use weighted regression and predict responses for new observations to obtain the prediction interval, see Calculating the prediction interval with weighted regression.

Ill-Conditioned Data

Ill-conditioned data relates to problems in the predictor variables, which can cause both statistical and computational difficulties. There are two types of problems: multicollinearity and a small coefficient of variation. The checks for ill-conditioned data in Minitab have been heavily influenced by Velleman et al. [33], [34].

Multicollinearity

Multicollinearity means that some predictors are correlated with other predictors. If this correlation is high, Minitab displays a warning message and continues computation. The predicted values and residuals still are computed with high statistical and numerical accuracy, but the standard errors of the coefficients will be large and their numerical accuracy may be affected. If the correlation of a predictor with other predictors is very high, Minitab eliminates the predictor from the model, and displays a message.

To identify predictors that are highly collinear, you can examine the correlation structure of the predictor variables and regress each suspicious predictor on the other predictors. You can also review the variance inflation factors (VIF), which measure how much the variance of an estimated regression coefficient increases if your predictors are correlated. If the

VIF < 1, there is no multicollinearity but if the VIF is > 1, predictors may be correlated. Montgomery and Peck suggest that if the VIF is 5 – 10, the regression coefficients are poorly estimated.

Some possible solutions to the problem of multicollinearity are:

- Eliminate predictors from the model, especially if deleting them has little effect on R^2.
- Change predictors by taking linear combinations of them using partial least squares regression or principal components analysis.
- If you are fitting polynomials, subtract a value near the mean of a predictor before squaring it.

Small coefficient of variation

Predictors with small coefficients of variation and are nearly constant, which can cause numerical problems. For example, the variable YEAR with values from 1970 to 1975 has a small coefficient of variation and numerical differences among the variables are contained in the fourth digit. The problem is compounded if YEAR is squared. You could subtract a constant from the data, replacing YEAR with YEARS SINCE 1970, which has values 0 to 5.

If the coefficient of variation is moderately small, some loss of statistical accuracy will occur. In this case, Minitab tells you that the predictor is nearly constant. If the coefficient of variation is very small, Minitab eliminates the predictor from the model, and displays a message.

More If your data are extremely ill-conditioned, Minitab removes one of the problematic columns from the model. You can use the TOLERANCE subcommand with REGRESS to force Minitab to keep that column in the model. Lowering the tolerance can be dangerous, possibly producing numerically inaccurate results. See Session command help for more information.

Detecting Autocorrelation in Residuals

In linear regression, it is assumed that the residuals are independent of (not correlated with) one another. If the independence assumption is violated, some model fitting results may be questionable. For example, positive correlation between error terms tends to inflate the t-values for coefficients, making predictors appear significant when they may not be.

Minitab provides two methods to determine if residuals are correlated:

- A graph of residuals versus data order (1 2 3 4... n) can provides a means to visually inspect residuals for autocorrelation. A positive correlation is indicated by a clustering of residuals with the same sign. A negative correlation is indicated by rapid changes in the signs of consecutive residuals.
- The Durbin-Watson statistic tests for the presence of autocorrelation in regression residuals by determining whether or not the correlation between two adjacent error terms is zero. The test is based upon an assumption that errors are generated by a first-order autoregressive process. If there are missing observations, these are omitted from the calculations, and only the nonmissing observations are used.

  To reach a conclusion from the test, you will need to compare the displayed statistic with lower and upper bounds in a table. If D > upper bound, no correlation exists; if D < lower bound, positive correlation exists; if D is in between the two bounds, the test is inconclusive. For additional information, see [4], [28].

To predict responses for new observations in regression

1 Choose **Stat > Regression > Regression > Options**.

2 In **Prediction intervals for new observations**, do any combination of the following:

   - Type numeric predictor values.
   - Enter stored constants containing numeric predictor values.
   - Enter columns of equal length containing numeric predictor values.

   The number of predictors and the order in which they are entered must match your original model. If you enter a constant and column(s), Minitab will assume that you want predicted values for all combinations of constant and column values.

4 In **Confidence level**, type a value or use the default, which is 95%.

5 Under **Storage**, check any of the prediction results to store them in your worksheet. Click **OK**.

Note If you use weighted regression and predict responses for new observations to obtain the prediction interval, see Calculating the prediction interval with weighted regression.

Regression – Results

Stat > Regression > Regression > Results

Control the display of output to the Session window.

Dialog box items

Control the Display of Results

Display nothing: Check to display nothing.

Regression equation, table of coefficients, s, R-squared, and basic analysis of variance: Check to display some basic regression output.

In addition, sequential sums of squares and the unusual observations in the table of fits and residuals: Check to display, in addition to the above, sequential sums of squares (added sums of squares explained by each additional predictor) and a table of unusual values.

In addition, the full table of fits and residuals: Check to display, in addition to the above, a table of fits and residuals for all observations.

Regression – Storage

Stat > Regression > Regression > Storage

You can store diagnostic measures and characteristics of the estimated regression equation.

Dialog box items

Diagnostic Measures

Residuals: Check to store the residuals.

Standard residuals.: Check to store the standardized residuals.

Deleted t residuals.: Check to store the Studentized residuals.

Hi (leverages): Check to store the leverages.

Cook's distance: Check to store Cook's distance.

DFITS: Check to store the DFITS.

Characteristics of Estimated Equation

Coefficients: Check to store the estimated coefficients of the regression equation.

Fits: Check to store the fitted values.

MSE: Check to store the mean square error. (This is also displayed in the analysis of variance table, under MS). Note that the square root of MSE equals s, which is also included in the output.

X'X inverse: Check to store inverse of **X'X** in a matrix. This matrix, when multiplied by MSE is the variance-covariance matrix of the coefficients. If you do a weighted regression (see Options) then this option stores the inverse of the **X'WX** matrix. See X'X inverse. To view the matrix, choose Data > Display Data.

R matrix: Check to store the **R** matrix of the **QR** or Cholesky decomposition. See R matrix. To view the matrix, choose Data > Display Data.

Identifying Outliers

Outliers are observations with larger than average response or predictor values. Minitab provides several ways to identify outliers, including residual plots and three stored statistics: leverages, Cook's distance, and DFITS, which are described below. It is important to identify outliers because they can significantly influence your model, providing potentially misleading or incorrect results. If you identify an outlier in your data, you should examine the observation to understand why it is unusual and identify an appropriate remedy.

- Leverage values provide information about whether an observation has unusual predictor values compared to the rest of the data. Leverages are a measure of the distance between the x-values for an observation and the mean of x-values for all observations. A large leverage value indicates that the x-values of an observation are far from the center of x-values for all observations. Observations with large leverage may exert considerable influence on the fitted value, and thus the regression model.

  Leverage values fall between 0 and 1. A leverage value greater than 2p/n or 3p/n, where p is the number of predictors plus the constant and n is the number of observations, is considered large and should be examined. Minitab identifies observations with leverage over 3p/n or .99, whichever is smaller, with an X in the table of unusual observations.

- Cook's distance or D is an overall measure of the combined impact of each observation on the fitted values. Because D is calculated using leverage values and standardized residuals, it considers whether an observation is unusual with respect to both x- and y-values. Geometrically, Cook's distance is a measure of the distance between the fitted values calculated with and without the i^{th} observation. Large values, which signify unusual observations, can occur because the observation has 1) a large residual and moderate leverage, 2) a large leverage and moderate residual, or 3) a large residual and leverage. Some statisticians recommend comparing D to the F-distribution (p, n-p). If D is greater than the F-value at the 50th percentile, then D is considered extreme and should be examined. Other statisticians recommend comparing the D statistics to one another, identifying values that are extremely large relative to the others. An easy way to compare D values is to graph them using Graph >Time Series, where the x-axis represents the observations, not an index or time period.

- DFITS provides another measure to determine whether an observation is unusual. It uses the leverage and deleted (Studentized) residual to calculate the difference between the fitted value calculated with and without the i^{th} observation. DFITS represents roughly the number of estimated standard deviations that the fitted value changes when the i^{th} observation is removed from the data. Some statisticians suggest that an observation with a DFITS value greater than sqrt(2p/n) is influential. Other statisticians recommend comparing DFITS values to one another, identifying values that are extremely large relative to the others. An easy way to compare DFITS is to graph the DFITS values using Graph >Time Series, where the x-axis represents the observations, not an index or time period.

Example of simple linear regression

You are a manufacturer who wants to obtain a quality measure on a product, but the procedure to obtain the measure is expensive. There is an indirect approach, which uses a different product score (Score 1) in place of the actual quality measure (Score 2). This approach is less costly but also is less precise. You can use regression to see if Score 1 explains a significant amount of variance in Score 2 to determine if Score 1 is an acceptable substitute for Score 2.

1 Open the worksheet EXH_REGR.MTW.

2 Choose **Stat > Regression > Regression**.

3 In **Response**, enter *Score2.*

4 In **Predictors**, enter *Score1.*

5 Click **OK.**

Session window output

Regression Analysis: Score2 versus Score1

```
The regression equation is
Score2 = 1.12 + 0.218 Score1


Predictor     Coef  SE Coef      T      P
Constant    1.1177   0.1093  10.23  0.000
Score1     0.21767  0.01740  12.51  0.000


S = 0.127419   R-Sq = 95.7%   R-Sq(adj) = 95.1%


Analysis of Variance

Source          DF      SS      MS       F      P
Regression       1  2.5419  2.5419  156.56  0.000
Residual Error   7  0.1136  0.0162
Total            8  2.6556


Unusual Observations

Obs  Score1  Score2     Fit  SE Fit  Residual  St Resid
  9    7.50  2.5000  2.7502  0.0519   -0.2502     -2.15R

R denotes an observation with a large standardized residual.
```

Interpreting the results

Minitab displays the results in the Session window by default.

- The p-value in the Analysis of Variance table (0.000), indicates that the relationship between Score 1 and Score 2 is statistically significant at an a-level of .05. This is also shown by the p-value for the estimated coefficient of Score 1, which is 0.000.
- The R^2 value shows that Score 1 explains 95.7% of the variance in Score 2, indicating that the model fits the data extremely well.
- Observation 9 is identified as an unusual observation because its standardized residual is less than –2. This could indicate that this observation is an outlier. See Identifying outliers.

- Because the model is significant and explains a large part of the variance in Score 2, the manufacturer decides to use Score 1 in place of Score 2 as a quality measure for the product.

Example of multiple regression

As part of a test of solar thermal energy, you measure the total heat flux from homes. You wish to examine whether total heat flux (HeatFlux) can be predicted by insulation, by the position of the focal points in the east, south, and north directions, and by the time of day. Data are from [27]. You found, using best subsets regression, that the best two-predictor model included the variables North and South and the best three-predictor added the variable East. You evaluate the three-predictor model using multiple regression.

1 Open the worksheet EXH_REGR.MTW.

2 Choose **Stat > Regression > Regression**.

3 In **Response**, enter *HeatFlux*.

4 In **Predictors**, enter *East South North*.

5 Click **Graphs.**

6 Under **Residuals for Plots,** choose **Standardized.**

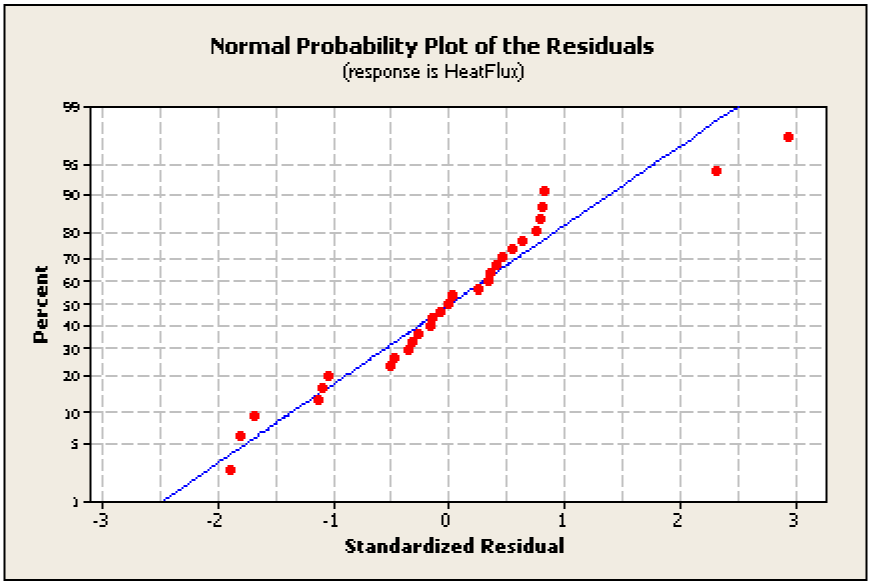
7 Under **Residual Plots**, choose **Individual Plots**. Check **Histogram of residuals**, **Normal plot of residuals**, and **Residuals versus fits**. Click **OK**.

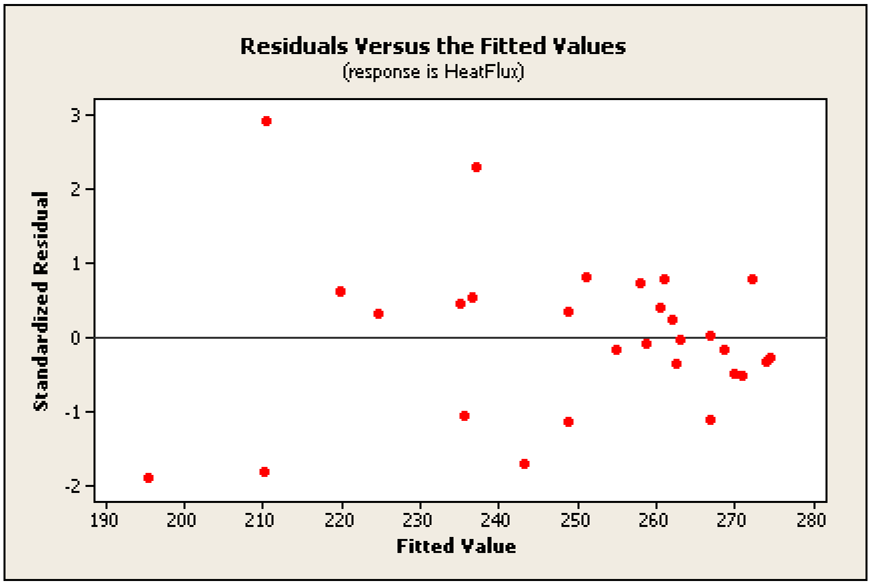
8 Click **Options**. Under **Display**, check **PRESS** and **predicted R-square**. Click **OK** in each dialog box.

Session window output

Regression Analysis: HeatFlux versus East, South, North

```
The regression equation is
HeatFlux = 389 + 2.12 East + 5.32 South - 24.1 North


Predictor     Coef  SE Coef       T      P
Constant    389.17    66.09    5.89  0.000
East         2.125    1.214    1.75  0.092
South       5.3185   0.9629    5.52  0.000
North      -24.132    1.869  -12.92  0.000


S = 8.59782   R-Sq = 87.4%   R-Sq(adj) = 85.9%

PRESS = 3089.67   R-Sq(pred) = 78.96%


Analysis of Variance

Source          DF       SS      MS      F      P
Regression       3  12833.9  4278.0  57.87  0.000
Residual Error  25   1848.1    73.9
Total           28  14681.9


Source  DF   Seq SS
East     1    153.8
South    1    349.5
North    1  12330.6


Unusual Observations

Obs  East  HeatFlux     Fit  SE Fit  Residual  St Resid
  4  33.1    230.70  210.20    5.03     20.50     2.94R
 22  37.8    254.50  237.16    4.24     17.34     2.32R

R denotes an observation with a large standardized residual.
```





Interpreting the results

Session window output

- The p-value in the Analysis of Variance table (0.000) shows that the model estimated by the regression procedure is significant at an α-level of 0.05. This indicates that at least one coefficient is different from zero.
- The p-values for the estimated coefficients of North and South are both 0.000, indicating that they are significantly related to HeatFlux. The p-value for East is 0.092, indicating that it is not related to HeatFlux at an α-level of 0.05. Additionally, the sequential sum of squares indicates that the predictor East doesn't explain a substantial amount of unique variance. This suggests that a model with only North and South may be more appropriate.
- The R^2 value indicates that the predictors explain 87.4% of the variance in HeatFlux. The adjusted R^2 is 85.9%, which accounts for the number of predictors in the model. Both values indicate that the model fits the data well.
- The predicted R^2 value is 78.96%. Because the predicted R^2 value is close to the R^2 and adjusted R^2 values, the model does not appear to be overfit and has adequate predictive ability.
- Observations 4 and 22 are identified as unusual because the absolute value of the standardized residuals are greater than 2. This may indicate they are outliers. See Checking your model, Identifying outliers, and Choosing a residual type.

Graph window output

- The histogram of the residuals exhibits a pattern consistent with a normal distribution. The histogram is most effective with data sets with more than 50 observations. The normal probability plot is easier to interpret for smaller samples.

- The normal probability plot shows an approximately linear pattern consistent with a normal distribution. The two points in the upper-right corner of the plot may be outliers. Brushing the graph identifies these points as 4 and 22, the same points that are labeled unusual observations in the output. See Checking your model and Identifying outliers.
- The plot of residuals versus the fitted values shows that the residuals get smaller (closer to the reference line) as the fitted values increase, which may indicate the residuals have nonconstant variance. See [9] for information on nonconstant variance.

Stepwise

Stepwise Regression

Stat > Regression > Stepwise

Stepwise regression removes and adds variables to the regression model for the purpose of identifying a useful subset of the predictors. Minitab provides three commonly used procedures: standard stepwise regression (adds and removes variables), forward selection (adds variables), and backward elimination (removes variables).

- When you choose the stepwise method, you can enter a starting set of predictor variables in **Predictors in initial model**. These variables are removed if their p-values are greater than the **Alpha to enter** value. If you want keep variables in the model regardless of their p-values, enter them in **Predictors to include in every model** in the main dialog box.
- When you choose the stepwise or forward selection method, you can set the value of the α for entering a new variable in the model in **Alpha to enter**.
- When you choose the stepwise or backward elimination method, you can set the value of α for removing a variable from the model in **Alpha to remove**.

See Using automatic selection procedures for a discussion of potential problems with stepwise regression.

Dialog box items

Response: Enter the Y, or response variable.

Predictors: Enter the columns containing the X, or predictor variables, to include in the model.

Predictors to include in every model: Indicate which (if any) predictors should never be removed from the model.

<Methods>

<Options>

Data – Stepwise Regression

Enter response and predictor variables in the worksheet in numeric columns of equal length so that each row in your worksheet contains measurements on one observation or subject. Minitab automatically omits rows with missing values from the calculations.

To do a stepwise regression

1. Choose **Stat > Regression > Stepwise**.
2. In **Response**, enter the numeric column containing the response (Y) data.
3. In **Predictors**, enter the numeric columns containing the predictor (X) variables.
4. If you like, use one or more of the dialog box options, then click **OK**.

Stepwise – Methods

Stat > Regression > Stepwise > Methods

Perform standard stepwise regression (adds and removes variables), forward selection (adds variables), or backward elimination (removes variables).

Dialog box items

Use alpha value: Choose to the use the alpha value as the criterion for adding or removing a variable to or from the model.

Use F values: Choose to use the F value as the criterion for adding or removing a variable to or from the model.

Stepwise (forward and backward): Choose standard stepwise regression.

Predictors in initial model: enter a starting set of predictor variables. These variables are removed if their p-values are greater than the **Alpha to enter** value. (If you want keep variables in the model regardless of their p-values, enter them in **Predictors to include in every model** in the main dialog box.)

Alpha to enter: Set the value of α for entering a new variable in the model.

F to enter: Set the value of F for entering a new variable in the model.

Alpha to remove: Set the value of α for removing a variable from the model.

F to remove: Set the value of F for removing a variable from the model.

Forward selection: Choose forward selection.

Alpha to enter: Set the value of α for entering a new variable in the model.

F to enter: Set the value of F for entering a new variable in the model.

Backward elimination: Choose backward elimination.

Alpha to remove: Set the value of α for removing a variable from the model.

F to remove: Set the value of F for removing a variable from the model.

Using Automatic Variable Selection Procedures

Variable selection procedures can be a valuable tool in data analysis, particularly in the early stages of building a model. At the same time, these procedures present certain dangers. Here are some considerations:

- Since the procedures automatically "snoop" through many models, the model selected may fit the data "too well." That is, the procedure can look at many variables and select ones which, by pure chance, happen to fit well.
- The three automatic procedures are heuristic algorithms, which often work very well but which may not select the model with the highest R^2 value (for a given number of predictors).
- Automatic procedures cannot take into account special knowledge the analyst may have about the data. Therefore, the model selected may not be the best from a practical point of view.

Stepwise – Options

Stat > Regression > Stepwise > Options

You can set the number of alternative predictors to display, indicate the number of steps between pauses so you can intervene, indicate whether to fit the intercept, or display predictive statistics in the Session window.

Dialog box items

Number of alternative predictors to show: Type a number to display the next best alternate predictors up to the number requested. If a new predictor is entered into the model, displays the predictor which was the second best choice, the third best choice, and so on.

Number of steps between pauses: Set the number of steps between prompts. This number can start at one with the default and maximum determined by the output width. Set a smaller value if you wish to intervene interactively more often. See Intervening in stepwise regression.

Fit intercept: Uncheck to exclude the intercept term from the regression model.

Display PRESS and predicted R-square: Check to display the PRESS statistic and predicted R^2.

Intervening in stepwise regression

Stepwise proceeds automatically by steps and then pauses. You can set the number of steps between pauses in the Options subdialog box.

The number of steps can start at one with the default and maximum determined by the output width. Set a smaller value if you wish to intervene more often. You must check Editor > Enable Commands in order to intervene and use the procedure interactively. If you do not, the procedure will run to completion without pausing.

At the pause, Minitab displays a MORE? prompt. At this prompt, you can continue the display of steps, terminate the procedure, or intervene by typing a subcommand.

To...	Type
display another "page" of steps (or until no more predictors can enter or leave the model)	YES
terminate the procedure	NO
enter a set of variables	ENTER C... C

remove a set of variables	REMOVE C... C
force a set of variables to be in model	FORCE C... C
display the next best alternate predictors	BEST K
set the number of steps between pauses	STEPS K
change F to enter	FENTER K
change F to remove	FREMOVE K
change α to enter	AENTER K
change α to remove	AREMOVE K

Example of stepwise regression

Students in an introductory statistics course participated in a simple experiment. Each student recorded his or her height, weight, gender, smoking preference, usual activity level, and resting pulse. They all flipped coins, and those whose coins came up heads ran in place for one minute. Afterward, the entire class recorded their pulses once more. You wish to find the best predictors for the second pulse rate.

1 Open the worksheet PULSE.MTW.

2 Press [CTRL] + [M] to make the Session window active.

3 Choose **Editor > Enable Commands** so Minitab displays session commands.

4 Choose **Stat > Regression > Stepwise**.

5 In **Response**, enter *Pulse2*.

6 In **Predictors**, enter *Pulse1 Ran–Weight*.

7 Click **Options**.

8 **In Number of steps between pauses**, type *2*. Click **OK** in each dialog box.

9 In the Session window, at the first **More**? prompt, type *Yes*.

10 In the Session window, at the first **More**? prompt, type *No*.

Session window output

Stepwise Regression: Pulse2 versus Pulse1, Ran, ...

```
  Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15


Response is Pulse2 on 6 predictors, with N = 92


Step             1      2
Constant     10.28  44.48

Pulse1       0.957  0.912
T-Value       7.42   9.74
P-Value      0.000  0.000

Ran                 -19.1
T-Value             -9.05
P-Value             0.000

S             13.5   9.82
R-Sq         37.97  67.71
R-Sq(adj)    37.28  66.98
Mallows C-p  103.2   13.5

More? (Yes, No, Subcommand, or Help)

SUBC> yes

Step              3
Constant      42.62
```

```
Pulse1        0.812
T-Value        8.88
P-Value       0.000

Ran           -20.1
T-Value      -10.09
P-Value       0.000

Sex             7.8
T-Value        3.74
P-Value       0.000

S              9.18
R-Sq          72.14
R-Sq(adj)     71.19
Mallows C-p     1.9

More? (Yes, No, Subcommand, or Help)

SUBC> no
```

Interpreting the results

This example uses six predictors. You requested that Minitab do two steps of the automatic stepwise procedure, display the results, and allow you to intervene.

The first "page" of output gives results for the first two steps. In step 1, the variable Pulse1 entered the model; in step 2, the variable Ran entered. No variables were removed on either of the first two steps. For each model, MINITAB displays the constant term, the coefficient and its t-value for each variable in the model, S (square root of MSE), and R^2.

Because you answered YES at the MORE? prompt, the automatic procedure continued for one more step, adding the variable Sex. At this point, no more variables could enter or leave, so the automatic procedure stopped and again allowed you to intervene. Because you do not want to intervene, you typed NO.

The stepwise output is designed to present a concise summary of a number of fitted models. If you want more information on any of the models, you can use the regression procedure.

Best Subsets

Best Subsets Regression

Stat > Regression > Best Subsets

Best subsets regression identifies the best-fitting regression models that can be constructed with the predictor variables you specify. Best subsets regression is an efficient way to identify models that achieve your goals with as few predictors as possible. Subset models may actually estimate the regression coefficients and predict future responses with smaller variance than the full model using all predictors [19].

Minitab examines all possible subsets of the predictors, beginning with all models containing one predictor, and then all models containing two predictors, and so on. By default, Minitab displays the two best models for each number of predictors.

For example, suppose you conduct a best subsets regression with three predictors. Minitab will report the best and second best one-predictor models, followed by the best and second best two-predictor models, followed by the full model containing all three predictors.

Dialog box items

Response: Enter the column containing the response (Y) variable.

Free predictors: Enter the columns containing the candidate predictor (X) variables. You can specify up to 31 variables, though large models require long computation time. See Large Data Sets for more information.

Predictors in all models: Select columns containing variables you want included as predictors in every model. Columns entered here must not be listed in **Free predictors**. If you are analyzing a large data set with more than 15 predictors, consider including certain predictors here in order to decrease the number of free variables and speed up the computations. The maximum number of variables which can be entered is equal to 100 minus the number of variables entered in Free predictors.

<Options>

Data – Best Subsets

Enter response and predictor variables in the worksheet in numeric columns of equal length so that each row in your worksheet contains measurements on one unit or subject. Minitab automatically omits rows with missing values from all models.

You can use as many as 31 free predictors. However, the analysis can take a long time when 15 or more free predictors are used. When analyzing a very large data set, forcing certain predictors to be in the model by entering them in Predictors in all models can decrease the length of time required to run the analysis. The total number of predictors (forced and free) in the analysis can not be more than 100.

To do a best subsets regression

1 Choose **Stat > Regression > Best Subsets**.

2 In **Response**, enter the numeric column containing the response (Y) data.

3 In **Free predictors**, enter from 1 to 31 numeric columns containing the candidate predictor (X) variables.

4 If you like, use one or more of the dialog box options, then click **OK**.

Best Subsets Regression – Options

Stat > Regression > Best Subsets > Options

Enter the minimum and maximum number of free predictors, how many "best" models to print, and whether or not to fit the intercept term.

Dialog box items

Free Predictor(s) in Each Model: By default, Best Subsets prints the best 1-predictor models, the best 2-predictor models, on up to the best m-predictor models. If you give a range, say from 5 to 12, then only the best 5-, 6-,..., 12-predictor models are printed. Note that this does not include the predictors you specified in **Predictors in all models** in the Best Subsets dialog box. For example, if you specified two variables for inclusion in all models, and then set the minimum and maximum to 5 and 12, respectively, models with 7 to 14 predictors will be printed.

Minimum: Specify the minimum number of free predictors to be added to the model.

Maximum: Specify the maximum number of free predictors to be added to the model.

Models of each size to print: Specify a number from 1 to 5. For example if you choose 3, Minitab will print information from the "best" 3 models of each size (if there are that many).

Fit intercept: Uncheck to exclude the intercept term from the regression models.

Example of best subsets regression

Total heat flux is measured as part of a solar thermal energy test. You wish to see how total heat flux is predicted by other variables: insulation, the position of the focal points in the east, south, and north directions, and the time of day. Data are from Montgomery and Peck [27], page 486.

1 Open the worksheet EXH_REGR.MTW.

2 Choose **Stat > Regression > Best Subsets**.

3 In **Response**, enter **Heatflux**.

4 In **Free Predictors**, enter **Insolation–Time**. Click **OK**.

Session window output

Best Subsets Regression: HeatFlux versus Insolation, East, ...

```
Response is HeatFlux

                                        I
                                        n
                                        s
                                        o
                                        l
                                        a   S N
                                        t E o o T
                                        i a u r i
                       Mallows          o s t t m
Vars  R-Sq  R-Sq(adj)      C-p       S  n t h h e
   1  72.1       71.0     38.5  12.328        X
   1  39.4       37.1    112.7  18.154  X
   2  85.9       84.8      9.1  8.9321      X X
   2  82.0       80.6     17.8  10.076        X X
   3  87.4       85.9      7.6  8.5978    X X X
   3  86.5       84.9      9.7  8.9110  X   X X
   4  89.1       87.3      5.8  8.1698  X X X X
   4  88.0       86.0      8.2  8.5550  X   X X X
   5  89.9       87.7      6.0  8.0390  X X X X X
```

Interpreting the results

Each line of the output represents a different model. Vars is the number of variables or predictors in the model. R^2 and adjusted R^2 are converted to percentages. Predictors that are present in the model are indicated by an X.

In this example, it isn't clear which model fits the data best. The model with all the variables has the highest adjusted R^2 (87.7%), a low Mallows Cp value (6.0), and the lowest S value (8.0390). The four-predictor model with all variables except Time has a lower Cp value (5.8), although S is slightly higher (8.16) and adjusted R^2 is slightly lower (87.3%). The best three-predictor model includes North, South, and East, with a slightly higher Cp value (7.6) and a lower adjusted R^2 (85.9%). The best two-predictor model might be considered the minimum fit. The multiple regression example indicates that adding the variable East does not improve the fit of the model.

Before choosing a model, you should always check to see if the models violate any regression assumptions using residual plots and other diagnostic tests. See Checking your model.

Fitted Line Plot

Fitted Line Plot

Stat > Regression > Fitted Line Plot

This procedure performs regression with linear and polynomial (second or third order) terms, if requested, of a single predictor variable and plots a regression line through the data, on the actual or log10 scale. Polynomial regression is one method for modeling curvature in the relationship between a response variable (Y) and a predictor variable (X) by extending the simple linear regression model to include X^2 and X^3 as predictors.

Dialog box items

Response [Y]: Select the column containing the Y, or response variable.

Predictor [X]: Select the column containing the X, or predictor variable.

Type of Regression Model

Linear: Choose to fit a linear regression model.

Quadratic: Choose to fit a quadratic model.

Cubic: Choose to fit a cubic model.

<Graphs>

<Options>

<Storage>

Data – Fitted Line Plot

Enter your response and single predictor variables in the worksheet in numeric columns of equal length so that each row in your worksheet contains measurements on one unit or subject. Minitab automatically omits rows with missing values from the calculations.

To do a fitted line plot

1 Choose **Stat > Regression > Fitted Line Plot**.

2 In **Response (Y)**, enter the numeric column containing the response data.

3 In **Predictor (X)**, enter the numeric column containing the predictor variable.

4 If you like, use one or more of the dialog box options, then click **OK**.

Fitted Line Plot – Options

Stat > Regression > Fitted Line Plot > Options

Allows you to transform the X and Y variables, display confidence and prediction bands around the fitted regression line, change the confidence level, and add a custom title.

Dialog box items

Transformations

Logten of Y: Check to use log10Y as the response variable.

Display logscale for Y variable: Check to plot the transformed Y variable on a log scale.

Logten of X: Check to use log10X as the predictor variable. If you use this option with polynomials of order greater than one, then the polynomial regression will be based on powers of the log10X.

Display logscale for X variable: Check to plot the transformed X variable on a log scale.

Display Options

Display confidence intervals: Check to display confidence bands about the fitted regression line at the confidence level specified in the box below.

Display prediction intervals: Check to display prediction bands about the fitted regression line at the confidence level specified in the box below.

Confidence level: Enter a number between 0 and 100 for the level of confidence you want to use for the confidence and prediction bands. The default is 95.0.

Title: To replace the default title with your own custom title, type the desired text in this box.

Fitted Line Plot – Storage

Stat > Regression > Fitted Line Plot > Storage

You can store the residuals, fits, and coefficients from the regression analysis. These values are stored in the next available columns, and are assigned column names.

Dialog box items

Residuals: Check to store the residuals. The residuals will be in log scale if you transformed the Y variable in the <Options> subdialog box.

Standardized residuals: Check to store the standardized residuals. The standardized residuals will be in log scale if you transformed the Y variable in the <Options> subdialog box.

Deleted t residuals: Check to store the Studentized residuals. The deleted t residuals will be in log scale if you transformed the Y variable in the <Options> subdialog box.

Fits: Check to store the fitted values. The fitted values will be in log scale if you transformed the Y variable in the <Options> subdialog box.

Coefficients: Check to store the coefficients of the regression equation.

Residuals in original units: Check to store the residuals for the transformed response in the original scale of the data. This is only available if you transform the Y variable in the <Options> subdialog box.

Fits in original units: Check to store the fitted values for the transformed response in the original scale of the data. This is only available if you transform the Y variable in the <Options> subdialog box.

Example of Fitted Regression Line

You are studying the relationship between a particular machine setting and the amount of energy consumed. This relationship is known to have considerable curvature, and you believe that a log transformation of the response variable will produce a more symmetric error distribution. You choose to model the relationship between the machine setting and the amount of energy consumed with a quadratic model.

1 Open the worksheet EXH_REGR.MTW.

2 Choose **Stat > Regression > Fitted Line Plot**.

3 In **Response (Y)**,enter **EnergyConsumption**.

4 In **Predictor (X)**, enter **MachineSetting**.

5 Under **Type of Regression Model**, choose **Quadratic**.

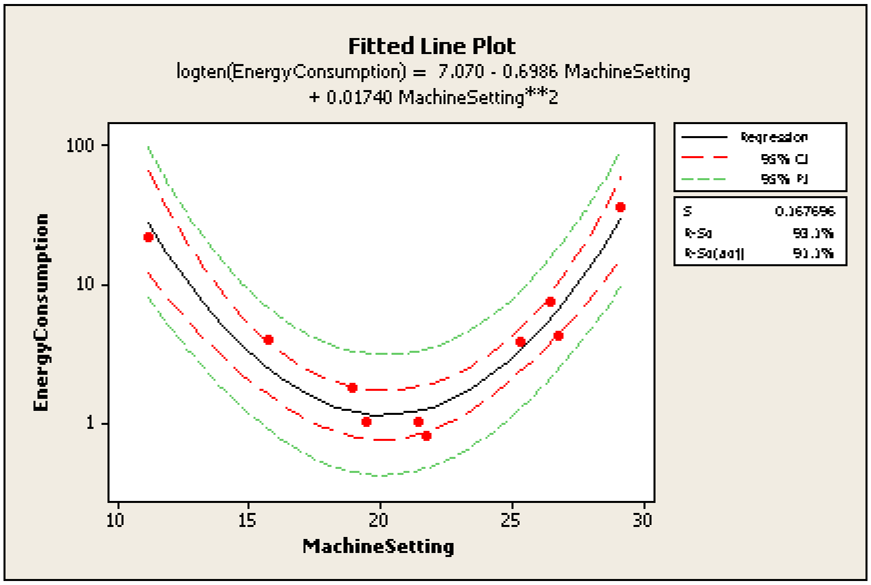
6 Click **Options**. Under **Transformations**, check **Logten of Y** and **Display logscale for Y variable.** Under **Display Options**, check **Display confidence interval** and **Display prediction interval**. Click **OK** in each dialog box.

Session window output

Polynomial Regression Analysis: EnergyConsumption versus MachineSetting

```
The regression equation is
logten(EnergyConsumption) = 7.070 - 0.6986 MachineSetting
                            + 0.01740 MachineSetting**2


S = 0.167696   R-Sq = 93.1%   R-Sq(adj) = 91.1%


Analysis of Variance

Source      DF       SS       MS      F      P
Regression   2  2.65326  1.32663  47.17  0.000
Error        7  0.19685  0.02812
Total        9  2.85012


Sequential Analysis of Variance

Source     DF       SS      F      P
Linear      1  0.03688   0.10  0.754
Quadratic   1  2.61638  93.04  0.000


Fitted Line: EnergyConsumption versus MachineSetting
```

Graph window output



Interpreting the results

The quadratic model (p-value = 0.000, or actually p-value < 0.0005) appears to provide a good fit to the data. The R^2 indicates that machine setting accounts for 93.1% of the variation in log10 of the energy consumed. A visual inspection of the plot reveals that the data are randomly spread about the regression line, implying no systematic lack-of-fit. The red-dashed lines are the 95% confidence limits for the log10 of energy consumed. The green-dashed lines are the 95% prediction limits for new observations.

Partial Least Squares (PLS)

Partial Least Squares Overview

Partial least squares (PLS) is a biased, non-least squares regression procedure that relates a set of predictor variables to multiple response variables. Use PLS with ill-conditioned data, when predictors are highly collinear or predictors outnumber observations and ordinary least squares regression either fails or produces coefficients with high standard errors.

Minitab uses the nonlinear iterative partial least squares (NIPALS) algorithm developed by Herman Wold [36]. The algorithm reduces the number of predictors using a technique similar to principal components analysis to extract a set of components that describes maximum correlation among the predictors and response variables. It then performs least squares regression on these uncorrelated components. In addition, cross-validation is often used to select the components that maximize the model's predictive ability. To perform partial least squares regression, see Stat > Regression > Partial Least Squares. See [14] for a tutorial on PLS.

Partial Least Squares

Stat > Regression > Partial Least Squares

Use partial least squares (PLS) to perform biased, non-least squares regression with one or more responses. PLS is particularly useful when your predictors are highly collinear or you have more predictors than observations and ordinary least squares regression either fails or produces coefficients with high standard errors. PLS reduces the number of predictors to a set of uncorrelated components and performs least squares regression on these components.

PLS fits multiple response variables in a single model. Because PLS models the responses in a multivariate way, the results may differ significantly from those calculated for the responses individually. Model multiple responses together only if they are correlated. For more information, see [14] and [20].

Dialog box items

Responses: Enter one or more columns containing the responses (Y).

Predictors: Enter one or more columns containing the predictors (X).

Maximum number of components: Type the number of components to calculate or cross-validate. By default, Minitab calculates or cross-validates 10 components or the number of predictors, whichever is less. You should not enter more components than there are predictors.

<Validation>

<Prediction>

<Graphs>

<Results>

<Storage>

Data – Partial Least Squares

Enter response and predictor variables in numeric columns of equal length, so that each row in your worksheet contains measurements on one observation or subject.

Minitab omits all observations that are missing responses or predictors.

To do partial least squares regression

1 Choose **Stat > Regression > Partial Least Squares**.

2 In **Responses**, enter one or more columns containing the responses (Y).

3 In **Predictors**, enter one or more columns containing the predictors (X).

4 In **Maximum number of components,** type a number or leave blank. By default, Minitab extracts or cross-validates 10 components.

5 If you like, use any dialog box options, then click **OK**.

Partial Least Squares – Validation

Stat > Regression > Partial Least Squares > Validation

You can use cross-validation to select the components that maximize your model's predictive ability.

Dialog box items

Cross-Validation

None: Choose to suppress cross-validation.

Leave-one-out: Choose to perform cross-validation leaving out one observation each time the model is recalculated.

Leave-group-out of size: Choose to perform cross-validation leaving out multiple observations (groups) each time the model is recalculated. Then, enter group size, which is 2 by default.

Leave out as specified in column: Choose to perform cross-validation using group identifiers (positive integers) to specify which observations are left out together each time the model is recalculated. Then, enter group identifier column, which must be equal in length to predictor and response columns and contain no missing values.

Using Cross-Validation

Cross-validation calculates the predictive ability of potential models to help you determine the appropriate number of components to retain in your model. Cross-validation is recommended if you do not know the optimal number of components. When the data contain multiple response variables, Minitab validates the components for all responses simultaneously. For more information, see [15].

Listed below are the methods for cross-validation:

- Leave-one-out: Calculates potential models leaving out one observation at a time. For large data sets, this method can be time-consuming, because it recalculates the models as many times as there are observations.
- Leave-group-out of size: Calculates the models leaving multiple observations out at a time, reducing the number of times it has to recalculate a model. This method is most appropriate when you have a large data set.
- Leave-out-as-specified-in-column: Calculates the models, simultaneously leaving out the observations that have matching numbers in the group identifier column, which you create in the worksheet. This method allows you to specify which observations are omitted together. For example, if the group identifier column includes numbers 1, 2, and 3, all observations with 1 are omitted together and the model is recalculated. Next, all observations with 2 are omitted and the model is recalculated, and so on. In this case, the model is recalculated a total of 3 times. The group identifier column must be the same length as your response and predictor columns and cannot contain missing values.

To cross-validate a PLS model

1 Choose **Stat > Regression > Partial Least Squares > Validation**.

2 Under Cross-Validation, do one of the following:

- To suppress cross-validation, choose **None**.
- To leave out one observation each time the model is recalculated, choose **Leave-one-out**.
- To leave out multiple observations each time the model is recalculated, choose **Leave-group-out of size**, then enter the number of observations to be left out together. By default, Minitab uses the group size of 2, which cuts in half the number of times the model is recalculated.
- To leave out specified observations together each time the model is recalculated, choose **Leave out as specified in column**, then enter the column containing the group identifiers.

3 Click **OK**.

Partial Least Squares – Prediction

Stat > Regression > Partial Least Squares > Prediction

You can calculate and store predicted response values using your PLS model for two purposes: testing prediction quality and predicting new responses.

- Testing prediction quality: You can apply your PLS model to a new data set that includes responses for each observation. Minitab calculates new response values and compares them to the actual response for each observation, calculating a test R^2, which indicates the model's predictive ability.
- Predicting new responses: You use the PLS model to calculate new response values for a set of predictors for which you have no response data. Without response data, a test R^2 cannot be calculated.

Dialog box items

New Observations (columns and/or constants)

Predictors: Type the numeric predictor values, or enter the columns or constants in which they are stored. The number of predictors must equal the number of predictors in the model. Columns must be of equal length.

Responses (Optional): Enter the numeric columns containing the response values. You cannot type the response values; they must be stored in columns. The number of response columns must equal the number of responses in the model and be the same length as the predictors containing new observations.

Confidence level: Type the desired confidence level (for example, type 90 for 90%). The default is 95%.

Storage

Fits: Check to store fitted values for new observations.

SE of fits: Check to store estimated standard errors of the fitted values.

Confidence limits: Check to store lower and upper limits of the confidence interval.

Prediction limits: Check to store lower and upper limits of the prediction interval.

To predict responses in PLS

1 Choose **Stat > Regression > Partial Least Squares > Prediction**.

2 In **Predictors**, do any combination of the following:

- Type numeric predictor values.
- Enter stored constants containing numeric predictor values.
- Enter columns of equal length containing numeric predictor values.

The number of predictors and the order in which they are entered must match your original model.

3 In **Responses (optional)**, enter columns of equal length containing the response values.

- Columns must be equal in length to columns entered in **Predictors** above.
- The number of responses must equal the number of responses in the PLS model.

4 In **Confidence level**, type a value or use the default, which is 95%.

5 Under **Storage**, check any of the prediction results to store them in your worksheet. Click **OK**.

Partial Least Squares – Graphs

Stat > Regression > Partial Least Squares > Graphs

You can display plots that examine the PLS model, the standardized residuals, and the components. See Understanding PLS graphs.

Dialog box items

Model Evaluation

Model selection plot: Check to plot the R^2 values for the fitted and cross-validated models.

Response plot: Check to plot the fitted and cross-validated responses versus the actual responses.

Coefficient plot: Check to plot the unstandardized regression coefficients for each predictor.

Std coefficient plot: Check to plot the standardized regression coefficients for each predictor

Distance plot: Check to plot each observation's distance from the x-model and distance from the y-model.

Residual Analysis

Residual histogram: Check to display a histogram of the standardized residuals.

Residual normal plot: Check to display a normal probability plot of the standardized residuals.

Residual versus fit: Check to display standardized residuals versus the fitted values.

Residual versus leverage: Check to display standardized residuals versus the leverages.

Residual fourpack: Choose to display a layout of a histogram of standardized residuals, a normal plot of standardized residuals, a plot of standardized residuals versus fits, and a plot of standardized residuals versus leverages.

Component Evaluation

Score plot: Check to plot the x-scores of the first component versus the second component.

3D score plot: Check to plot the x-scores of the first component, second and third components.

Loading plot: Check to plot the x-loadings of the first component versus the second component.

Residual X plot: Check to display a line plot of the residual x-values.

Calculated X plot: Check to display a line plot of the fitted x-values.

Understanding PLS Graphs

Use the table below to learn more about PLS graphs. To create any of these graphs, see Partial Least Squares – Graphs.

PLS graph	Definition	Use to...
Model selection plot	Scatterplot of the R^2 and predicted R^2 values as a function of the number of components. The vertical line indicates number of components in the optimal model.	Compare the modeling and predictive powers of models with different numbers of components.
Response plot	Scatterplot of the fitted and cross-validated responses versus the actual responses.	Show how well the model fits and predicts. Large differences in fitted and cross-validated values indicate leverage points.
Coefficient plot	Projected scatterplot of the unstandardized regression coefficients.	View sign and the magnitude of the relationship between predictors and responses.
Standardized coefficient plot	Projected scatterplot of the standardized regression coefficients.	View sign and the magnitude of the relationship between predictors and responses when predictors are not on the same scale.
Distance plot	Scatterplot of each observation's distance from the x-model and distance from y-model.	Identify leverage points and outliers.
Residual histogram	Histogram of the standardized residuals.	Check the normality of your residuals. Histograms should show a bell-shaped distribution.
Residual normal plot	Scatterplot of the standardized residuals versus the normal scores.	Check the normality of your residuals. Points should follow a straight line.
Residual versus fit plot	Scatterplot of the standardized residuals versus the fitted responses.	Identify outliers and check for patterns in the residuals.
Residual versus leverage plot	Scatterplot of the standardized residuals versus leverages.	Identify outliers and leverage points simultaneously.

Residual fourpack	Layout of residual histogram, residual normal plot, residual versus fit plot, and residual versus order plot on one page.	View residual plots simultaneously.
Score plot	Scatterplot of the x-scores from the first and second components.	Display the overall configuration of the data using the first two components to identify leverage points or clusters of points.
3D score plot	3D scatterplot of the x-scores from the first, second, and third components.	Display the overall configuration of the data using the first three components to identify leverage points or clusters of points.
Loading plot	Connected scatterplot of the x-loadings from the first and second components.	Display the correlation between the loadings of each predictor on the first and second components. Compare the importance of predictors to the model.
Residual X plot	Connected scatterplot of the x-residuals, in which each line represents an observation and has as many points as predictors.	Identify observations or predictors that are poorly explained by the model.
Calculated X plot	Connected scatterplot of the x-calculated values, in which each line represents an observation and has as many points as predictors.	Identify observations or predictors that are poorly explained by the model.

Partial Least Squares – Results

Stat > Regression > Partial Least Squares > Results

You can control the output displayed in the Session window. By default, Minitab displays the number of components in the model and the Analysis of variance table.

Dialog box items

Model selection: Check to display the fitted and cross-validated sums of squares, the x-variance value, and the R^2 and predicted R^2 values for each model calculated. If you do not choose to cross-validate the model, the cross-validated sums of squares and the predicted R^2 do not appear on the output.

Coefficients: Check to display standardized and unstandardized regression coefficients for the predictors.

Fits and residuals: Check to display the calculated fitted values, residuals, and standardized residuals and the cross-validated fitted values and cross-validated residuals. If you do not choose to cross-validate the model, the cross-validated fitted values and residuals do not appear on the output.

Leverages and distance: Check to display leverages, distances from x-model and distances from y-model.

X scores: Check to display the x-scores for each observation.

X loadings: Check to display the loadings for each predictor.

X residuals: Check to display the residuals for the x-matrix.

X calculated: Check to display the calculated values for x-matrix.

Y scores: Check to display the y-scores for each observation.

Y loadings: Check to display the loadings for each response.

Y residuals: Check to display the residuals for the y-matrix. The y-residuals are the same as the residuals for the model.

Y calculated: Check to display the calculated values for the y-matrix. The y-calculated values are the same as the fitted values for the model.

Partial Least Squares – Storage

Stat > Regression > Partial Least Squares > Storage

You can store information from the selected model in the worksheet.

Dialog box items

Coefficients: Check to store the coefficients. Minitab stores a column for each response.

Standardized coefficients: Check to store the standardized coefficients. Minitab stores a column for each response.

Fits: Check to store the fitted values. Minitab stores a column for each response.

Cross-validated fits: Check to store the cross-validated fitted values. Minitab stores a column for each response.

Residuals: Check to store the residuals. Minitab stores a column for each response.

Cross-validated residuals: Check to store the cross-validated residuals. Minitab stores a column for each response.

Standardized residuals: Check to store the standardized residuals. Minitab stores a column for each response.

Leverages: Check to store leverages.

Distances from X: Check to store the distances from the x-model.

Distances from Y: Check to store the distances from the y-model.

X scores: Check to store the x-scores for the predictors in a matrix.

X loadings: Check to store the loadings for predictors in a matrix.

X weights: Check to store the weights for predictors in a matrix.

X residuals: Check to store the x-residuals for the predictors. Minitab stores a column for each predictor.

X calculated: Check to store the fitted values for the predictors. Minitab stores a column for each predictor.

Y scores: Check to store the y-scores for the responses in a matrix.

Y loadings: Check to store the loadings for the responses in a matrix.

Y residuals: Check to store the y-residuals for the responses. Minitab stores a column for each response. The y-residuals are the same as the residuals for the model.

Y calculated: Check to store the y-calculated values for the responses. Minitab stores a column for each response. The y-calculated values are the same as the fitted values for the model.

Example of partial least squares regression

You are a wine producer who wants to know how the chemical composition of your wine relates to sensory evaluations. You have 37 Pinot Noir wine samples, each described by 17 elemental concentrations (Cd, Mo, Mn, Ni, Cu, Al, Ba, Cr, Sr, Pb, B, Mg, Si, Na, Ca, P, K) and a score on the wine's aroma from a panel of judges. You want to predict the aroma score from the 17 elements and determine that PLS is an appropriate technique because the ratio of samples to predictors is low. Data are from [12].

1 Open the worksheet WINEAROMA.MTW.

2 Choose **Stat > Regression > Partial Least Squares**.

3 In **Responses**, enter *Aroma*.

4 In **Predictors**, enter *Cd-K*.

5 In **Maximum number of components**, type *17*.

6 Click **Validation**, then choose **Leave-one-out**. Click **OK**.

7 Click **Graphs**, then check **Model selection plot**, **Response plot**, **Std Coefficient plot**, **Distance plot**, **Residual versus leverage plot**, and **Loading plot**. Uncheck **Coefficient plot**. Click **OK** in each dialog box.

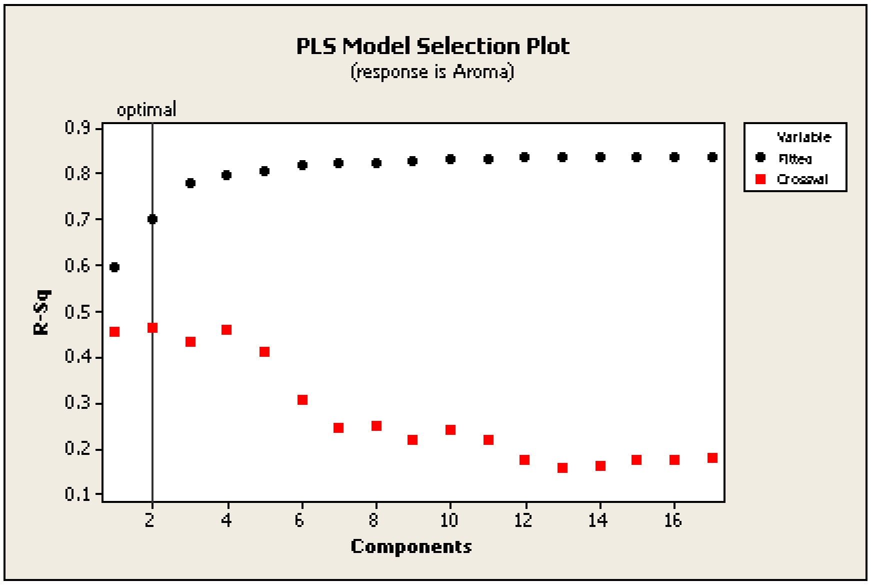
Session window output

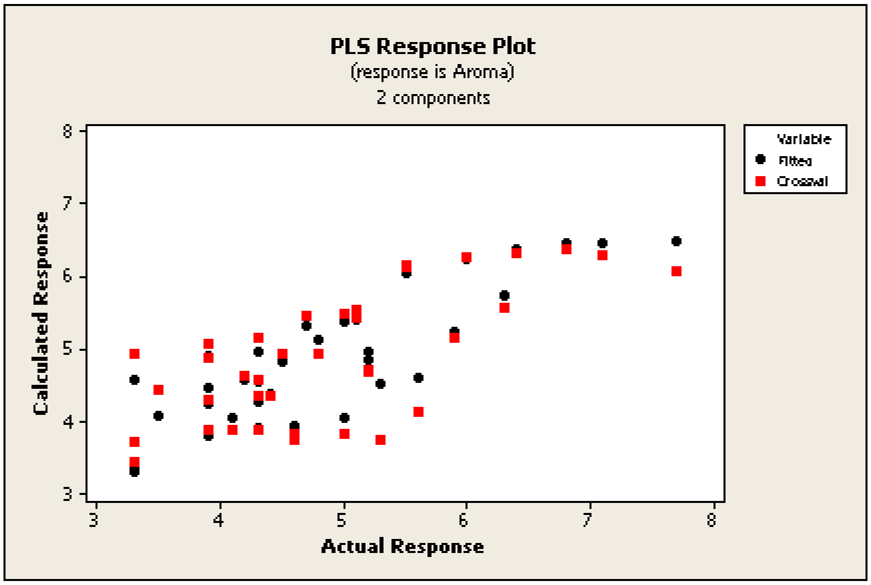
PLS Regression: Aroma versus Cd, Mo, Mn, Ni, Cu, Al, Ba, Cr, ...

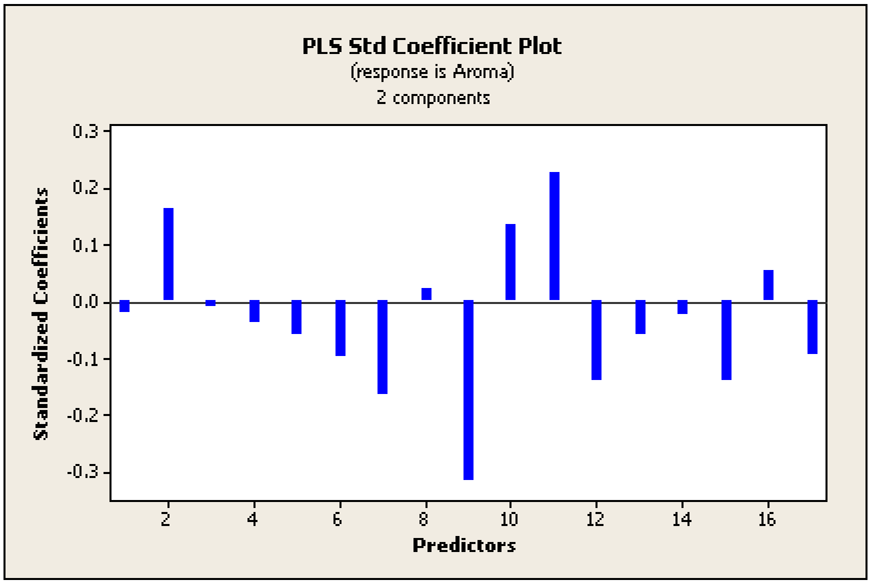
```
Number of components selected by cross-validation: 2
Number of observations left out per group: 1
Number of components cross-validated: 17


Analysis of Variance for Aroma

Source          DF       SS       MS      F      P
Regression       2  28.8989  14.4494  39.93  0.000
Residual Error  34  12.3044   0.3619
Total           36  41.2032


Model Selection and Validation for Aroma

Components  X Variance  Error SS      R-Sq    PRESS  R-Sq (pred)
         1    0.225149   16.5403  0.598569  22.3904     0.456585
         2    0.366697   12.3044  0.701374  22.1163     0.463238
         3                8.9938  0.781720  23.3055     0.434377
         4                8.2761  0.799139  22.2610     0.459726
         5                7.8763  0.808843  24.1976     0.412726
         6                7.4542  0.819087  28.5973     0.305945
         7                7.2448  0.824168  31.0924     0.245389
         8                7.1581  0.826274  30.9149     0.249699
```

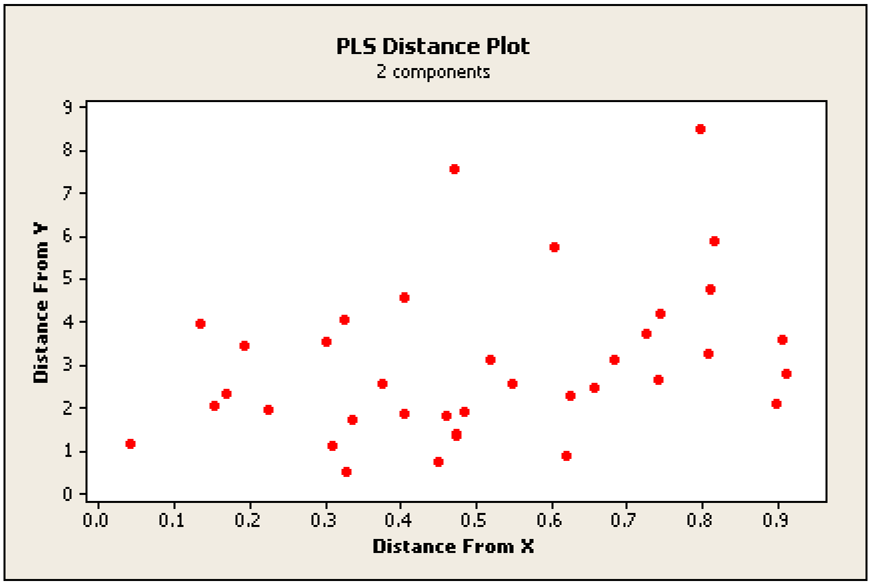
```
   9                 6.9711  0.830811  32.1611     0.219451
  10                 6.8324  0.834178  31.3590     0.238920
  11                 6.7488  0.836207  32.1908     0.218732
  12                 6.6955  0.837501  34.0891     0.172660
  13                 6.6612  0.838333  34.7985     0.155442
  14                 6.6435  0.838764  34.5011     0.162660
  15                 6.6335  0.839005  34.0829     0.172811
  16                 6.6296  0.839100  34.0143     0.174476
  17                 6.6289  0.839117  33.8365     0.178789
```

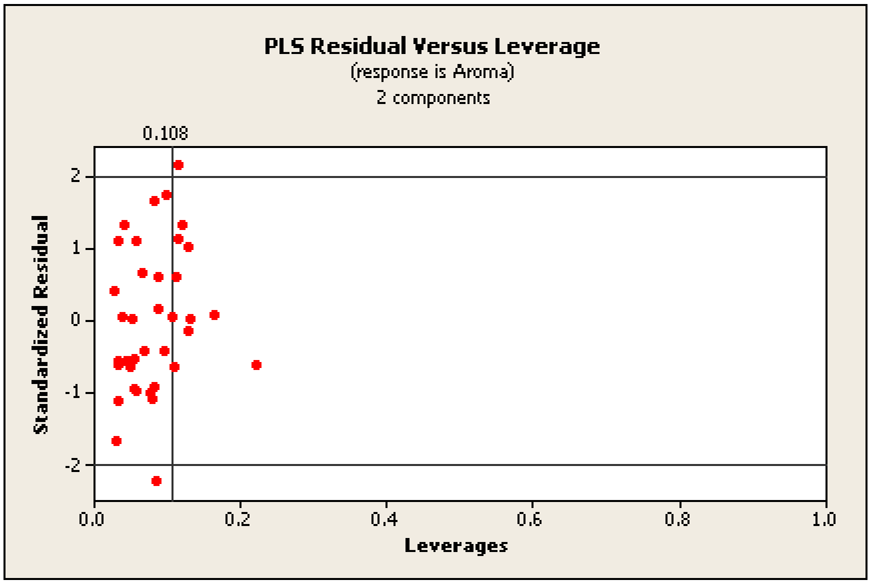
Graph window output

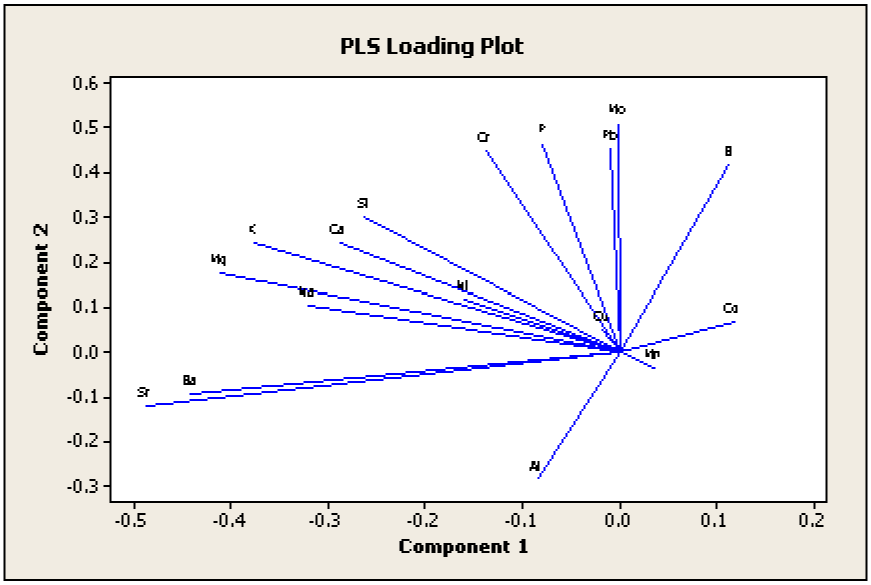





PLS Std Coefficient Plot
(response is Aroma)
2 components
Standardized Coefficients
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
2
4
6
8
10
12
14
16
Predictors


PLS Distance Plot
2 components
Distance From Y
9
8
7
6
5
4
3
2
1
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Distance From X





Interpreting the results

Session window output

- The first line of the output shows the number of components in optimal model, which is defined as the model with the highest predicted R^2. Minitab selected the two-component model as the optimal model, with a predicted R^2 of .46.
- Because you fit the same number of components as there are predictors (17), you can compare the goodness-of-fit and goodness-of-prediction statistics for the PLS models and the least squares solution.
- Minitab displays one Analysis of Variance table per response based on the optimal model. The p-value for aroma is 0.000, which is less than 0.05, providing sufficient evidence that the model is significant.
- Use the Model Selection and Validation table to select the optimal number of components for your model. Depending on your data or field of study, you may determine that a model other than the one selected by cross-validation is more appropriate.
  - The model with two components, which was selected by cross-validation, has an R^2 of 70.1% and a predicted R^2 of 46.3%. You can also see that the four-component model has a higher R^2 (79.9%), and a slightly lower predicted R^2 (46%). Because the predicted R^2 does not decrease measurably with two additional components, the four-component model is not overfit and could be used instead of the two-component model. (The remainder of this example is based on the two-component model.)

- By comparing the predicted R^2 of the two-component PLS model to the predicted R^2 of the 17-component least squares model, you can see that the PLS model predicts data much more accurately than the least squares model does. The predicted R^2 of the two-component model is 46%, while the predicted R^2 of the 17-component model is only 18%.

- The X-variance indicates the amount of variance in the predictors that is explained by the model. In this example, the 2-component model explains 36.7% of the variance in the predictors.

Graph window output

- The model selection plot is a graphical display of the Model Selection and Validation table. The vertical line indicates that the optimal model has two components. You can see that the predictive ability of all models with more than four components decreases significantly, including the 17-component least squares solution, which has a predicted R^2 of only 18%.
- Because the points are in a linear pattern, from the bottom left-hand corner to the top right-hand corner, the response plot indicates that the model fits the data adequately. Although there are differences between the fitted and cross-validated fitted responses, none are severe enough to indicate an extreme leverage point.
- The coefficient plot displays the standardized coefficients for the predictors. You can use this plot to interpret the magnitude and sign of the coefficients. The elements Sr, B, Mo, Ba, Mg, Pb, and Ca have the largest standardized coefficients and the biggest impact on aroma. The elements Mo, Cr, Pb, and B are positively related to aroma, while Cd, Ni, Cu, Al, BA, and Sr are negatively related.
- The loading plot compares the relative influence of the predictors on the response. In this example, Cu and Mn have very short lines, indicating that they have low x-loadings and are not related to aroma. The elements Sr, Mg, and Ba have long lines, indicating that they have higher loadings and are more related to aroma.
- The distance plot and the residual versus leverage plot display outliers and leverages. By brushing the distance plot, you can see that compared to the rest of the data:
  - observations 14 and 32 have a greater distance value on the y-axis
  - observations in rows 7, 12, and 23 have greater distance value on the x-axis

  The residual versus leverage plot confirms these findings, showing that:
  - observations 14 and 32 are outliers, because they are outside the horizontal reference lines
  - observations 7, 12, and 23 have extreme leverage values, because they are to the right of the vertical reference line

Binary Logistic Regression

Logistic Regression Overview

Both logistic regression and least squares regression investigate the relationship between a response variable and one or more predictors. A practical difference between them is that logistic regression techniques are used with categorical response variables, and linear regression techniques are used with continuous response variables.

Minitab provides three logistic regression procedures that you can use to assess the relationship between one or more predictor variables and a categorical response variable of the following types:

Variable type	Number of categories	Characteristics	Examples
Binary	2	two levels	success, failure yes, no
Ordinal	3 or more	natural ordering of the levels	none, mild, severe fine, medium, coarse
Nominal	3 or more	no natural ordering of the levels	blue, black, red, yellow sunny, rainy, cloudy

Both logistic and least squares regression methods estimate parameters in the model so that the fit of the model is optimized. Least squares minimizes the sum of squared errors to obtain parameter estimates, whereas logistic regression obtains maximum likelihood estimates of the parameters using an iterative-reweighted least squares algorithm [25].

Binary Logistic Regression

Stat > Regression > Binary Logistic Regression

Use binary logistic regression to perform logistic regression on a binary response variable. A binary variable only has two possible values, such as presence or absence of a particular disease. A model with one or more predictors is fit using an iterative reweighted least squares algorithm to obtain maximum likelihood estimates of the parameters [25].

Binary logistic regression has also been used to classify observations into one of two categories, and it may give fewer classification errors than discriminant analysis for some cases [10], [29].

Dialog box items

Response: Choose if the response data has been entered as raw data or as two columns–one containing the response values and one column containing the frequencies. Then enter the column containing the response values.

> **Frequency (optional):** If the data has been entered as two columns – one containing the response values and one column containing the frequencies – enter the column containing the frequencies in the text box.

Success: Choose if the response data has been entered as two columns – one containing the number of successes and one column containing the number of trials. Then enter the column containing the number of successes in the text box.

> **Trial:** Enter the column containing the number of trials.

Success: Choose if the response data has been entered as two columns – one containing the number of successes and one column containing the number of failures. Then enter the column containing the number of successes in the text box.

> **Failure:** Enter the column containing the number of failures.

Failure: Choose if the response data has been entered as two columns – one containing the number of failures and one column containing the number of trials. Then enter the column containing the number of failures in the text box..

> **Trial:** Enter the column containing the number of trials.

Model: Specify the terms to be included in the model.

Factors (optional): Specify which of the predictors are factors. Minitab assumes all variables in the model are covariates unless specified to be factors here. Continuous predictors must be modeled as covariates; categorical predictors must be modeled as factors.

<Graphs>

<Options>

<Results>

<Storage>

Data – Binary Logistic Regression

Your data must be arranged in your worksheet in one of five ways: as raw data, as frequency data, as successes and trials, as successes and failures, or as failures and trials. See Entering data for response variables.

Factors, covariates, and response data can be numeric, text, or date/time. The **reference level** and the **reference event** depend on the data type. See Setting reference levels and reference events for details.

The predictors may either be factors (nominal variables) or covariates (continuous variables). Factors may be crossed or nested. Covariates may be crossed with each other or with factors, or nested within factors.

The model can include up to 9 factors and 50 covariates. Unless you specify a predictor in the model as a factor, the predictor is assumed to be a covariate. Model continuous predictors as covariates and categorical predictors as factors. See Specifying the model terms.

Minitab automatically omits all observations with missing values from all calculations.

Entering Data for Response Variables

Data used for input to the logistic regression procedures may be arranged in two different ways in your worksheet: as raw (categorical) data, or as frequency (collapsed) data. For binary logistic regression, there are three additional ways to arrange the data in your worksheet: as successes and trials, as successes and failures, or as failures and trials. These ways are illustrated here for the same data.

The response entered as raw data or as frequency data :

Raw Data: one row for each observation

C1	C2	C3	C4
Response		**Factor**	**Covar**
0		1	12
1		1	12
1		1	12
.		.	.
.		.	.
.		.	.
1		1	12
0		2	12
1		2	12
.		.	.
.		.	.
.		.	.
1		2	12
.		.	.
.		.	.

Frequency Data: one row for each combination of factor and covariate

C1	C2	C3	C4
Response	**Count**	**Factor**	**Covar**
0	1	1	12
1	19	1	12
0	1	2	12
1	19	2	12
0	5	1	24
1	15	1	24
0	4	2	24
1	16	2	24
0	7	1	50
1	13	1	50
0	8	2	50
1	12	2	50
0	11	1	125
1	2	1	125
0	9	2	125
1	11	2	125
0	19	1	200
1	1	1	200
0	18	2	200
1	2	2	200

The binary response entered as the number of successes, failures, or trials. Enter one row for each combination of factor and covariate.

Successes and Trials

C1	C2	C3	C4
S	**T**	**Factor**	**Covar**
19	20	1	12
19	20	2	12
15	20	1	24
16	20	2	24
13	20	1	50
12	20	2	50
9	20	1	125
11	20	2	125
1	20	1	200
2	20	2	200

Successes and Failures

C1	C2	C3	C4
S	**F**	**Factor**	**Covar**
19	1	1	12
19	1	2	12
15	5	1	24
16	4	2	24
13	7	1	50
12	8	2	50
9	11	1	125
11	9	2	125
1	19	1	200
2	18	2	200

Failures and Trials

C1	C2	C3	C4
F	**T**	**Factor**	**Covar**
1	20	1	12
1	20	2	12
5	20	1	24
4	20	2	24
7	20	1	50
8	20	2	50
11	20	1	125
9	20	2	125
19	20	1	200
18	20	2	200

Specifying the Model in Logistic Regression

The logistic regression procedures can fit models with:

- up to 9 factors and up to 50 covariates
- crossed or nested factors
- covariates that are crossed with each other or with factors, or nested within factors

Model continuous predictors as covariates and categorical predictors as factors. Here are some examples. A is a factor and X is a covariate.

Model terms

A X A∗X	fits a full model with a covariate crossed with a factor
A | X	an alternative way to specify the previous model
A X X∗X	fits a model with a covariate crossed with itself making a squared term
A X(A)	fits a model with a covariate nested within a factor

The model for logistic regression is a generalization of the model used in Minitab's general linear model (GLM) procedure. Any model fit by GLM can also be fit by the logistic regression procedures. For a discussion of specifying models in general, see Specifying the Model Terms and Specifying Reduced Models. In the logistic regression commands, Minitab assumes any variable in the model is a covariate unless the variable is specified as a factor. In contrast, GLM assumes that any variable in the model is a factor unless the variable is specified as a covariate. Be sure to specify which predictors are factors in the main dialog box.

Model restrictions

Logistic regression models in Minitab have the same restrictions as GLM models:

- There must be enough data to estimate all the terms in your model, so that the model is full rank. Minitab will automatically determine if your model is **full rank** and display a message. In most cases, eliminating some unimportant high-order interactions in your model should solve your problem.
- The model must be hierarchical. In a hierarchical model, if an interaction term is included, all lower order interactions and main effects that comprise the interaction term must appear in the model.

To do a binary logistic regression

1 Choose **Stat > Regression > Binary Logistic Regression**.

2 Do one of the following:

   - If your data is in raw form, choose **Response** and enter the column containing the response variable.
   - If your data is in frequency form, choose **Response** and enter the column containing the response variable. In **Frequency**, enter the column containing the count or frequency variable.
   - If your data is in success-trial, success-failure, or failure-trial form, choose **Success** with **Trial**, **Success** with **Failure**, or **Failure** with **Trial**, and enter the respective columns in the accompanying boxes.

   See Entering data for response variables.

3 In **Model**, enter the model terms. See Specify the model in logistic regression.

4 If you like, use one or more of the dialog box options, then click **OK**.

Binary Logistic Regression – Graphs

Stat > Regression > Binary Logistic Regression > Graphs

Generates various diagnostic plots. You can use these plots to pinpoint factor/covariate patterns that are poorly fit by model or have a large influence on the estimated logistic regression coefficients. See Regression diagnostics and residual analysis for more information.

Involving Event Probability These graphs all use the estimated event probability as the x-axis variable.

Delta chi-square vs probability: Choose to plot delta Pearson chi-square versus the estimated event probability for each distinct factor/covariate pattern.

Delta deviance vs probability: Choose to plot delta deviance versus the estimated event probability for each distinct factor/covariate pattern.

Delta beta (Standardized) vs probability: Choose to plot the delta β (based on standardized Pearson residual) versus the estimated event probability for each distinct factor/covariate pattern.

Delta beta vs probability: Choose to plot delta β versus the estimated event probability for each distinct factor/covariate pattern.

Involving Leverage (Hi) These graphs all use the leverage as the x-axis variable.

Delta chi-square vs leverage: Choose to plot delta Pearson chi-square versus the leverage for each distinct factor/covariate pattern.

Delta deviance vs leverage: Choose to plot delta deviance versus the leverage for each distinct factor/covariate pattern.

Delta beta (Standardized) vs leverage: Choose to plot the delta β (based on standardized Pearson residual) versus the leverage for each distinct factor/covariate pattern.

Delta beta vs leverage: Choose to plot delta β versus the leverage for each distinct factor/covariate pattern.

Regression Diagnostics and Residual Analysis

Following any modeling procedure, it is a good idea to assess the validity of your model. Logistic regression has a collection of diagnostic plots, goodness-of-fits tests, and other diagnostic measures to do this. These residuals and diagnostic statistics allow you to identify factor/covariate patterns that are either poorly fit by the model, have a strong influence upon the estimated parameters, or which have a high leverage. Minitab provides different options for each of these, as listed in the following table. Hosmer and Lemeshow [21] suggest that you interpret these diagnostics jointly to understand any potential problems with the model.

To identify...	Use...	Which measures...
poorly fit factor/covariate patterns	Pearson residual	the difference between the actual and predicted observation
	standardized Pearson residual	the difference between the actual and predicted observation, but standardized to have σ = 1
	deviance residual	deviance residuals, a component of deviance chi-square
	delta chi-square	changes in the Pearson chi-square when the jth factor/covariate pattern is removed
	delta deviance	changes in the deviance when the jth factor/covariate pattern is removed
factor/covariate patterns with a strong influence on parameter estimates	delta beta	changes in the coefficients when the jth factor/covariate pattern is removed–based on Pearson residuals
	delta beta based on standardized Pearson residuals	changes in the coefficients when the jth factor/covariate pattern is removed–based on standardized Pearson residuals
factor/covariate patterns with a large leverage	leverage (Hi)	leverages of the jth factor/covariate pattern, a measure of how unusual predictor values are

The graphs available in the Graphs subdialog box allow you to visualize some of these diagnostics jointly; you can plot a measure useful for identifying poorly fit factor/covariate patterns (delta chi-square or delta deviance) or a measure useful for identifying a factor/covariate pattern with a strong influence on parameter estimates (one of the delta beta statistics) versus either the estimated event probability or leverage. The estimated event probability is the probability of the event, given the data and model. Leverages are used to assess how unusual the predictor values are (see Identifying outliers). See [16] for a further discussion of diagnostic plots. You can use Minitab's graph brushing capabilities to identify points. See Brushing Graphs.

Binary Logistic Regression – Options

Stat > Regression > Binary Logistic Regression > Options

Provides options that allow you to choose a link function, the reference level and event, specify starting values for the estimated coefficients, and setting the number of groups for the Hosmer and Lemeshow test.

Dialog box items

Link Functions Minitab provides three link functions, allowing you to fit a broad class of binary response models. The link function is the inverse of a distribution function.

Logit: Choose to use the logit link function. This is the default.

Normit/Probit: Choose to use the normit link function.

Gompit/Complementary log-log: Choose to use the gompit link function (also called the complementary log-log function). The gompit function is the inverse of the Gompertz distribution function.

Reference Options

Event: Enter the reference event of the response. For information on why you might want to change the reference event, see Interpreting parameter estimates.

Reference factor level (enter factor followed by level): Specify the reference factor level. Enter the factor column followed by the reference level. (Text and date/time levels must be enclose in quotes.) For information on why you might want to change the reference factor level, see Interpreting parameter estimates.

Algorithm Options

Starting estimates for algorithm: Specify the column containing the initial values for model parameters. The column containing the initial values must have one row for each estimated coefficient in the model. The value in the first row is assigned to the constant, the value in the second row is assigned to the first predictor defined in the model, the value in the third row is assigned to the second predictor, etc. There must be one row for each degree of freedom.

Estimates for validation model: Specify the column containing the estimated model parameters. Minitab will then fit the validation model.

Maximum number of iterations: Enter the maximum number of iterations you want Minitab to perform to reach convergence. The default value is 20. Minitab obtains maximum likelihood estimates through an iterative process. If the maximum number of iterations is reached before convergence, the command terminates.

Option for Hosmer-Lemeshow Test

Number of Groups: Enter the number of groups for the Hosmer-Lemeshow test. The default is 10. See [21] for details.

Interpreting Estimated Coefficients in Binary Logistic Regression

The interpretation of the estimated coefficients depends on: the link function, reference event, and reference factor levels (see Setting reference and event levels). The estimated coefficient associated with a predictor (factor or covariate) represents the change in the link function for each unit change in the predictor, while all other predictors are held constant. A unit change in a factor refers to a comparison of a certain level to the reference level.

The logit link provides the most natural interpretation of the estimated coefficients and is therefore the default link in Minitab. A summary of the interpretation follows:

- The odds of a reference event is the ratio of P(event) to P(not event). The estimated coefficient of a predictor (factor or covariate) is the estimated change in the log of P(event)/P(not event) for each unit change in the predictor, assuming the other predictors remain constant.
- The estimated coefficients can also be used to calculate the odds ratio, or the ratio between two odds. Exponentiating the estimated coefficient of a factor yields the ratio of P(event)/P(not event) for a certain factor level compared to the reference level. The odds ratios at different values of the covariate can be constructed relative to zero. In the covariate case, it may be more meaningful to interpret the odds and not the odds ratio. Note that an estimated coefficient of zero or an odds ratio of one both imply the same thing–the factor or covariate has no effect.

To change how you view the estimated coefficients, you can change the event or reference levels in the Options subdialog box. See Setting reference and event levels.

Setting Reference Levels and Reference Events

Reference levels for factors

Minitab needs to assign one factor level as the reference level. The estimated coefficients are then interpreted relative to this reference level. Minitab designates the reference level based on the data type:

- For numeric factors, the reference level is the level with the least numeric value.
- For date/time factors, the reference level is the level with the earliest date/time.
- For text factors, the reference level is the level that is first in alphabetical order.

You can change the default reference level in the Options subdialog box. To change the reference level for a factor, specify the factor variable followed by the new reference level in the **Reference factor level** box. You can specify reference levels for more than one factor at the same time. If the levels are text or date/time, enclose them in quotes.

If you have defined a value order for a text factor, the default rule above does not apply. Minitab designates the first value in the defined order as the reference value. See Ordering Text Categories.

Logistic regression creates a set of design variables for each factor in the model. If there are k levels, there will be k-1 design variables and the reference level will be coded as 0. Here are two examples of the default coding scheme:

Factor A with 4 levels

(1 2 3 4)

		A1	**A2**	**A3**
reference level is 1	1	0	0	0
	2	1	0	0
	3	0	1	0
	4	0	0	1

Factor B with 3 levels

(Humidity Pressure Temp)

		B1	**B2**
reference level is Humidity	Humidity	0	0
	Pressure	1	0
	Temp	0	1

Reference event for the response variable

Minitab needs to designate one of the response values as the reference event. Minitab defines the reference event based on the data type:

- For numeric factors, the reference event is the greatest numeric value.
- For date/time factors, the reference event is the most recent date/time.
- For text factors, the reference event is the last in alphabetical order.

You can change the default reference event in the Options subdialog box. To change the event, specify the new event value in the **Event** box. To change the order of response values in ordinal logistic regression, specify the new order in the **Order of the response values** box.

If you have defined a value order for a text factor, the default rule above does not apply. Minitab designates the last value in the defined order as the reference event. See Ordering Text Categories.

Entering Initial Values for Estimated Coefficients

There are several scenarios for which you might enter values for estimated coefficients. For example, you may wish to give starting estimates so that the algorithm converges to a solution, or you may wish to validate a model with an independent sample.

- Convergence – The maximum likelihood solution may not converge if the starting estimates are not in the neighborhood of the true solution. If the algorithm does not converge to a solution, you can specify what you think are good starting values for parameter estimates in **Starting estimates for algorithm** in the **Options** subdialog box.
- Validation – You may also wish to validate the model with an independent sample. This is done by splitting the data into two subsets. Use the first set to estimate and store the coefficients. If you enter these estimates in **Estimates for validation model** in the **Options** subdialog box, Minitab will use these values as the parameter estimates rather than calculating new parameter estimates. Then, you can assess the model fit for the independent sample.

In both cases, enter a column with the first entry being the constant estimate, and the remaining entries corresponding to the model terms in the order in which they appear in the **Model** box or the output.

Groups for the Hosmer-Lemeshow Goodness-of-Fit Test

The Hosmer-Lemeshow statistic in binary logistic regression is the chi-square goodness-of-fit statistic from a 2 x (the number of groups) table. The default number of groups is 10. This may work for a large number of problems but if the number of distinct factor/covariate patterns is small or large you may wish to adjust the number of groups. Hosmer and Lemeshow suggest using a minimum of six groups. See [21] for details.

Binary Logistic Regression – Results

Stat > Regression > Binary Logistic Regression > Results

You can control the amount of displayed output.

Dialog box items

Control the Display of Results

Display nothing: Choose to suppress all printed output, but do all requested storage and display graphs.

Response information, regression table, log-likelihood, and test that all slopes equal 0: Choose to print the response information, logistic regression table, log-likelihood, and test that all slopes equal 0.

In addition, 3 goodness-of-fit tests, table of observed and expected frequencies, and measures of association: Choose to print the Pearson, deviance, and Hosmer-Lemeshow goodness-of-fit tests, the table of frequencies, and measures of association in addition to the output described above.

In addition, lists of factor level values, tests for terms with more than 1 degree of freedom, and 2 additional goodness-of-fit tests: Choose to print factor level values, tests for terms with more than 1 degree of freedom, and two Brown goodness-of-fit-tests in addition to all of the output mentioned above.

Show log-likelihood for each iteration of algorithm: Check to display the log-likelihood at each iteration of the parameter estimation process.

Binary Logistic Regression – Storage

Stat > Regression > Binary Logistic Regression > Storage

Stores diagnostic measures, characteristics of the estimated equation, and aggregated data.

Dialog box items

Diagnostic Measures

Pearson residuals: Check to store the Pearson residuals.

Standardized Pearson residuals: Check to store the standardized Pearson residuals.

Deviance residuals: Check to store the deviance residuals.

Delta chi-square: Check to store the change in the chi-square.

Delta deviance: Check to store the change in the deviance statistic.

Delta beta (standardized): Check to store the change in the standardized estimated coefficients.

Delta beta: Check to store the change in the estimated coefficients.

Leverage (Hi): Check to store the leverages.

Characteristics of Estimated Equation

Event probability: Check to store the predicted event probabilities. Minitab stores the probability of success in the first column, and optionally, the probability of failure in the second column.

Coefficients: Check to store the estimated coefficients for the fitted model down a column in the order that they appear in the model. See Interpreting estimated coefficients in binary logistic regression.

Standard error of coefficients: Check to store the standard errors of the estimated coefficients down a column in the order that they appear in the model.

Variance/covariance matrix: Check to store a (d x d) matrix $(\mathbf{X'}\ \mathbf{W}\ \mathbf{X})^{-1}$, where d is the number of parameters in the model. The $(\mathbf{X'}\ \mathbf{W}\ \mathbf{X})^{-1}$ matrix is the variance-covariance matrix of the estimated coefficients.

Log-likelihood for last iteration: Check to store the last log-likelihood in a constant.

Aggregated Data

Number of occurrences of the event: Check to store the number of occurrences for the jth factor/covariate pattern. For Binary Logistic Regression, you may specify one or two columns--the first column stores the number of successes, and the second column stores the number of failures for the jth factor/covariate pattern.

Number of trials: Check to store the number of trials for each factor/covariate pattern.

Example of Binary Logistic Regression

You are a researcher who is interested in understanding the effect of smoking and weight upon resting pulse rate. Because you have categorized the response–pulse rate–into low and high, a binary logistic regression analysis is appropriate to investigate the effects of smoking and weight upon pulse rate.

1 Open the worksheet EXH_REGR.MTW.

2 Choose **Stat > Regression > Binary Logistic Regression**.

3 In **Response**, enter **RestingPulse**. In **Model**, enter **Smokes Weight**. In **Factors** (**optional**), enter **Smokes**.

4 Click **Graphs**. Check **Delta chi-square vs probability** and **Delta chi-square vs leverage**. Click **OK**.

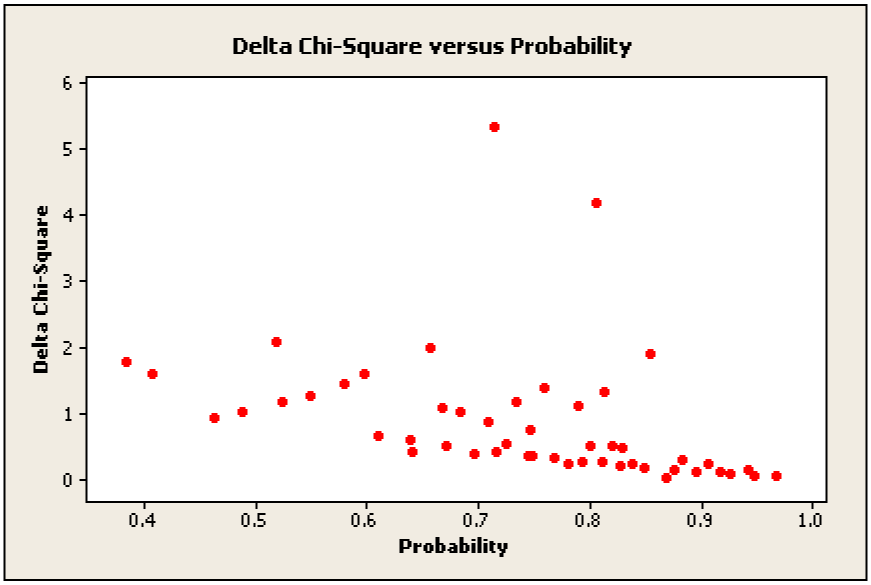
5 Click **Results**. Choose **In addition, list of factor level values, tests for terms with more than 1 degree of freedom, and 2 additional goodness-of-fit tests**. Click **OK** in each dialog box.

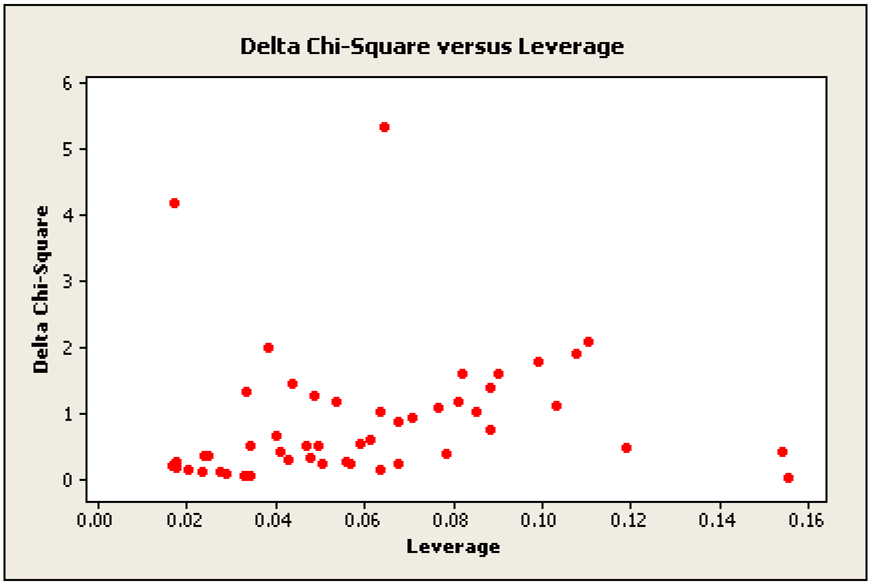
Session window output

Binary Logistic Regression: RestingPulse versus Smokes, Weight

```
Link Function: Logit


Response Information

Variable      Value  Count
RestingPulse  Low       70  (Event)
              High      22
              Total     92


Factor Information

Factor  Levels  Values
Smokes       2  No, Yes


Logistic Regression Table

                                                Odds     95% CI
Predictor       Coef    SE Coef      Z      P  Ratio  Lower  Upper
Constant    -1.98717    1.67930  -1.18  0.237
Smokes
 Yes        -1.19297   0.552980  -2.16  0.031   0.30   0.10   0.90
Weight     0.0250226  0.0122551   2.04  0.041   1.03   1.00   1.05


Log-Likelihood = -46.820
Test that all slopes are zero: G = 7.574, DF = 2, P-Value = 0.023


Goodness-of-Fit Tests

Method                 Chi-Square  DF      P
Pearson                   40.8477  47  0.724
Deviance                  51.2008  47  0.312
Hosmer-Lemeshow            4.7451   8  0.784
Brown:
General Alternative        0.9051   2  0.636
Symmetric Alternative      0.4627   1  0.496


Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

                             Group

Value    1    2    3    4    5    6    7     8    9   10  Total
Low
  Obs    4    6    6    8    8    6    8    12   10    2     70
  Exp  4.4  6.4  6.3  6.6  6.9  7.2  8.3  12.9  9.1  1.9
High
  Obs    5    4    3    1    1    3    2     3    0    0     22
  Exp  4.6  3.6  2.7  2.4  2.1  1.8  1.7   2.1  0.9  0.1
Total    9   10    9    9    9    9   10    15   10    2     92


Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs       Number  Percent  Summary Measures
Concordant    1045     67.9  Somers' D              0.38
Discordant     461     29.9  Goodman-Kruskal Gamma  0.39
Ties            34      2.2  Kendall's Tau-a        0.14
Total         1540    100.0
```

Graph window output





Interpreting the results

The Session window output contains the following seven parts:

Response Information – displays the number of missing observations and the number of observations that fall into each of the two response categories. The response value that has been designated as the reference event is the first entry under Value and labeled as the event. In this case, the reference event is low pulse rate (see Factor variables and reference levels).

Factor Information – displays all the factors in the model, the number of levels for each factor, and the factor level values. The factor level that has been designated as the reference level is first entry under Values, the subject does not smoke (see Factor variables and reference levels).

Logistic Regression Table – shows the estimated coefficients, standard error of the coefficients, z-values, and p-values. When you use the logit link function, you also see the odds ratio and a 95% confidence interval for the odds ratio.

- From the output, you can see that the estimated coefficients for both Smokes (z = -2.16, p = 0.031) and Weight (z = 2.04, p = 0.041) have p-values less than 0.05, indicating that there is sufficient evidence that the coefficients are not zero using an α-level of 0.05.
- The estimated coefficient of -1.193 for Smokes represents the change in the log of P(low pulse)/P(high pulse) when the subject smokes compared to when he/she does not smoke, with the covariate Weight held constant. The

estimated coefficient of 0.0250 for Weight is the change in the log of P(low pulse)/P(high pulse) with a 1 unit (1 pound) increase in Weight, with the factor Smokes held constant.

- Although there is evidence that the estimated coefficient for Weight is not zero, the odds ratio is very close to one (1.03), indicating that a one pound increase in weight minimally effects a person's resting pulse rate. A more meaningful difference would be found if you compared subjects with a larger weight difference (for example, if the weight unit is 10 pounds, the odds ratio becomes 1.28, indicating that the odds of a subject having a low pulse increases by 1.28 times with each 10 pound increase in weight).
- For Smokes, the negative coefficient of -1.193 and the odds ratio of 0.30 indicate that subjects who smoke tend to have a higher resting pulse rate than subjects who do not smoke. Given that subjects have the same weight, the odds ratio can be interpreted as the odds of smokers in the sample having a low pulse being 30% of the odds of non-smokers having a low pulse.

Next, the last Log-Likelihood from the maximum likelihood iterations is displayed along with the statistic G. This statistic tests the null hypothesis that all the coefficients associated with predictors equal zero versus these coefficients not all being equal to zero. In this example, G = 7.574, with a p-value of 0.023, indicating that there is sufficient evidence that at least one of the coefficients is different from zero, given that your accepted α-level is greater than 0.023.

- Note that for factors with more than 1 degree of freedom, Minitab performs a multiple degrees of freedom test with a null hypothesis that all the coefficients associated with the factor are equal to 0 versus them not all being equal to 0. This example does not have a factor with more than 1 degree of freedom.

Goodness-of-Fit Tests – displays Pearson, deviance, and Hosmer-Lemeshow goodness-of-fit tests. In addition, two Brown tests-general alternative and symmetric alternative-are displayed because you have chosen the logit link function and the selected option in the Results subdialog box. The goodness-of-fit tests, with p-values ranging from 0.312 to 0.724, indicate that there is insufficient evidence to claim that the model does not fit the data adequately. If the p-value is less than your accepted α-level, the test would reject the null hypothesis of an adequate fit.

Table of Observed and Expected Frequencies – allows you to see how well the model fits the data by comparing the observed and expected frequencies. There is insufficient evidence that the model does not fit the data well, as the observed and expected frequencies are similar. This supports the conclusions made by the Goodness of Fit Tests.

Measures of Association – displays a table of the number and percentage of concordant, discordant, and tied pairs, as well as common rank correlation statistics. These values measure the association between the observed responses and the predicted probabilities.

- The table of concordant, discordant, and tied pairs is calculated by pairing the observations with different response values. Here, you have 70 individuals with a low pulse and 22 with a high pulse, resulting in 70 * 22 = 1540 pairs with different response values. Based on the model, a pair is concordant if the individual with a low pulse rate has a higher probability of having a low pulse, discordant if the opposite is true, and tied if the probabilities are equal. In this example, 67.9% of pairs are concordant and 29.9% are discordant. You can use these values as a comparative measure of prediction, for example in comparing fits with different sets of predictors or with different link functions.
- Somers' D, Goodman-Kruskal Gamma, and Kendall's Tau-a are summaries of the table of concordant and discordant pairs. These measures most likely lie between 0 and 1 where larger values indicate that the model has a better predictive ability. In this example, the measure range from 0.14 to 0.39 which implies less than desirable predictive ability.

Plots – In the example, you chose two diagnostic plots-delta Pearson χ^2 versus the estimated event probability and delta Pearson χ^2 versus the leverage. Delta Pearson χ^2 for the jth factor/covariate pattern is the change in the Pearson χ^2 when all observations with that factor/covariate pattern are omitted. These two graphs indicate that two observations are not well fit by the model (high delta χ^2). A high delta χ^2 can be caused by a high leverage and/or a high Pearson residual. In this case, a high Pearson residual caused the large delta χ^2, because the leverages are less than 0.1. Hosmer and Lemeshow indicate that delta χ^2 or delta deviance greater than 3.84 is large.

If you choose Editor > Brush, brush these points, and then click on them, they will be identified as data values 31 and 66. These are individuals with a high resting pulse, who do not smoke, and who have smaller than average weights (Weight = 116, 136 pounds). You might further investigate these cases to see why the model did not fit them well.

Ordinal Logistic Regression

Ordinal Logistic Regression

Stat > Regression > Ordinal Logistic Regression

Use ordinal logistic regression to perform logistic regression on an ordinal response variable. Ordinal variables are categorical variables that have three or more possible levels with a natural ordering, such as strongly disagree, disagree, neutral, agree, and strongly agree. A model with one or more predictors is fit using an iterative-reweighted least squares algorithm to obtain maximum likelihood estimates of the parameters [25].

Parallel regression lines are assumed, and therefore, a single slope is calculated for each covariate. In situations where this assumption is not valid, nominal logistic regression, which generates separate logit functions, is more appropriate.

Dialog box items

Response: Choose if the response data has been entered as raw data or as two columns – one containing the response values and one column containing the frequencies. Then enter the column containing the number response values in the text box.

with frequency (optional): If the data has been entered as two columns – one containing the response values and one column containing the frequencies – enter the column containing the frequencies in the text box.

Model: Specify the terms to be included in the model.

Factors (optional): Specify which of the predictors are factors. Minitab assumes all variables in the model are covariates unless specified to be factors here. Continuous predictors must be modeled as covariates; categorical predictors must be modeled as factors.

<Options>

<Results>

<Storage>

Data – Ordinal Logistic Regression

Your data may be arranged in one of two ways: as raw data or as frequency data. See Entering data for response variables.

Factors, covariates, and response data can be numeric, text, or date/time. The reference level and the reference event depend on the data type. See Setting reference level and events for details.

The predictors may either be factors (nominal variables) or covariates (continuous variables). Factors may be crossed or nested. Covariates may be crossed with each other or with factors, or nested within factors.

The model can include up to 9 factors and 50 covariates. Unless you specify a predictor in the model as a factor, the predictor is assumed to be a covariate. Model continuous predictors as covariates and categorical predictors as factors. See Specifying the model in logistic regression for more information.

Minitab automatically omits observations with missing values from all calculations.

Entering Data for Response Variables

Data used for input to the logistic regression procedures may be arranged in two different ways in your worksheet: as raw (categorical) data, or as frequency (collapsed) data. For binary logistic regression, there are three additional ways to arrange the data in your worksheet: as successes and trials, as successes and failures, or as failures and trials. These ways are illustrated here for the same data.

The response entered as raw data or as frequency data :

Raw Data: one row for each observation

C1	C2	C3	C4
Response		Factor	Covar
0		1	12
1		1	12
1		1	12
.		.	.
.		.	.
.		.	.
1		1	12
0		2	12
1		2	12
.		.	.
.		.	.
.		.	.

Frequency Data: one row for each combination of factor and covariate

C1	C2	C3	C4
Response	Count	Factor	Covar
0	1	1	12
1	19	1	12
0	1	2	12
1	19	2	12
0	5	1	24
1	15	1	24
0	4	2	24
1	16	2	24
0	7	1	50
1	13	1	50
0	8	2	50
1	12	2	50

1	2	12	0	11	1	125
.	.	.	1	2	1	125
.	.	.	0	9	2	125
			1	11	2	125
			0	19	1	200
			1	1	1	200
			0	18	2	200
			1	2	2	200

The binary response entered as the number of successes, failures, or trials. Enter one row for each combination of factor and covariate.

Successes and Trials

C1	C2	C3	C4
S	T	Factor	Covar
19	20	1	12
19	20	2	12
15	20	1	24
16	20	2	24
13	20	1	50
12	20	2	50
9	20	1	125
11	20	2	125
1	20	1	200
2	20	2	200

Successes and Failures

C1	C2	C3	C4
S	F	Factor	Covar
19	1	1	12
19	1	2	12
15	5	1	24
16	4	2	24
13	7	1	50
12	8	2	50
9	11	1	125
11	9	2	125
1	19	1	200
2	18	2	200

Failures and Trials

C1	C2	C3	C4
F	T	Factor	Covar
1	20	1	12
1	20	2	12
5	20	1	24
4	20	2	24
7	20	1	50
8	20	2	50
11	20	1	125
9	20	2	125
19	20	1	200
18	20	2	200

To do an ordinal logistic regression

1 Choose **Stat > Regression > Ordinal Logistic Regression**.

2 Do one of the following:

- If you have raw response data, in **Response**, enter the numeric column containing the response data.
- If you have frequency data, in **Response**, enter the numeric column containing the response values. Then, in **Frequency**, enter the variable containing the counts.

See Entering data for response variables.

3 In **Model**, enter the model terms. See Specify the model terms.

4 If you like, use one or more of the dialog box options, then click **OK**.

Ordinal Logistic Regression – Options

Stat > Regression > Ordinal Logistic Regression > Options

Provides options that allow you to choose a link function, the reference level and event, and specify starting values.

Dialog box items

Link Functions Minitab provides three link functions, allowing you to fit a broad class of binary response models. The link function is the inverse of a distribution function. See Link Functions for more information.

Logit: Choose to use the logit link function. This is the default.

Normit/Probit: Choose to use the normit link function.

Gompit/Complementary log-log: Choose to use the gompit link function (also called the complementary log-log function). The gompit function is the inverse of the Gompertz distribution function.

Reference Options

Order of the response values: Enter the order of the response values from lowest to highest if you want to change the reference event. See Interpreting the parameter estimates.

Reference factor level (enter factor followed by level): Specify the reference factor level by entering the factor column followed by the reference level. (Text and date/time levels must be enclose in quotes.) See Interpreting the parameter estimates.

Algorithm Options

Starting estimates for algorithm: Specify the column containing the initial values for model parameters. The column containing the initial values must have one row for each estimated coefficient in the model. The starting values for all the constants appear first, followed by starting values the predictors in the model. There must be one row for each degree of freedom.

Estimates for validation model: Specify the column containing the estimated model parameters. Minitab will then fit the validation model.

Maximum number of iterations: Enter the maximum number of iterations you want Minitab to perform to reach convergence. The default value is 20. Minitab obtains maximum likelihood estimates through an iterative process. If the maximum number of iterations is reached before convergence, the command terminates.

Interpreting Estimated Coefficients in Ordinal Logistic Regression

The interpretation of the estimated coefficients depends on: the link function (see link function), reference event, and reference factor levels (see Setting reference levels and reference events). The estimated coefficient associated with a predictor (factor or covariate) represents the change in the link function for each unit change in the predictor, while all other predictors are held constant. A unit change in a factor refers to a comparison of a certain level to the reference level.

The logit link provides the most natural interpretation of the estimated coefficients and is therefore the default link in Minitab. A summary of the interpretation follows:

- The odds of a reference event is the ratio of P(event) to P(not event). The estimated coefficient of a predictor (factor or covariate) is the estimated change in the log of P(event)/P(not event) for each unit change in the predictor, assuming the other predictors remain constant.
- The estimated coefficient can also be used to calculate the odds ratio, or the ratio between two odds. Exponentiating the estimated coefficient of a factor yields the ratio of P(event)/P(not event) for a certain factor level compared to the reference level. The odds ratios at different values of the covariate can be constructed relative to zero. In the covariate case, it may be more meaningful to interpret the odds and not the odds ratio. Note that a coefficient of zero or an odds ratio of one both imply the same thing–the factor or covariate has no effect.

To change how you view the estimated coefficients, you can change the event or reference levels in the Options subdialog box. See Setting reference levels and reference events. For example, if your responses were coded Low, Medium, and High, rather than 1, 2, 3, the default alphabetical ordering of the responses would be improper and you should change the order.

Setting Reference Levels and Reference Events

Reference levels for factors

Minitab needs to assign one factor level as the reference level. The estimated coefficients are then interpreted relative to this reference level. Minitab designates the reference level based on the data type:

- For numeric factors, the reference level is the level with the least numeric value.
- For date/time factors, the reference level is the level with the earliest date/time.
- For text factors, the reference level is the level that is first in alphabetical order.

You can change the default reference level in the Options subdialog box. To change the reference level for a factor, specify the factor variable followed by the new reference level in the **Reference factor level** box. You can specify reference levels for more than one factor at the same time. If the levels are text or date/time, enclose them in quotes.

If you have defined a value order for a text factor, the default rule above does not apply. Minitab designates the first value in the defined order as the reference value. See Ordering Text Categories.

Logistic regression creates a set of design variables for each factor in the model. If there are k levels, there will be k-1 design variables and the reference level will be coded as 0. Here are two examples of the default coding scheme:

Factor A with 4 levels

(1 2 3 4)

		A1	**A2**	**A3**
reference level is 1	1	0	0	0
	2	1	0	0
	3	0	1	0
	4	0	0	1

Factor B with 3 levels

(Humidity Pressure Temp)

		B1	**B2**
reference level is Humidity	Humidity	0	0
	Pressure	1	0
	Temp	0	1

Reference event for the response variable

Minitab needs to designate one of the response values as the reference event. Minitab defines the reference event based on the data type:

- For numeric factors, the reference event is the greatest numeric value.
- For date/time factors, the reference event is the most recent date/time.
- For text factors, the reference event is the last in alphabetical order.

You can change the default reference event in the Options subdialog box. To change the event, specify the new event value in the **Event** box. To change the order of response values in ordinal logistic regression, specify the new order in the **Order of the response values** box.

If you have defined a value order for a text factor, the default rule above does not apply. Minitab designates the last value in the defined order as the reference event. See Ordering Text Categories.

Entering Initial Values for Estimated Coefficients

There are several scenarios for which you might enter values for estimated coefficients. For example, you may wish to give starting estimates so that the algorithm converges to a solution, or you may wish to validate a model with an independent sample.

- Convergence – The maximum likelihood solution may not converge if the starting estimates are not in the neighborhood of the true solution. If the algorithm does not converge to a solution, you can specify what you think are good starting values for parameter estimates in **Starting estimates for algorithm** in the **Options** subdialog box.
- Validation – You may also wish to validate the model with an independent sample. This is done by splitting the data into two subsets. Use the first set to estimate and store the coefficients. If you enter these estimates in **Estimates for validation model** in the **Options** subdialog box, Minitab will use these values as the parameter estimates rather than calculating new parameter estimates. Then, you can assess the model fit for the independent sample.

In both cases, enter a column with the first entry being the constant estimate, and the remaining entries corresponding to the model terms in the order in which they appear in the **Model** box or the output.

Ordinal Logistic Regression – Results

Stat > Regression > Ordinal Logistic Regression > Results

Allows you to control the amount of printed output.

Dialog box items

Control the Display of Results

Display nothing: Choose to suppress all printed output, but do all requested storage and display graphs.

Response information, regression table, log-likelihood, and test that all slopes equal 0: Choose to print the response information, logistic regression table, log-likelihood, and test that all slopes equal 0.

In addition, 2 goodness-of-fit tests, and measures of association: Choose to print the Pearson and deviance tests, and measures of association in addition to the output described above.

In addition, lists of factor level values, and tests for terms with more than 1 degree of freedom: Choose to print factor level values, and tests for terms with more than 1 degree of freedom.

Show log-likelihood for each iteration of algorithm: Check to display the log-likelihood at each iteration of the parameter estimation process.

Ordinal Logistic Regression – Storage

Stat > Regression > Ordinal Logistic Regression > Storage

Stores characteristics of the estimated equation, event probabilities, and aggregated data.

Dialog box items

Characteristics of Estimated Equation

Coefficients: Check to store the estimated coefficients for the fitted model in a column in the order that they appear in the model. See Interpreting estimated coefficients.

Standard error of coefficients: Check to store the standard errors of the estimated coefficients down a column in the order that they appear in the model.

Variance/covariance matrix: Check to store a (d x d) matrix ($\mathbf{X' W X}$)$^{-1}$, where d is the number of parameters in the model. The ($\mathbf{X' W X}$)$^{-1}$ matrix is the variance-covariance matrix of the estimated coefficients.

Log-likelihood for last iteration: Check to store the last log-likelihood in a constant.

Event probabilities and/or aggregated data

Number of trials: Check to store the number of trials for each factor/covariate pattern.

For the remaining three storage items

Enter the number of events: Enter the number of levels (distinct values) of the response variable.

Event probabilities: Check to store the predicted event probabilities, one probability for each distinct value of the response.

Cumulative event probabilities: Check to store the cumulative event probabilities.

Number of occurrences: Check to store the number of occurrences for the jth factor/covariate pattern.

Ordinal Logistic Regression Link Functions

Minitab provides three link functions–logit (the default), normit (also called probit), and gompit (also called complementary log-log)–allowing you to fit a broad class of ordinal response models. These are the inverse of the cumulative logistic distribution function (logit), the inverse of the cumulative standard normal distribution function (normit), and the inverse of the Gompertz distribution function (gompit).

You want to choose a link function that results in a good fit to your data. Goodness-of-fit statistics can be used to compare the fits using different link functions. Certain link functions may be used for historical reasons or because they have a special meaning in a discipline.

An advantage of the logit link function is that it provides an estimate of the odds ratios. The logit link function is the default. For a comparison of link functions, see [19].

Example of Ordinal Logistic Regression

Suppose you are a field biologist and you believe that adult population of salamanders in the Northeast has gotten smaller over the past few years. You would like to determine whether any association exists between the length of time a hatched salamander survives and level of water toxicity, as well as whether there is a regional effect. Survival time is coded as 1 if < 10 days, 2 = 10 to 30 days, and 3 = 31 to 60 days.

1 Open the worksheet EXH_REGR.MTW.

2 Choose **Stat > Regression > Ordinal Logistic Regression**.

3 In **Response**, enter **Survival**. In **Model**, enter **Region ToxicLevel**. In **Factors** (**optional**), enter **Region**.

4 Click **Results**. Choose **In addition, list of factor level values, and tests for terms with more than 1 degree of freedom**. Click **OK** in each dialog box.

Session window output

Ordinal Logistic Regression: Survival versus Region, ToxicLevel

```
Link Function: Logit


Response Information

Variable  Value  Count
Survival  1         15
          2         46
          3         12
          Total     73


Factor Information
Factor  Levels  Values
Region       2  1, 2


Logistic Regression Table

                                                Odds     95% CI
Predictor       Coef    SE Coef      Z      P  Ratio  Lower  Upper
Const(1)    -7.04343    1.68017  -4.19  0.000
Const(2)    -3.52273    1.47108  -2.39  0.017
Region
 2          0.201456   0.496153   0.41  0.685   1.22   0.46   3.23
ToxicLevel  0.121289  0.0340510   3.56  0.000   1.13   1.06   1.21


Log-Likelihood = -59.290
Test that all slopes are zero: G = 14.713, DF = 2, P-Value = 0.001


Goodness-of-Fit Tests

Method    Chi-Square   DF      P
Pearson      122.799  122  0.463
Deviance     100.898  122  0.918


Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs       Number  Percent  Summary Measures
Concordant    1127     79.3  Somers' D              0.59
Discordant     288     20.3  Goodman-Kruskal Gamma  0.59

Ties             7      0.5  Kendall's Tau-a        0.32
Total         1422    100.0
```

Interpreting the results

The Session window contains the following five parts:

Response Information displays the number of observations that fall into each of the response categories, and the number of missing observations. The ordered response values, from lowest to highest, are shown. Here, we use the default coding scheme which orders the values from lowest to highest: 1 is < 10 days, 2 = 10 to 30 days, and 3 = 31 to 60 days (see Reference event for the response variable on page).

Factor Information displays all the factors in the model, the number of levels for each factor, and the factor level values. The factor level that has been designated as the reference level is first entry under Values, region 1 (see Reference event for the response variable on page).

Logistic Regression Table shows the estimated coefficients, standard error of the coefficients, z-values, and p-values. When you use the logit link function, you see the calculated odds ratio, and a 95% confidence interval for the odds ratio.

- The values labeled Const(1) and Const(2) are estimated intercepts for the logits of the cumulative probabilities of survival for <10 days, and for 10-30 days, respectively. Because the cumulative probability for the last response value is 1, there is not need to estimate an intercept for 31-60 days.
- The coefficient of 0.2015 for Region is the estimated change in the logit of the cumulative survival time probability when the region is 2 compared to region being 1, with the covariate ToxicLevel held constant. Because the p-value for estimated coefficient is 0.685, there is insufficient evidence to conclude that region has an effect upon survival time.
- There is one estimated coefficient for each covariate, which gives parallel lines for the factor levels. Here, the estimated coefficient for the single covariate, ToxicLevel, is 0.121, with a p-value of < 0.0005. The p-value indicates that for most α-levels, there is sufficient evidence to conclude that the toxic level affects survival. The positive coefficient, and an odds ratio that is greater than one indicates that higher toxic levels tend to be associated with lower values of survival. Specifically, a one-unit increase in water toxicity results in a 13% increase in the odds that a salamander lives less than or equal to 10 days versus greater than 30 days and that the salamander lives less than or equal to 30 days versus greater than 30 days.
- Next displayed is the last Log-Likelihood from the maximum likelihood iterations along with the statistic G. This statistic tests the null hypothesis that all the coefficients associated with predictors equal zero versus at least one coefficient is not zero. In this example, G = 14.713 with a p-value of 0.001, indicating that there is sufficient evidence to conclude that at least one of the estimated coefficients is different from zero.

Goodness-of-Fit Tests displays both Pearson and deviance goodness-of-fit tests. In our example, the p-value for the Pearson test is 0.463, and the p-value for the deviance test is 0.918, indicating that there is insufficient evidence to claim that the model does not fit the data adequately. If the p-value is less than your selected α-level, the test rejects the null hypothesis that the model fits the data adequately.

Measures of Association display a table of the number and percentage of concordant, discordant and tied pairs, and common rank correlation statistics. These values measure the association between the observed responses and the predicted probabilities.

- The table of concordant, discordant, and tied pairs is calculated by pairing the observations with different response values. Here, we have 15 1's, 46 2's, and 12 3's, resulting in 15 x 46 + 15 x 12 + 46 x 12 = 1422 pairs of different response values. For pairs involving the lowest coded response value (the 1–2 and 1–3 value pairs in the example), a pair is concordant if the cumulative probability up to the lowest response value (here 1) is greater for the observation with the lowest value. This works similarly for other value pairs. For pairs involving responses coded as 2 and 3 in our example, a pair is concordant if the cumulative probability up to 2 is greater for the observation coded as 2. The pair is discordant if the opposite is true. The pair is tied if the cumulative probabilities are equal. In our example, 79.3% of pairs are concordant, 20.3% are discordant, and 0.5% are ties. You can use these values as a comparative measure of prediction. For example, you can use them in evaluating predictors and different link functions.
- Somers' D, Goodman-Kruskal Gamma, and Kendall's Tau-a are summaries of the table of concordant and discordant pairs. The numbers have the same numerator: the number of concordant pairs minus the number of discordant pairs. The denominators are the total number of pairs with Somers' D, the total number of pairs excepting ties with Goodman-Kruskal Gamma, and the number of all possible observation pairs for Kendall's Tau-a. These measures most likely lie between 0 and 1 where larger values indicate a better predictive ability of the model.

Nominal Logistic Regression

Nominal Logistic Regression

Stat > Regression > Nominal Logistic Regression

Use nominal logistic regression performs logistic regression on a nominal response variable using an iterative-reweighted least squares algorithm to obtain maximum likelihood estimates of the parameters [25]. Nominal variables are categorical variables that have three or more possible levels with no natural ordering. For example, the levels in a food tasting study may include crunchy, mushy, and crispy.

Dialog box items

Response: Choose if the response data has been entered as raw data or as two columns – one containing the response values and one column containing the frequencies. Then enter the column containing the response values.

with frequency (optional): If the data has been entered as two columns – one containing the response values and one column containing the frequencies – enter the column containing the frequencies in the text box.

Model: Specify the terms to be included in the model. See Specifying the Model.

Factors (optional): Specify which of the predictors are factors. Minitab assumes all variables in the model are covariates unless specified to be factors here. Continuous predictors must be modeled as covariates; categorical predictors must be modeled as factors.

<Options>

<Results>

<Storage>

Data – Nominal Logistic Regression

Your data may be arranged in one of two ways: as raw data or as frequency data. See Entering data for response variables.

Factors, covariates, and response data can be numeric, text, or date/time. The **reference level** and the **reference event** depend on the data type. See Setting reference levels and reference events for details.

The predictors may either be factors (nominal variables) or covariates (continuous variables). Factors may be crossed or nested. Covariates may be crossed with each other or with factors, or nested within factors.

The model can include up to 9 factors and 50 covariates. Unless you specify a predictor in the model as a factor, the predictor is assumed to be a covariate. Model continuous predictors as covariates and categorical predictors as factors. See Specifying the model.

Minitab automatically omits observations with missing values from all calculations.

Entering Data for Response Variables

Data used for input to the logistic regression procedures may be arranged in two different ways in your worksheet: as raw (categorical) data, or as frequency (collapsed) data. For binary logistic regression, there are three additional ways to arrange the data in your worksheet: as successes and trials, as successes and failures, or as failures and trials. These ways are illustrated here for the same data.

The response entered as raw data or as frequency data :

Raw Data: one row for each observation

C1	C2	C3	C4
Response		**Factor**	**Covar**
0		1	12
1		1	12
1		1	12
.		.	.
.		.	.
.		.	.
1		1	12
0		2	12
1		2	12
.		.	.
.		.	.
.		.	.
1		2	12
.		.	.
.		.	.

Frequency Data: one row for each combination of factor and covariate

C1	C2	C3	C4
Response	**Count**	**Factor**	**Covar**
0	1	1	12
1	19	1	12
0	1	2	12
1	19	2	12
0	5	1	24
1	15	1	24
0	4	2	24
1	16	2	24
0	7	1	50
1	13	1	50
0	8	2	50
1	12	2	50
0	11	1	125
1	2	1	125
0	9	2	125
1	11	2	125
0	19	1	200
1	1	1	200
0	18	2	200
1	2	2	200

The binary response entered as the number of successes, failures, or trials. Enter one row for each combination of factor and covariate.

Successes and Trials

C1	C2	C3	C4
S	T	Factor	Covar
19	20	1	12
19	20	2	12
15	20	1	24
16	20	2	24
13	20	1	50
12	20	2	50
9	20	1	125
11	20	2	125
1	20	1	200
2	20	2	200

Successes and Failures

C1	C2	C3	C4
S	F	Factor	Covar
19	1	1	12
19	1	2	12
15	5	1	24
16	4	2	24
13	7	1	50
12	8	2	50
9	11	1	125
11	9	2	125
1	19	1	200
2	18	2	200

Failures and Trials

C1	C2	C3	C4
F	T	Factor	Covar
1	20	1	12
1	20	2	12
5	20	1	24
4	20	2	24
7	20	1	50
8	20	2	50
11	20	1	125
9	20	2	125
19	20	1	200
18	20	2	200

To do a nominal logistic regression

1 Choose **Stat > Regression > Nominal Logistic Regression**.

2 Do one of the following:

- If you have raw response data, in **Response**, enter the numeric column containing the response data.
- If you have frequency data, in **Response**, enter the numeric column containing the response values. Then, in **Frequency**, enter the variable containing the counts.

See Entering data for response variables.

3 In **Model**, enter the model terms. See Specifying a model in logistic regression.

4 If you like, use one or more of the options listed below, then click **OK**.

Specifying the Model in Logistic Regression

The logistic regression procedures can fit models with:

- up to 9 factors and up to 50 covariates
- crossed or nested factors
- covariates that are crossed with each other or with factors, or nested within factors

Model continuous predictors as covariates and categorical predictors as factors. Here are some examples. A is a factor and X is a covariate.

Model terms

A X A*X	fits a full model with a covariate crossed with a factor
A | X	an alternative way to specify the previous model
A X X*X	fits a model with a covariate crossed with itself making a squared term
A X(A)	fits a model with a covariate nested within a factor

The model for logistic regression is a generalization of the model used in Minitab's general linear model (GLM) procedure. Any model fit by GLM can also be fit by the logistic regression procedures. For a discussion of specifying models in general, see Specifying the Model Terms and Specifying Reduced Models. In the logistic regression commands, Minitab assumes any variable in the model is a covariate unless the variable is specified as a factor. In contrast, GLM assumes that any variable in the model is a factor unless the variable is specified as a covariate. Be sure to specify which predictors are factors in the main dialog box.

Model restrictions

Logistic regression models in Minitab have the same restrictions as GLM models:

- There must be enough data to estimate all the terms in your model, so that the model is full rank. Minitab will automatically determine if your model is **full rank** and display a message. In most cases, eliminating some unimportant high-order interactions in your model should solve your problem.
- The model must be hierarchical. In a hierarchical model, if an interaction term is included, all lower order interactions and main effects that comprise the interaction term must appear in the model.

Nominal Logistic Regression – Options

Stat > Regression > Nominal Logistic Regression > Options

Provides options that allow you to choose the reference level and event, and specify starting values.

Dialog box items

Reference Options

Reference event: Enter the reference event if you want to change the default event. (Text and date/time levels must be enclose in quotes.) For information on why you might want to change the reference event, see Interpreting the parameter estimates.

Reference factor level (enter factor followed by level): Enter the factor column followed by the reference level if you want to change the default level. (Text and date/time levels must be enclose in quotes.) For information on why you might want to change the reference event, see Interpreting the parameter estimates.

Algorithm Options

Starting estimates for algorithm: Enter the column containing the initial values for model parameters. Specify initial values for model parameters or parameter estimates for a validation model. See Entering initial values for parameter estimates.

Estimates for validation model: Enter the column containing the estimated model parameters. Minitab will then fit the validation model. See Entering initial values for parameter estimates.

Maximum number of iterations: Change the maximum number of iterations that Minitab will perform to reach convergence. The default value is 20. Minitab's logistic regression commands obtain maximum likelihood estimates through an iterative process. If the maximum number of iterations is reached before convergence, the command terminates.

Interpreting Estimated Coefficients in Nominal Logistic Regression

The interpretation of the estimated coefficients depends upon the designated reference event and reference factor levels (see Setting reference levels and events). The estimated coefficient associated with a predictor represents the change in the particular logit for each unit change in the predictor, assuming that all other factors and covariates are held constant. A one unit change in a factor refers to a comparison of a certain level to the reference level.

If there are k distinct response values, Minitab estimates k-1 sets of estimated coefficients. These are the estimated differences in log odds or logits of levels of the response variable relative to the reference event. Each set contains a constant and coefficients for the factors and the covariates. Note that these sets of parameter estimates give nonparallel lines for the response value. The interpretation of the parameter estimates is as follows:

- The coefficient of a predictor (factor or covariate) is the estimated change in the log of P(response level)/ P(reference event) for each unit change in the predictor, assuming the other predictors remain constant.
- The coefficient can also be used to calculate the odds ratio, or the ratio between two odds. Exponentiating the estimated coefficient of a factor yields the ratio of P(response level)/P(reference event) for a certain factor level compared to the reference level. The odds ratios at different values of the covariate can be constructed relative to zero. In the covariate case, it may be more meaningful to interpret the odds and not the odds ratio. Note that a coefficient of zero or an odds ratio of one both imply the same thing–the factor or covariate has no effect.

To change how you view the parameter estimates, you can change the event or reference levels in the Options subdialog box. See Setting reference and event levels.

Setting Reference Levels and Reference Events

Reference levels for factors

Minitab needs to assign one factor level as the reference level. The estimated coefficients are then interpreted relative to this reference level. Minitab designates the reference level based on the data type:

- For numeric factors, the reference level is the level with the least numeric value.
- For date/time factors, the reference level is the level with the earliest date/time.
- For text factors, the reference level is the level that is first in alphabetical order.

You can change the default reference level in the Options subdialog box. To change the reference level for a factor, specify the factor variable followed by the new reference level in the **Reference factor level** box. You can specify reference levels for more than one factor at the same time. If the levels are text or date/time, enclose them in quotes.

If you have defined a value order for a text factor, the default rule above does not apply. Minitab designates the first value in the defined order as the reference value. See Ordering Text Categories.

Logistic regression creates a set of design variables for each factor in the model. If there are k levels, there will be k-1 design variables and the reference level will be coded as 0. Here are two examples of the default coding scheme:

Factor A with 4 levels

(1 2 3 4)

		A1	**A2**	**A3**
reference level is 1	1	0	0	0
	2	1	0	0
	3	0	1	0
	4	0	0	1

Factor B with 3 levels

(Humidity Pressure Temp)

		B1	**B2**
reference level is Humidity	Humidity	0	0
	Pressure	1	0
	Temp	0	1

Reference event for the response variable

Minitab needs to designate one of the response values as the reference event. Minitab defines the reference event based on the data type:

- For numeric factors, the reference event is the greatest numeric value.
- For date/time factors, the reference event is the most recent date/time.
- For text factors, the reference event is the last in alphabetical order.

You can change the default reference event in the Options subdialog box. To change the event, specify the new event value in the **Event** box. To change the order of response values in ordinal logistic regression, specify the new order in the **Order of the response values** box.

If you have defined a value order for a text factor, the default rule above does not apply. Minitab designates the last value in the defined order as the reference event. See Ordering Text Categories.

Entering Initial Values for Estimated Coefficients

There are several scenarios for which you might enter values for estimated coefficients. For example, you may wish to give starting estimates so that the algorithm converges to a solution, or you may wish to validate a model with an independent sample.

- Convergence – The maximum likelihood solution may not converge if the starting estimates are not in the neighborhood of the true solution. If the algorithm does not converge to a solution, you can specify what you think are good starting values for parameter estimates in **Starting estimates for algorithm** in the **Options** subdialog box.
- Validation – You may also wish to validate the model with an independent sample. This is done by splitting the data into two subsets. Use the first set to estimate and store the coefficients. If you enter these estimates in **Estimates for validation model** in the **Options** subdialog box, Minitab will use these values as the parameter estimates rather than calculating new parameter estimates. Then, you can assess the model fit for the independent sample.

In both cases, enter a column with the first entry being the constant estimate, and the remaining entries corresponding to the model terms in the order in which they appear in the **Model** box or the output.

Nominal Logistic Regression – Results

Stat > Regression > Nominal Logistic Regression > Results

Allows you to control the amount of printed output.

Dialog box items

Control the Display of Results

Display nothing: Choose to suppress all printed output, but do all requested storage and display graphs.

Response information, regression table, log-likelihood, and test that all slopes equal 0: Choose to print the response information, logistic regression table, log-likelihood, and test that all slopes equal 0.

In addition, 2 goodness-of-fit tests: Choose to print the Pearson and deviance tests in addition to the output described above.

In addition, list of factor level values, and tests for terms with more than 1 degree of freedom: Choose to print factor level values, and tests for terms with more than 1 degree of freedom.

Show log-likelihood for each iteration of algorithm: Check to display the log-likelihood at each iteration of the parameter estimation process.

Nominal Logistic Regression – Storage

Stat > Regression > Nominal Logistic Regression > Storage

Stores characteristics of the estimated equation, event probabilities, and aggregated data.

Dialog box items

Characteristics of Estimated Equation

Coefficients: Check to store the estimated coefficients for the fitted model in a column in the order that they appear in the model. See Interpreting the estimated coefficients.

Standard error of coefficients: Check to store the standard errors of the estimated coefficients down a column in the order that they appear in the model.

Variance/covariance matrix: Check to store a (d x d) matrix, $(\mathbf{X' W X})^{-1,}$ where d is the number of parameters in the model. The $(\mathbf{X' W X})^{-1}$ matrix is the variance-covariance matrix of the estimated coefficients.

Log-likelihood for last iteration: Check to store the last log-likelihood in a constant.

Event probabilities and/or aggregated data

Number of trials: Check to store the number of trials for each factor/covariate pattern.

For the remaining two storage items

Enter the number of events: Enter the number of levels (distinct values) of the response variable.

Event probabilities: Check to store the predicted event probabilities, one probability for each distinct value of the response.

Number of occurrences: Check to store the number of occurrences for the jth factor/covariate pattern.

Example of Nominal Logistic Regression

Suppose you are a grade school curriculum director interested in what children identify as their favorite subject and how this is associated with their age or the teaching method employed. Thirty children, 10 to 13 years old, had classroom instruction in science, math, and language arts that employed either lecture or discussion techniques. At the end of the school year, they were asked to identify their favorite subject. We use nominal logistic regression because the response is categorical but possesses no implicit categorical ordering.

1 Open the worksheet EXH_REGR.MTW.

2 Choose **Stat > Regression > Nominal Logistic Regression**.

3 In **Response**, enter **Subject**. In **Model**, enter **TeachingMethod Age**. In **Factors** (**optional**), enter **TeachingMethod**.

4 Click **Results**. Choose **In addition, list of factor level values, and tests for terms with more than 1 degree of freedom**. Click **OK** in each dialog box.

Session window output

Nominal Logistic Regression: Subject versus TeachingMethod, Age

```
Response Information

Variable  Value    Count
Subject   science     10  (Reference Event)
          math        11
          arts         9
          Total       30


Factor Information

Factor          Levels  Values
TeachingMethod       2  discuss, lecture


Logistic Regression Table

                                                                    95%
                                                             Odds    CI
Predictor                     Coef  SE Coef      Z      P  Ratio  Lower
Logit 1: (math/science)
Constant                  -1.12266  4.56425  -0.25  0.806
TeachingMethod
 lecture                 -0.563115  0.937591  -0.60  0.548   0.57   0.09
Age                       0.124674  0.401079   0.31  0.756   1.13   0.52
Logit 2: (arts/science)
Constant                  -13.8485  7.24256  -1.91  0.056
TeachingMethod
```

```
 lecture                    2.76992   1.37209   2.02  0.044  15.96   1.08
Age                         1.01354  0.584494   1.73  0.083   2.76   0.88



Predictor                 Upper
Logit 1: (math/science)
Constant
TeachingMethod
 lecture                   3.58
Age                        2.49
Logit 2: (arts/science)
Constant
TeachingMethod
 lecture                 234.91
Age                        8.66



Log-Likelihood = -26.446
Test that all slopes are zero: G = 12.825, DF = 4, P-Value = 0.012


Goodness-of-Fit Tests

Method    Chi-Square  DF      P
Pearson      6.95295  10  0.730
Deviance     7.88622  10  0.640
```

Interpreting the results

The Session window output contains the following five parts:

Response Information displays the number of observations that fall into each of the response categories (science, math, and language arts), and the number of missing observations. The response value that has been designated as the reference event is the first entry under Value. Here, the default coding scheme defines the reference event as science using reverse alphabetical order.

Factor Information displays all the factors in the model, the number of levels for each factor, and the factor level values. The factor level that has been designated as the reference level is the first entry under Values. Here, the default coding scheme defines the reference level as discussion using alphabetical order.

Logistic Regression Table shows the estimated coefficients (parameter estimates), standard error of the coefficients, z-values, and p-values. You also see the odds ratio and a 95% confidence interval for the odds ratio. The coefficient associated with a predictor is the estimated change in the logit with a one unit change in the predictor, assuming that all other factors and covariates are the same.

- If there are k response distinct values, Minitab estimates k−1 sets of parameter estimates, here labeled as Logit(1) and Logit(2). These are the estimated differences in log odds or logits of math and language arts, respectively, compared to science as the reference event. Each set contains a constant and coefficients for the factor(s), here teaching method, and the covariate(s), here age. The TeachingMethod coefficient is the estimated change in the logit when TeachingMethod is lecture compared to the teaching method being discussion, with Age held constant. The Age coefficient is the estimated change in the logit with a one year increase in age with teaching method held constant. These sets of parameter estimates gives nonparallel lines for the response values.

- The first set of estimated logits, labeled Logit(1), are the parameter estimates of the change in logits of math relative to the reference event, science. The p-values of 0.548 and 0.756 for TeachingMethod and Age, respectively, indicate that there is insufficient evidence to conclude that a change in teaching method from discussion to lecture or in age affected the choice of math as favorite subject as compared to science.

- The second set of estimated logits, labeled Logit(2), are the parameter estimates of the change in logits of language arts relative to the reference event, science. The p-values of 0.044 and 0.083 for TeachingMethod and Age, respectively, indicate that there is sufficient evidence, if the p-values are less than your acceptable α-level, to conclude that a change in teaching method from discussion to lecture or in age affected the choice of language arts as favorite subject compared to science. The positive coefficient for teaching method indicates students given a lecture style of teaching tend to prefer language arts over science compared to students given a discussion style of teaching. The estimated odds ratio of 15.96 implies that the odds of choosing language arts over science is about 16 times higher for these students when the teaching method changes from discussion to lecture. The positive coefficient associated with age indicates that students tend to like language arts over science as they become older.

Next displayed is the last Log-Likelihood from the maximum likelihood iterations along with the statistic G. G is the difference in −2 log-likelihood for a model which only has the constant terms and the fitted model shown in the **Logistic** Regression **Table**. G is the test statistic for testing the null hypothesis that all the coefficients associated with predictors equal 0 versus them not all being zero. G = 12.825 with a p-value of 0.012, indicating that at $\alpha = 0.05$, there is sufficient evidence for at least one coefficient being different from 0.

Goodness-of-Fit Tests displays Pearson and deviance goodness-of-fit tests. In our example, the p-value for the Pearson test is 0.730 and the p-value for the deviance test is 0.640, indicating that there is insufficient evidence for the model not fitting the data adequately. If the p-value is less than your selected α level, the test would indicate sufficient evidence for an inadequate fit.

Analysis of Variance

Overview

Analysis of Variance Overview

Analysis of variance (ANOVA) is similar to regression in that it is used to investigate and model the relationship between a response variable and one or more independent variables. However, analysis of variance differs from regression in two ways: the independent variables are qualitative (categorical), and no assumption is made about the nature of the relationship (that is, the model does not include coefficients for variables). In effect, analysis of variance extends the two-sample t-test for testing the equality of two population means to a more general null hypothesis of comparing the equality of more than two means, versus them not all being equal. Several of Minitab's ANOVA procedures, however, allow models with both qualitative and quantitative variables.

Minitab's ANOVA capabilities include procedures for fitting ANOVA models to data collected from a number of different designs, for fitting MANOVA models to designs with multiple response, for fitting ANOM (analysis of means) models, and graphs for testing equal variances, for confidence interval plots, and graphs of main effects and interactions.

ANOVA

Stat > ANOVA

Allows you to perform analysis of variance, test for equality of variances, and generate various plots.

Select one of the following commands:

One-Way – performs a one-way analysis of variance, with the response in one column, subscripts in another and performs multiple comparisons of means

One-Way (Unstacked) – performs a one-way analysis of variance, with each group in a separate column

Two-way – performs a two-way analysis of variance for balanced data

Analysis of Means – displays an Analysis of Means chart for normal, binomial, or Poisson data

Balanced ANOVA – analyzes balanced ANOVA models with crossed or nested and fixed or random factors

General Linear Model – analyzes balanced or unbalanced ANOVA models with crossed or nested and fixed or random factors. You can include covariates and perform multiple comparisons of means.

Fully Nested ANOVA – analyzes fully nested ANOVA models and estimates variance components

Balanced MANOVA – analyzes balanced MANOVA models with crossed or nested and fixed or random factors

General MANOVA – analyzes balanced or unbalanced MANOVA models with crossed or nested and fixed or random factors. You can also include covariates.

Test for Equal Variances – performs Bartlett's and Levene's tests for equality of variances

Interval Plot – produces graphs that show the variation of group means by plotting standard error bars or confidence intervals

Main Effects Plot – generates a plot of response main effects

Interactions Plot – generates an interaction plots (or matrix of plots)

More complex ANOVA models

Minitab offers a choice of three procedures for fitting models based upon designs more complicated than one- or two-way designs. Balanced ANOVA and General Linear Model are general procedures for fitting ANOVA models that are discussed more completely in Overview of Balanced ANOVA and GLM.

- Balanced ANOVA performs univariate (one response) analysis of variance when you have a balanced design (though one-way designs can be unbalanced). Balanced designs are ones in which all cells have the same number of observations. Factors can be crossed or nested, fixed or random. You can also use General Linear Models to analyze balanced, as well as unbalanced, designs.
- General linear model (GLM) fits the general linear model for univariate responses. In matrix form this model is Y = XB + E, where Y is the response vector, X contains the predictors, B contains parameters to be estimated, and E represents errors assumed to be normally distributed with mean vector 0 and variance σ. Using the general linear model, you can perform a univariate analysis of variance with balanced and unbalanced designs, analysis of covariance, and regression. GLM also allows you to examine differences among means using multiple comparisons.
- Fully nested ANOVA fits a fully nested (hierarchical) analysis of variance and estimates variance components. All factors are implicitly assumed to be random.

Special analytical graphs

- Test for equal variances performs Bartlett's (or F-test if 2 levels) and Levene's hypothesis tests for testing the equality or homogeneity of variances. Many statistical procedures, including ANOVA, are based upon the assumption that samples from different populations have the same variance.
- Interval plot creates a plot of means with either error bars or confidence intervals when you have a one-way design.
- Main effects plot creates a main effects plot for either raw response data or fitted values from a model-fitting procedure. The points in the plot are the means at the various levels of each factor with a reference line drawn at the grand mean of the response data. Use the main effects plot to compare magnitudes of marginal means.
- Interactions plot creates a single interaction plot if two factors are entered, or a matrix of interaction plots if 3 to 9 factors are entered. An interactions plot is a plot of means for each level of a factor with the level of a second factor held constant. Interactions plots are useful for judging the presence of interaction, which means that the difference in the response at two levels of one factor depends upon the level of another factor. Parallel lines in an interactions plot indicate no interaction. The greater the departure of the lines from being parallel, the higher the degree of interaction. To use an interactions plot, data must be available from all combinations of levels.

Use Factorial Plots generate main effects plots and interaction plots specifically for 2-level factorial designs, such as those generated by Create Factorial Design and Create RS Design.

Examples of ANOVA

Minitab Help offers examples of the following analysis of variance procedures:

One-way Analysis of Variance: Stacked Data

Two-way Analysis of Variance

Analysis of Means: Two-Way with Normal Data

Analysis of Means: Binomial response data

Analysis of Means: Poisson response data

ANOVA: Two Crossed Factors

ANOVA: Repeated Measures Design

ANOVA: Mixed Model with Restricted and Unrestricted Cases

GLM: Multiple comparisons with an unbalanced nested design

GLM: Fitting linear and quadratic effects

Fully Nested ANOVA

Balanced MANOVA

Test for Equal Variances

Interval Plot

Main Effects Plot

Interaction Plots: with Two Factors

Interaction Plots: with more than Two Factors

References for ANOVA

[1] R.E. Bechhofer and C.W. Dunnett (1988). "Percentage points of multivariate Student t distributions," *Selected Tables in Mathematical Studies*, Vol.11. American Mathematical Society.

[2] M.B. Brown and A.B. Forsythe (1974). "Robust Tests for the Equality of Variance,"*Journal of the American Statistical Association*, 69, 364–367.

[3] H.L. Harter (1970). *Order Statistics and Their Uses in Testing and Estimation*, Vol.1. U.S. Government Printing Office.

[4] A.J. Hayter (1984). "A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative," *Annals of Statistics*, 12, 61–75.

[5] D.L. Heck (1960). "Charts of Some Upper Percentage Points of the Distribution of the Largest Characteristic Root," *The Annals of Statistics*, 625–642.

[6] C.R. Hicks (1982). *Fundamental Concepts in the Design of Experiments*, Third Edition. CBC College Publishing.

[7] Y. Hochberg and A.C. Tamhane (1987). *Multiple Comparison Procedures*. John Wiley & Sons.

[8] J.C. Hsu (1984). "Constrained Two-Sided Simultaneous Confidence Intervals for Multiple Comparisons with the Best," *Annals of Statistics*, 12, 1136–1144.

[9] J.C. Hsu (1996). *Multiple Comparisons, Theory and methods.* Chapman & Hall.

[10] R. Johnson and D. Wichern (1992). *Applied Multivariate Statistical Methods*,Third Edition. Prentice Hall.

[11] H. Levene (1960). *Contributions to Probability and Statistics*. Stanford University Press, CA.

[12] T.M. Little (1981). "Interpretation and Presentation of Result," *HortScience*, 19, 637-640.

[13] G.A. Milliken and D.E. Johnson (1984). *Analysis of Messy Data,* Volume I. Van Nostrand Reinhold.

[14] D.C. Montgomery (1991). *Design and Analysis of Experiments*, Third Edition. John Wiley & Sons.

[15] D. Morrison (1967). *Multivariate Statistical Methods.* McGraw-Hill.

[16] L.S. Nelson (1974). "Factors for the Analysis of Means," *Journal of Quality Technology*, 6, 175–181.

[17] L.S. Nelson (1983). "Exact Critical Values for Use with the Analysis of Means", *Journal of Quality Technology,* 15, 40–44.

[18] P.R. Nelson (1983). "A Comparison of Sample Sizes for the Analysis of Means and the Analysis of Variance," *Journal of Quality Technology*, 15, 33–39.

[19] J. Neter, W. Wasserman and M.H. Kutner (1985). *Applied Linear Statistical Models*, Second Edition. Irwin, Inc.

[20] R.A. Olshen (1973). "The conditional level of the F-test," *Journal of the American Statistical Association*, 68, 692–698.

[21] E.R. Ott (1983). "Analysis of Means–A Graphical Procedure," *Journal of Quality Technology*, 15, 10–18.

[22] E.R. Ott and E.G. Schilling (1990). *Process Quality Control–Troubleshooting and Interpretation of Data*, 2nd Edition. McGraw-Hill.

[23] P.R. Ramig (1983). "Applications of the Analysis of Means," *Journal of Quality Technology*, 15, 19–25.

[24] E.G. Schilling (1973). "A Systematic Approach to the Analysis of Means," *Journal of Quality Technology*, 5, 93–108, 147–159.

[25] S.R. Searle, G. Casella, and C.E. McCulloch (1992). *Variance Components.* John Wiley & Sons.

[26] N.R. Ullman (1989). "The Analysis of Means (ANOM) for Signal and Noise," *Journal of Quality Technology*, 21, 111–127.

[27] E. Uusipaikka (1985). "Exact simultaneous confidence intervals for multiple comparisons among three or four mean values," *Journal of the American Statistical Association*, 80, 196–201.

[28] B.J. Winer (1971). *Statistical Principles in Experimental Design*, Second Edition. McGraw-Hill.

Acknowledgment

We are grateful for assistance in the design and implementation of multiple comparisons from Jason C. Hsu, Department of Statistics, Ohio State University and for the guidance of James L. Rosenberger, Statistics Department, The Pennsylvania State University, in developing the Balanced ANOVA, Analysis of Covariance, and General Linear Models procedures.

One-Way

One-way and two-way ANOVA models

- One-way analysis of variance tests the equality of population means when classification is by one variable. The classification variable, or factor, usually has three or more levels (one-way ANOVA with two levels is equivalent to a t-test), where the **level** represents the treatment applied. For example, if you conduct an experiment where you measure durability of a product made by one of three methods, these methods constitute the levels. The one-way procedure also allows you to examine differences among means using multiple comparisons.
- Two-way analysis of variance performs an analysis of variance for testing the equality of populations means when classification of treatments is by two variables or factors. In two-way ANOVA, the data must be balanced (all cells must have the same number of observations) and factors must be fixed.

  If you wish to specify certain factors to be random, use Balanced ANOVA if your data are balanced; use General Linear Models if your data are unbalanced or if you wish to compare means using multiple comparisons.

One-Way Analysis of Variance

Stat > ANOVA > One-way

Performs a one-way analysis of variance, with the dependent variable in one column, subscripts in another. If each group is entered in its own column, use Stat > ANOVA > One-Way (Unstacked).

You can also perform multiple comparisons and display graphs of your data.

Dialog box items

Response: Enter the column containing the response.

Factor: Enter the column containing the factor levels.

Store Residuals: Check to store residuals in the next available column.

Store fits: Check to store the fitted values in the next available column.

Confidence level: Enter the confidence level. For example, enter 90 for 90%. The default is 95%.

<Comparisons>

<Graphs>

Data – One-Way with Stacked Data

The response variable must be numeric. Stack the response data in one column with another column of level values identifying the population (stacked case). The factor level (group) column can be numeric, text, or date/time. If you wish to change the order in which text categories are processed from their default alphabetical order, you can define your own order. See Ordering Text Categories. You do not need to have the same number of observations in each level. You can use Make Patterned Data to enter repeated factor levels.

Note If your response data are entered in separate worksheet columns, use Stat > ANOVA > One-Way (Unstacked).

To perform a one-way analysis of variance with stacked data

1 Choose **Stat > ANOVA > One-Way**.

2 In **Response**, enter the column containing the response.

3 In **Factor**, enter the column containing the factor levels.

4 If you like, use any dialog box options, then click **OK**.

Discussion of Multiple Comparisons

The multiple comparisons are presented as a set of confidence intervals, rather than as a set of hypothesis tests. This allows you to assess the practical significance of differences among means, in addition to statistical significance. As usual, the null hypothesis of no difference between means is rejected if and only if zero is not contained in the confidence interval.

The selection of the appropriate multiple comparison method depends on the desired inference. It is inefficient to use the Tukey all-pairwise approach when Dunnett or MCB is suitable, because the Tukey confidence intervals will be wider and the hypothesis tests less powerful for a given family error rate. For the same reasons, MCB is superior to Dunnett if you want to eliminate levels that are not the best and to identify those that are best or close to the best. The choice of Tukey versus Fisher methods depends on which error rate, family or individual, you wish to specify.

Individual error rates are exact in all cases. Family error rates are exact for equal group sizes. If group sizes are unequal, the true family error rate for Tukey, Fisher, and MCB will be slightly smaller than stated, resulting in conservative confidence intervals [4,22]. The Dunnett family error rates are exact for unequal sample sizes.

The results of the one-way F-test and multiple comparisons can conflict. For example, it is possible for the F-test to reject the null hypothesis of no differences among the level means, and yet all the Tukey pairwise confidence intervals contain zero. Conversely, it is possible for the F-test to fail to reject, and yet have one or more of the Tukey pairwise confidence intervals not include zero. The F-test has been used to protect against the occurrence of false positive differences in means. However, Tukey, Dunnett, and MCB have protection against false positives built in, while Fisher only benefits from this protection when all means are equal. If the use of multiple comparisons is conditioned upon the significance of the F-test, the error rate can be higher than the error rate in the unconditioned application of multiple comparisons [15].

Comparisons – One-Way Multiple Comparisons with Stacked Data

Stat > ANOVA > One-Way > Comparisons

Provides confidence intervals for the differences between means, using four different methods: Tukey's, Fisher's, Dunnett's, and Hsu's MCB. Tukey and Fisher provide confidence intervals for all pairwise differences between level means. Dunnett provides a confidence interval for the difference between each treatment mean and a control mean. Hsu's MCB provides a confidence interval for the difference between each level mean and the best of the other level means. Tukey, Dunnett and Hsu's MCB tests use a family error rate, whereas Fisher's LSD procedure uses an individual error rate.

Which multiple comparison test to use depends on the desired inference. It is inefficient to use the Tukey all-pairwise approach when Dunnett or Hsu's MCB is suitable, because the Tukey confidence intervals will be wider and the hypothesis tests less powerful for a given family error rate. For the same reasons, Hsu's MCB is superior to Dunnett if you want to eliminate levels that are not the best and to identify those that are best or close to the best. The choice of Tukey versus Fisher depends on which error rate, family or individual, you wish to specify.

Dialog box items

Tukey's, family error rate: Check to obtain confidence intervals for all pairwise differences between level means using Tukey's method (also called Tukey-Kramer in the unbalanced case), and then enter a family error rate between 0.5 and 0.001. Values greater than or equal to 1.0 are interpreted as percentages. The default error rate is 0.05.

Fisher's, individual error rate: Check to obtain confidence intervals for all pairwise differences between level means using Fisher's LSD procedure, and then enter an individual rate between 0.5 and 0.001. Values greater than or equal to 1.0 are interpreted as percentages. The default error rate is 0.05.

Dunnett's family error rate: Check to obtain a two-sided confidence interval for the difference between each treatment mean and a control mean, and then enter a family error rate between 0.5 and 0.001. Values greater than or equal to 1.0 are interpreted as percentages. The default error rate is 0.05.

Control group level: Enter the value for the control group factor level. (IMPORTANT: For text variables, you must enclose factor levels in double quotes, even if there are no spaces in them.)

Hsu's MCB, family error rate: Check to obtain a confidence interval for the difference between each level mean and the best of the other level means [9]. There are two choices for "best." If the smallest mean is considered the best, set K = –1; if the largest is considered the best, set K = 1. Specify a family error rate between 0.5 and 0.001. Values greater than or equal to 1.0 are interpreted as percentages. The default error rate is 0.05.

Largest is best: Choose to have the largest mean considered the best.

Smallest is best: Choose to have the smallest mean considered the best.

One-Way Analysis of Variance – Graphs

Stat > ANOVA > One-way > Graphs

Displays an individual value plot, a boxplot, and residual plots. You do not have to store the residuals in order to produce the residual plots.

Dialog box items

Individual value plot: Check to display an individual value plot of each sample.

Boxplots of data: Check to display a boxplot of each sample.

Residual Plots

Individual plots: Choose to display one or more plots.

Histogram of residuals: Check to display a histogram of the residuals.

Normal plot of residuals: Check to display a normal probability plot of the residuals.

Residuals versus fits: Check to plot the residuals versus the fitted values.

Residuals versus order: Check to plot the residuals versus the order of the data. The row number for each data point is shown on the x-axis–for example, 1 2 3 4... n.

Four in one: Choose to display a layout of a histogram of residuals, normal plot of residuals, plot of residuals versus fits, and plot of residuals versus order.

Residuals versus the variables: Enter one or more columns containing the variables against which you want to plot the residuals. Minitab displays a separate graph for each column.

Example of a one-way analysis of variance with multiple comparisons

You design an experiment to assess the durability of four experimental carpet products. You place a sample of each of the carpet products in four homes and you measure durability after 60 days. Because you wish to test the equality of means and to assess the differences in means, you use the one-way ANOVA procedure (data in stacked form) with multiple comparisons. Generally, you would choose one multiple comparison method as appropriate for your data. However, two methods are selected here to demonstrate Minitab's capabilities.

1. Open the worksheet EXH_AOV.MTW.
2. Choose **Stat > ANOVA > One-Way**.
3. In **Response**, enter *Durability*. In **Factor**, enter *Carpet*.
4. Click **Comparisons**. Check **Tukey's, family error rate**. Check **Hsu's MCB, family error rate** and enter *10*.
5. Click **OK** in each dialog box.

Session window output

One-way ANOVA: Durability versus Carpet

```
Source  DF     SS    MS     F      P
Carpet   3  146.4  48.8  3.58  0.047
Error   12  163.5  13.6
Total   15  309.9

S = 3.691   R-Sq = 47.24%   R-Sq(adj) = 34.05%


                         Individual 95% CIs For Mean Based on
                         Pooled StDev
Level  N    Mean  StDev  ---------+---------+---------+---------+
1      4  14.483  3.157            (-------*-------)
2      4   9.735  3.566  (-------*--------)
3      4  12.808  1.506         (-------*-------)
4      4  18.115  5.435                   (-------*-------)
                         ---------+---------+---------+---------+
                               10.0      15.0      20.0      25.0

Pooled StDev = 3.691


Hsu's MCB (Multiple Comparisons with the Best)

Family error rate = 0.1
Critical value = 1.87

Intervals for level mean minus largest of other level means

Level    Lower  Center  Upper    --+---------+---------+---------+-------
1       -8.511  -3.632  1.246            (-------*-------)
2      -13.258  -8.380  0.000    (-------*-------------)
3      -10.186  -5.308  0.000         (-------*--------)
4       -1.246   3.632  8.511                        (-------*-------)
                                 --+---------+---------+---------+-------
                               -12.0      -6.0       0.0       6.0


Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of Carpet

Individual confidence level = 98.83%


Carpet = 1 subtracted from:

Carpet    Lower  Center   Upper  ------+---------+---------+---------+---
2       -12.498  -4.748   3.003      (------*-------)
3        -9.426  -1.675   6.076         (------*-------)
4        -4.118   3.632  11.383              (-------*------)

                                 ------+---------+---------+---------+---
                                     -10         0        10        20


Carpet = 2 subtracted from:

Carpet   Lower  Center   Upper  ------+---------+---------+---------+---
3       -4.678   3.073  10.823             (-------*-------)
4        0.629   8.380  16.131                   (------*-------)
                                ------+---------+---------+---------+---
                                    -10         0        10        20


Carpet = 3 subtracted from:

Carpet   Lower  Center   Upper  ------+---------+---------+---------+---
```

```
4       -2.443   5.308  13.058                (------*-------)
                                ------+---------+---------+---------+---
                                    -10         0        10        20
```

Interpreting the results

In the ANOVA table, the p-value (0.047) for Carpet indicates that there is sufficient evidence that not all the means are equal when alpha is set at 0.05. To explore the differences among the means, examine the multiple comparison results.

Hsu's MCB comparisons

Hsu's MCB (Multiple Comparisons with the Best) compares each mean with the best (largest) of the other means. Minitab compares the means of carpets 1, 2, and 3 to the carpet 4 mean because it is the largest. Carpet 1 or 4 may be best because the corresponding confidence intervals contain positive values. No evidence exists that carpet 2 or 3 is the best because the upper interval endpoints are 0, the smallest possible value.

Note You can describe the potential advantage or disadvantage of any of the contenders for the best by examining the upper and lower confidence intervals. For example, if carpet 1 is best, it is no more than 1.246 better than its closest competitor, and it may be as much as 8.511 worse than the best of the other level means.

Tukey's comparisons

Tukey's test provides 3 sets of multiple comparison confidence intervals:

- Carpet 1 mean subtracted from the carpet 2, 3, and 4 means: The first interval in the first set of the Tukey's output (–12.498, –4.748, 3.003) gives the confidence interval for the carpet 1 mean subtracted from the carpet 2 mean. You can easily find confidence intervals for entries not included in the output by reversing both the order and the sign of the interval values. For example, the confidence interval for the mean of carpet 1 minus the mean of carpet 2 is (–3.003, 4.748, 12.498). For this set of comparisons, none of the means are statistically different because all of the confidence intervals include 0.
- Carpet 2 mean subtracted from the carpet 3 and 4 means: The means for carpets 2 and 4 are statistically different because the confidence interval for this combination of means (0.629, 8.380, 16.131) excludes zero.
- Carpet 3 mean subtracted from the carpet 4 mean: Carpets 3 and 4 are not statistically different because the confidence interval includes 0.

By not conditioning upon the F-test, differences in treatment means appear to have occurred at family error rates of 0.10. If Hsu's MCB method is a good choice for these data, carpets 2 and 3 might be eliminated as a choice for the best. When you use Tukey's method, the mean durability for carpets 2 and 4 appears to be different.

One-Way (Unstacked)

One-way and two-way ANOVA models

- One-way analysis of variance tests the equality of population means when classification is by one variable. The classification variable, or factor, usually has three or more levels (one-way ANOVA with two levels is equivalent to a t-test), where the **level** represents the treatment applied. For example, if you conduct an experiment where you measure durability of a product made by one of three methods, these methods constitute the levels. The one-way procedure also allows you to examine differences among means using multiple comparisons.
- Two-way analysis of variance performs an analysis of variance for testing the equality of populations means when classification of treatments is by two variables or factors. In two-way ANOVA, the data must be balanced (all cells must have the same number of observations) and factors must be fixed.

  If you wish to specify certain factors to be random, use Balanced ANOVA if your data are balanced; use General Linear Models if your data are unbalanced or if you wish to compare means using multiple comparisons.

One-Way Analysis of Variance (Unstacked)

Stat > ANOVA > One-Way (Unstacked)

Performs a one-way analysis of variance, with each group in a separate column. If your response data are stacked in one column with another column of level values identifying the population, use Stat > ANOVA > One-Way.

You can also perform multiple comparisons and display graphs of your data.

Dialog box items

Responses [in separate columns]: Enter the columns containing the separate response variables.

Store residuals: Check to store residuals in the next available columns. The number of residual columns will match the number of response columns.

Store fits: Check to store the fitted values in the next available column.

Confidence level: Enter the confidence level. For example, enter 90 for 90%. The default is 95%.

<Comparisons>

<Graphs>

Data – One-Way (Unstacked)

The response variable must be numeric. Enter the sample data from each population into separate columns of your worksheet.

Note If your response data are stacked in one column with another column of level values identifying the population, use Stat > ANOVA > One-Way.

To perform a one-way analysis of variance with unstacked data

1 Choose **Stat > ANOVA > One-Way (Unstacked)**.

2 In **Responses (in separate columns)**, enter the columns containing the separate response variables.

3 If you like, use any dialog box options, then click **OK**.

Comparisons – One-Way Multiple Comparisons with Unstacked Data

Stat > ANOVA > One-Way (Unstacked) > Comparisons

Use to generate confidence intervals for the differences between means, using four different methods: Tukey's, Fisher's, Dunnett's, and Hsu's MCB. Tukey's and Fisher's methods provide confidence intervals for all pairwise differences between level means. Dunnett's method provides a confidence interval for the difference between each treatment mean and a control mean. Hsu's MCB method provides a confidence interval for the difference between each level mean and the best of the other level means. Tukey's, Dunnett's, and Hsu's MCB tests use a family error rate, whereas Fisher's LSD procedure uses an individual error rate.

Which multiple comparison test to use depends on the desired inference. Using Tukey's all-pairwise approach is inefficient when Dunnett's or Hsu's MCB is suitable, because Tukey's confidence intervals are wider and the hypothesis tests less powerful for a given family error rate. For the same reasons, Hsu's MCB is superior to Dunnett's if you want to eliminate levels that are not the best and to identify those that are best or close to the best. The choice of Tukey's versus Fisher's depends on which error rate, family or individual, you wish to specify.

Dialog box items

Tukey's, family error rate: Check to obtain confidence intervals for all pairwise differences between level means using Tukey's method (also called Tukey-Kramer in the unbalanced case), then enter a family error rate between 0.5 and 0.001. Values greater than or equal to 1.0 are interpreted as percentages. The default error rate is 0.05.

Fisher's, individual error rate: Check to obtain confidence intervals for all pairwise differences between level means using Fisher's LSD procedure, then enter an individual rate between 0.5 and 0.001. Values greater than or equal to 1.0 are interpreted as percentages. The default error rate is 0.05.

Dunnett's family error rate: Check to obtain a two-sided confidence interval for the difference between each treatment mean and a control mean, then enter a family error rate between 0.5 and 0.001. Values greater than or equal to 1.0 are interpreted as percentages. The default error rate is 0.05.

Control group level: Enter the column with the control group data.

Hsu's MCB, family error rate: Check to obtain a confidence interval for the difference between each level mean and the best of the other level means [9]. There are two choices for "best." If the smallest mean is considered the best, set K = –1; if the largest is considered the best, set K = 1. Specify a family error rate between 0.5 and 0.001. Values greater than or equal to 1.0 are interpreted as percentages. The default error rate is 0.05.

Largest is best: Choose to have the largest mean considered the best.

Smallest is best: Choose to have the smallest mean considered the best.

One-Way Analysis of Variance – Graphs

Stat > ANOVA > One-Way (Unstacked) > Graphs

Displays individual value plots and boxplots for each sample and residual plots.

Dialog box items

Individual value plot: Check to display an individual value plot of each sample. The sample mean is shown on each dotplot.

Boxplots of data: Check to display a boxplot of each sample. The sample mean is shown on each boxplot.

Residual Plots

Individual plots: Choose to display one or more plots.

Histogram of residuals: Check to display a histogram of the residuals.

Normal plot of residuals: Check to display a normal probability plot of the residuals.

Residuals versus fits: Check to plot the residuals versus the fitted values.

Three in one: Choose to display a layout of a histogram of residuals, normal plot of residuals, and a plot of residuals versus fits.

Discussion of Multiple Comparisons

The multiple comparisons are presented as a set of confidence intervals, rather than as a set of hypothesis tests. This allows you to assess the practical significance of differences among means, in addition to statistical significance. As usual, the null hypothesis of no difference between means is rejected if and only if zero is not contained in the confidence interval.

The selection of the appropriate multiple comparison method depends on the desired inference. It is inefficient to use the Tukey all-pairwise approach when Dunnett or MCB is suitable, because the Tukey confidence intervals will be wider and the hypothesis tests less powerful for a given family error rate. For the same reasons, MCB is superior to Dunnett if you want to eliminate levels that are not the best and to identify those that are best or close to the best. The choice of Tukey versus Fisher methods depends on which error rate, family or individual, you wish to specify.

Individual error rates are exact in all cases. Family error rates are exact for equal group sizes. If group sizes are unequal, the true family error rate for Tukey, Fisher, and MCB will be slightly smaller than stated, resulting in conservative confidence intervals [4,22]. The Dunnett family error rates are exact for unequal sample sizes.

The results of the one-way F-test and multiple comparisons can conflict. For example, it is possible for the F-test to reject the null hypothesis of no differences among the level means, and yet all the Tukey pairwise confidence intervals contain zero. Conversely, it is possible for the F-test to fail to reject, and yet have one or more of the Tukey pairwise confidence intervals not include zero. The F-test has been used to protect against the occurrence of false positive differences in means. However, Tukey, Dunnett, and MCB have protection against false positives built in, while Fisher only benefits from this protection when all means are equal. If the use of multiple comparisons is conditioned upon the significance of the F-test, the error rate can be higher than the error rate in the unconditioned application of multiple comparisons [15].

Two-Way

One-way and two-way ANOVA models

- One-way analysis of variance tests the equality of population means when classification is by one variable. The classification variable, or factor, usually has three or more levels (one-way ANOVA with two levels is equivalent to a t-test), where the **level** represents the treatment applied. For example, if you conduct an experiment where you measure durability of a product made by one of three methods, these methods constitute the levels. The one-way procedure also allows you to examine differences among means using multiple comparisons.
- Two-way analysis of variance performs an analysis of variance for testing the equality of populations means when classification of treatments is by two variables or factors. In two-way ANOVA, the data must be balanced (all cells must have the same number of observations) and factors must be fixed.

  If you wish to specify certain factors to be random, use Balanced ANOVA if your data are balanced; use General Linear Models if your data are unbalanced or if you wish to compare means using multiple comparisons.

Two-Way Analysis of Variance

Stat > ANOVA > Two-Way

A two-way analysis of variance tests the equality of populations means when classification of treatments is by two variables or factors. For this procedure, the data must be balanced (all cells must have the same number of observations) and factors must be fixed.

To display cell means and standard deviations, use Cross Tabulation and Chi-Square.

If you wish to specify certain factors to be random, use Balanced ANOVA if your data are balanced. Use General Linear Model if your data are unbalanced or if you wish to compare means using multiple comparisons.

Dialog box items

Response: Enter the column containing the response variable.

Row Factor: Enter one of the factor level columns.

Display means: Check to compute marginal means and confidence intervals for each level of the row factor.

Column factor: Enter the other factor level column.

Display means: Check to compute marginal means and confidence intervals for each level of the column factor.

Store residuals: Check to store the residuals.

Store fits: Check to store the fitted value for each group.

Confidence level: Enter the level for the confidence intervals for the individual means. For example, enter 90 for 90%. The default is 95%.

Fit additive model: Check to fit a model without an interaction term. In this case, the fitted value for cell (i,j) is (mean of observations in row i) + (mean of observations in row j) – (mean of all observations).

<Graphs>

Data – Two-Way

The response variable must be numeric and in one worksheet column. You must have a single factor level column for each of the two factors. These can be numeric, text, or date/time. If you wish to change the order in which text categories are processed from their default alphabetical order, you can define your own order. See Ordering Text Categories. You must have a balanced design (same number of observations in each treatment combination) with fixed factors. You can use Make Patterned Data to enter repeated factor levels.

To perform a two-way analysis of variance

1. Choose **Stat > ANOVA > Two-Way**.
2. In **Response**, enter the column containing the response variable.
3. In **Row Factor**, enter one of the factor level columns.
4. In **Column Factor**, enter the other factor level column.
5. If you like, use any dialog box options, then click **OK**.

Two-Way Analysis of Variance – Graphs

Stat > ANOVA > Two-Way > Graphs

Displays residual plots. You do not have to store the residuals in order to produce these plots.

Dialog box items

Individual value plot: Check to display an individual value plot of each sample.

Boxplots of data: Check to display a boxplot of each sample.

Residual Plots

Individual plots: Choose to display one or more plots.

Histogram of residuals: Check to display a histogram of the residuals.

Normal plot of residuals: Check to display a normal probability plot of the residuals.

Residuals versus fits: Check to plot the residuals versus the fitted values.

Residuals versus order: Check to plot the residuals versus the order of the data. The row number for each data point is shown on the x-axis – for example, 1 2 3 4... n.

Four in one: Choose to display a layout of a histogram of residuals, a normal plot of residuals, a plot of residuals versus fits, and a plot of residuals versus order.

Residuals versus the variables: Enter one or more columns containing the variables against which you want to plot the residuals. Minitab displays a separate graph for each column.

Example of a Two-Way Analysis of Variance

You as a biologist are studying how zooplankton live in two lakes. You set up twelve tanks in your laboratory, six each with water from one of the two lakes. You add one of three nutrient supplements to each tank and after 30 days you count the zooplankton in a unit volume of water. You use two-way ANOVA to test if the population means are equal, or equivalently, to test whether there is significant evidence of interactions and main effects.

1 Open the worksheet EXH_AOV.MTW.

2 Choose **Stat > ANOVA > Two-Way**.

3 In **Response**, enter *Zooplankton*.

4 In **Row factor**, enter *Supplement*. Check **Display means**.

5 In **Column factor**, enter *Lake*. Check **Display means**. Click **OK**.

Session window output

Two-way ANOVA: Zooplankton versus Supplement, Lake

```
Source        DF       SS       MS     F      P
Supplement     2  1918.50  959.250  9.25  0.015
Lake           1    21.33   21.333  0.21  0.666
Interaction    2   561.17  280.583  2.71  0.145
Error          6   622.00  103.667
Total         11  3123.00

S = 10.18   R-Sq = 80.08%   R-Sq(adj) = 63.49%


                   Individual 95% CIs For Mean Based on
                   Pooled StDev
Supplement   Mean  --+---------+---------+---------+-------
1           43.50     (-------*-------)
2           68.25                     (--------*-------)
3           39.75  (--------*-------)
                   --+---------+---------+---------+-------
                    30        45        60        75


                   Individual 95% CIs For Mean Based on
                   Pooled StDev
Lake         Mean  -----+---------+---------+---------+----
Dennison  51.8333      (----------------*----------------)
Rose      49.1667  (----------------*----------------)
                   -----+---------+---------+---------+----
                     42.0      48.0      54.0      60.0
```

Interpreting the results

The default output for two-way ANOVA is the analysis of variance table. For the zooplankton data, there is no significant evidence for a supplement∗lake interaction effect or a lake main effect if your acceptable a value is less than 0.145 (the p-value for the interaction F-test). There is significant evidence for supplement main effects, as the F-test p-value is 0.015.

As requested, the means are displayed with individual 95% confidence intervals. Supplement 2 appears to have provided superior plankton growth in this experiment. These are t-distribution confidence intervals calculated using the error degrees of freedom and the pooled standard deviation (square root of the mean square error). If you want to examine simultaneous differences among means using multiple comparisons, use General Linear Model.

Analysis of Means

Overview of Analysis of Means

Analysis of Means (ANOM) is a graphical analog to ANOVA, and tests the equality of population means. ANOM [16] was developed to test main effects from a designed experiment in which all factors are fixed. This procedure is used for one-factor designs. Minitab uses an extension of ANOM or ANalysis Of Mean treatment Effects (ANOME) [24] to test the significance of mean treatment effects for two-factor designs.

An ANOM chart can be described in two ways: by its appearance and by its function. In appearance, it resembles a Shewhart control chart. In function, it is similar to ANOVA for detecting differences in population means [13]. The null hypotheses for ANOM and ANOVA are the same: both methods test for a lack of homogeneity among means. However, the alternative hypotheses are different [16]. The alternative hypothesis for ANOM is that one of the population means is different from the other means, which are equal. The alternative hypothesis for ANOVA is that the variability among population means is greater than zero.

For most cases, ANOVA and ANOM will likely give similar results. However, there are some scenarios where the two methods might be expected to differ:

- If one group of means is above the grand mean and another group of means is below the grand mean, the F-test for ANOVA might indicate evidence for differences where ANOM might not.
- If the mean of one group is separated from the other means, the ANOVA F-test might not indicate evidence for differences whereas ANOM might flag this group as being different from the grand mean.

Refer to [21], [22], [23], and [24] for an introduction to the analysis of means.

ANOM can be used if you assume that the response follows a normal distribution, similar to ANOVA, and the design is one-way or two-way. You can also use ANOM when the response follows either a binomial distribution or a Poisson distribution.

Analysis of Means

Stat > ANOVA > Analysis of Means

Draws an Analysis of Means chart (ANOM) for normal, binomial, and Poisson data and optionally prints a summary table for normal and binomial data.

Dialog box items

Response: Enter the column containing the response variable. The meaning of and limitations for the response variable vary depending on whether your data follow a normal, binomial, or Poisson distribution. See Data for Analysis of Means for more information.

Distribution of Data:

Normal: Choose if the response data follow a normal distribution (measurement data).

Factor 1: Enter the column containing the levels for the first factor. If you enter a single factor, Analysis of Means produces a single plot showing the means for each level of the factor.

Factor 2 (optional): Enter the column containing the levels for the second factor. If you enter two factors, Analysis of Means produces three plots–one showing the interaction effects, one showing the main effects for the first factor, and one showing the main effects for the second factor.

Binomial: Choose if the response data follow a binomial distribution. The sample size must be constant (balanced design), and the sample size must be large enough to ensure that the normal approximation to the binomial is valid. This usually implies that $np > 5$ and $n(1–p) > 5$, where p is the proportion of defects.

Sample size: Enter a number or a stored constant to specify the sample size.

Poisson: Choose if the response data follow a Poisson distribution. The Poisson distribution can be adequately approximated by a normal distribution if the mean of the Poisson distribution is at least five. Therefore, when the Poisson mean is large enough, you can test for equality of population means using this procedure.

Alpha level: Enter a value for the error rate, or alpha-level. The number you enter must be between 0 and 1. The decision lines on the ANOM chart are based on an experiment-wide error rate, similar to what you might use when making pairwise comparisons or contrasts in an ANOVA.

Title: Type a new title to replace the plot's default title.

Data – Analysis of Means (normal)

Your response data may be numeric or date/time and must be entered into one column. Factor columns may be numeric, text, or date/time and may contain any values. The response and factor columns must be of the same length. Minitab's capability to enter patterned data can be helpful in entering numeric factor levels; see Make Patterned Data to enter repeated factor levels. If you wish to change the order in which text categories are processed from their default alphabetical order, you can define your own order; see Ordering Text Categories.

One-way designs may be balanced or unbalanced and can have up to 100 levels. Two-way designs must be balanced and can have up to 50 levels for each factor. All factors must be fixed.

Rows with missing data are automatically omitted from calculations. If you have two factors, the design must be balanced after omitting rows with missing values.

Data – Analysis of Means (binomial)

The response data are the numbers of defectives (or defects) found in each sample, with a maximum of 500 samples. You must enter these data in one column.

Because the decision limits in the ANOM chart are based upon the normal distribution, one of the assumptions that must be met when the response data are binomial is that the sample size is large enough to ensure that the normal approximation to the binomial is valid. A general rule of thumb is to only use ANOM if $np > 5$ and $n(1 - p) > 5$, where n is the sample size and p is the proportion of defectives. The second assumption is that all of the samples are the same size. See [24] for more details.

A sample with a missing response value (∗) is automatically omitted from the analysis.

Data – Analysis of Means (Poisson)

The response data are the numbers of defects found in each sample. You can have up to 500 samples.

The Poisson distribution can be adequately approximated by a normal distribution if the mean of the Poisson distribution is at least 5. When the Poisson mean is large enough, you can apply analysis of means to data from a Poisson distribution to test if the population means are equal to the grand mean.

A sample with a missing response value (∗) is automatically omitted from the analysis.

To perform an analysis of means

1 Choose **Stat > ANOVA > Analysis of Means**.

2 In **Response**, enter a numeric column containing the response variable.

3 Under **Distribution of Data**, choose **Normal**, **Binomial**, or **Poisson**.

- If you choose **Normal**, you can analyze either a one-way or two-way design. For a one-way design, enter the column containing the factor levels in **Factor 1**. For a two-way design, enter the columns containing the factor levels in **Factor 1** and **Factor 2**.
- If you choose **Binomial**, type a number in **Sample size**.

4 If you like, use one or more of the dialog box options, then click **OK**.

Example of a two-way analysis of means

You perform an experiment to assess the effect of three process time levels and three strength levels on density. You use analysis of means for normal data and a two-way design to identify any significant interactions or main effects.

1 Open the worksheet EXH_AOV.MTW.

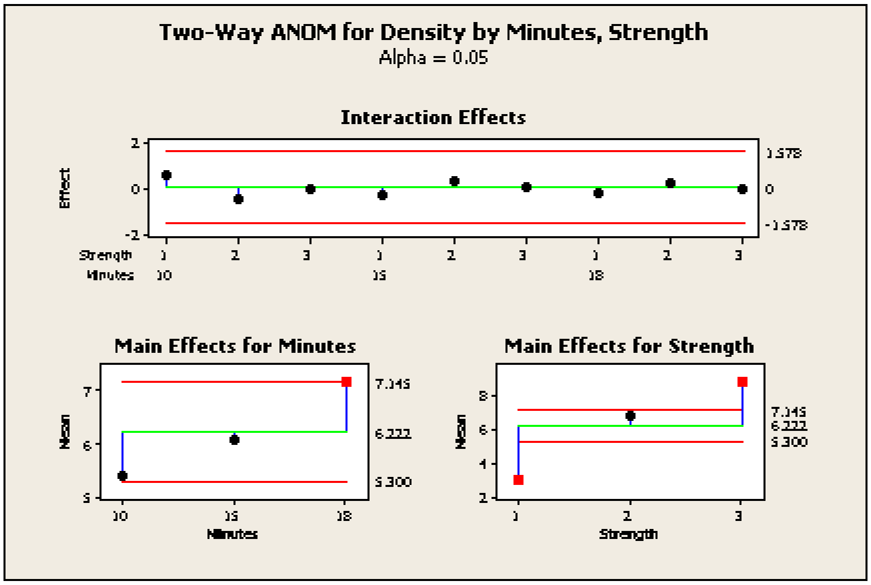
2 Choose **Stat > ANOVA > Analysis of Means**.

3 In **Response**, enter *Density*.

4 Choose **Normal**.

5 In **Factor 1**, enter *Minutes*. In **Factor 2**, enter *Strength*. Click **OK**.

Graph window output



Interpreting the results

Minitab displays three plots with a two-way ANOM to show the interaction effects, main effects for the first factor, and main effects for the second factor. ANOM plots have a center line and decision limits. If a point falls outside the decision limits, then there is significant evidence that the mean represented by that point is different from the grand mean. With a two-way ANOM, look at the interaction effects first. If there is significant evidence for interaction, it usually does not make sense to consider main effects, because the effect of one factor depends upon the level of the other.

In this example, the interaction effects are well within the decision limits, signifying no evidence of interaction. Now you can look at the main effects. The lower two plots show the means for the levels of the two factors, with the main effect being the difference between the mean and the center line. The point representing the level 3 mean of the factor Minutes is displayed by a red asterisk, which indicates that there is significant evidence that the level 3 mean is different from the grand mean at $\alpha = 0.05$. You may wish to investigate any point near or above the decision limits. The main effects for levels 1 and 3 of factor Strength are well outside the decision limits of the lower left plot, signifying that there is evidence that these means are different from the grand mean at $\alpha = 0.05$.

Example of analysis of means for binomial response data

You count the number of rejected welds from samples of size 80 in order to identify samples whose proportions of rejects are out of line with the other samples. Because the data are binomial (two possible outcomes, constant proportion of success, and independent samples) you use analysis of means for binomial data.

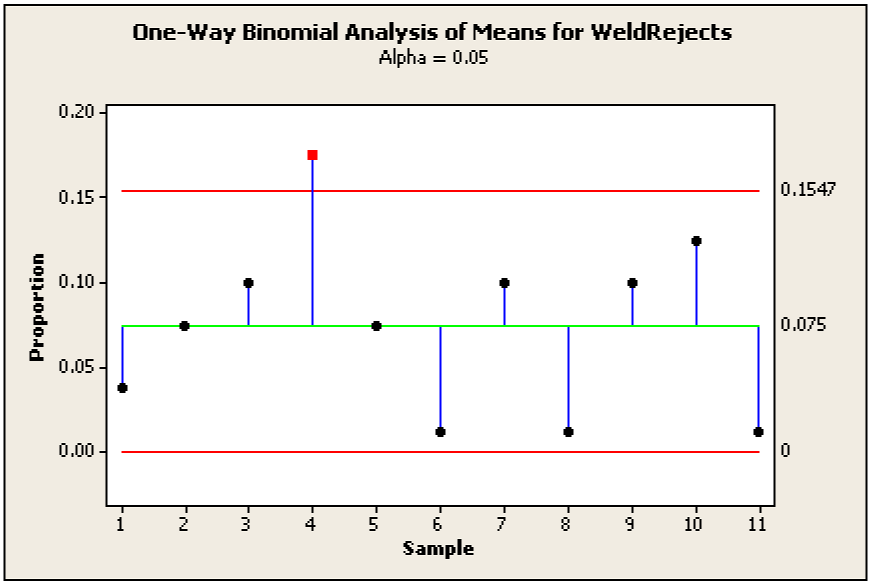
1 Open the worksheet EXH_AOV.MTW.

2 Choose **Stat > ANOVA > Analysis of Means**.

3 In **Response**, enter *WeldRejects*.

4 Choose **Binomial** and type *80* in **Sample size**. Click **OK.**

Graph window output



Interpreting the results

The plot displays the proportion of defects for each sample, a center line representing the average proportion, and upper and lower decision limits. If the point representing a sample falls outside the decision limits, there is significant evidence that the sample mean is different from the average

In this example, the proportion of defective welds in sample four is identified as unusually high because the point representing this sample falls outside the decision limits.

Example of analysis of means for Poisson response data

As production manager of a toy manufacturing company, you want to monitor the number of defects per sample of motorized toy cars. You monitor 20 samples of toy cars and create an analysis of means chart to examine the number of defects in each sample.

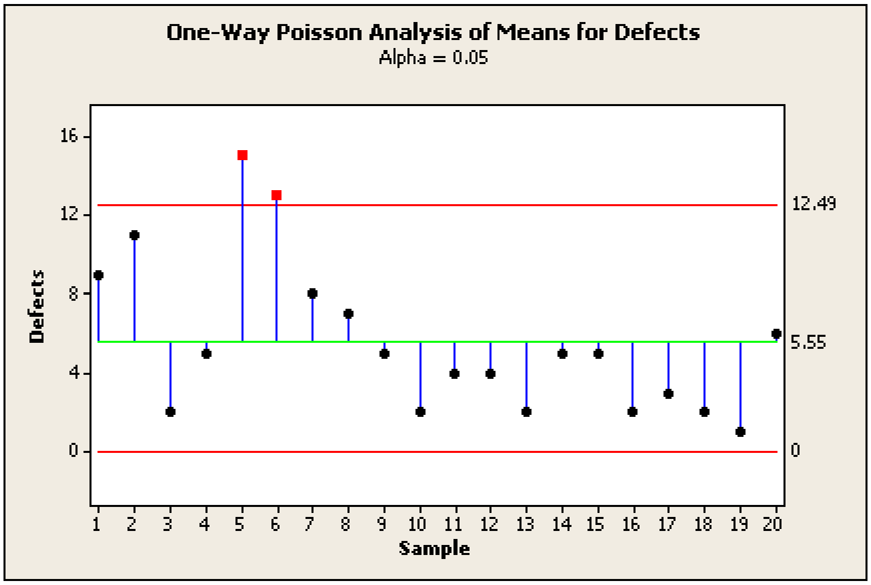
1 Open the worksheet TOYS.MTW.

2 Choose **Stat > ANOVA > Analysis of Means**.

3 In **Response**, enter *Defects*.

4 Choose **Poisson**, then click **OK**.

Graph window output



Interpreting the results

The plot displays the number of defects for each sample, a center line representing the average number of defects, and upper and lower decision limits. If the point representing a sample falls outside the decision limits, there is significant evidence exists that the sample mean is different from the average

In this example, the number of defective motorized toy cars in samples five and six is identified as being unusually high because the points representing these samples fall outside the decision limits.

Balanced ANOVA

Overview of Balanced ANOVA and GLM

Balanced ANOVA and general linear model (GLM) are ANOVA procedures for analyzing data collected with many different experimental designs. Your choice between these procedures depends upon the experimental design and the available options. The experimental design refers to the selection of units or subjects to measure, the assignment of treatments to these units or subjects, and the sequence of measurements taken on the units or subjects. Both procedures can fit univariate models to balanced data with up to 31 factors. Here are some of the other options:

	Balanced ANOVA	**GLM**
Can fit unbalanced data	no	yes
Can specify factors as random and obtain expected means squares	yes	yes
Fits covariates	no	yes
Performs multiple comparisons	no	yes
Fits restricted/unrestricted forms of mixed model	yes	unrestricted only

You can use balanced ANOVA to analyze data from balanced designs. See Balanced designs. You can use GLM to analyze data from any balanced design, though you cannot choose to fit the restricted case of the mixed model, which only balanced ANOVA can fit. See Restricted and unrestricted form of mixed models.

To classify your variables, determine if your factors are:

- crossed or nested
- fixed or random
- covariates

For information on how to specify the model, see Specifying the model terms, Specifying terms involving covariates, Specifying reduced models, and Specifying models for some specialized designs.

For easy entering of repeated factor levels into your worksheet, see Using patterned data to set up factor levels.

Balanced Analysis of Variance

Stat > ANOVA > Balanced ANOVA

Use Balanced ANOVA to perform univariate analysis of variance for each response variable.

Your design must be balanced, with the exception of one-way designs. *Balanced* means that all treatment combinations (cells) must have the same number of observations. See Balanced designs. Use General Linear Model to analyze balanced and unbalanced designs.

Factors may be crossed or nested, fixed or random. You may include up to 50 response variables and up to 31 factors at one time.

Dialog box items

Responses: Enter the columns containing the response variables.

Model: Enter the terms to be included in the model. See Specifying a Model for more information.

Random factors: Enter any columns containing random factors. Do not include model terms that involve other factors.

<Options>

<Graphs>

<Results>

<Storage>

Data – Balanced ANOVA

You need one column for each response variable and one column for each factor, with each row representing an observation. Regardless of whether factors are crossed or nested, use the same form for the data. Factor columns may be numeric, text, or date/time. If you wish to change the order in which text categories are processed from their default alphabetical order, you can define your own order. See Ordering Text Categories. You may include up to 50 response variables and up to 31 factors at one time.

Balanced data are required except for one-way designs. The requirement for balanced data extends to nested factors as well. Suppose A has 3 levels, and B is nested within A. If B has 4 levels within the first level of A, B must have 4 levels within the second and third levels of A. Minitab will tell you if you have unbalanced nesting. In addition, the subscripts used to indicate the 4 levels of B within each level of A must be the same. Thus, the four levels of B cannot be (1 2 3 4) in level 1 of A, (5 6 7 8) in level 2 of A, and (9 10 11 12) in level 3 of A.

If any response or factor column specified contains missing data, that entire observation (row) is excluded from all computations. The requirement that data be balanced must be preserved after missing data are omitted. If an observation is missing for one response variable, that row is eliminated for all responses. If you want to eliminate missing rows separately for each response, perform a separate ANOVA for each response.

To perform a balanced ANOVA

1 Choose **Stat > ANOVA > Balanced ANOVA**.

2 In **Responses**, enter up to 50 numeric columns containing the response variables.

3 In **Model**, type the model terms you want to fit. See Specifying the Model Terms.

4 If you like, use any dialog box options, then click **OK**.

Balanced designs

Your design must be balanced to use balanced ANOVA, with the exception of a one-way design. A *balanced* design is one with equal numbers of observations at each combination of your treatment levels. A quick test to see whether or not you have a balanced design is to use **Stat > Tables > Cross Tabulation and Chi-Square**. Enter your classification variables and see if you have equal numbers of observations in each cell, indicating balanced data.

Restricted and unrestricted form of mixed models

A mixed model is one with both fixed and random factors. There are two forms of this model: one requires the crossed, mixed terms to sum to zero over subscripts corresponding to fixed effects (this is called the restricted model), and the other does not. See Example of both restricted and unrestricted forms of the mixed model. Many textbooks use the restricted model. Most statistics programs use the unrestricted model. Minitab fits the unrestricted model by default, but you can choose to fit the restricted form. The reasons to choose one form over the other have not been clearly defined in the statistical literature. Searle et al. [25] say "that question really has no definitive, universally acceptable answer," but also say that one "can decide which is more appropriate to the data at hand," without giving guidance on how to do so.

Your choice of model form does not affect the sums of squares, degrees of freedom, mean squares, or marginal and cell means. It does affect the expected mean squares, error terms for F-tests, and the estimated variance components. See Example of both restricted and unrestricted forms of the mixed model.

Specifying a model

Specify a model in the Model text box using the form **Y = expression**. The **Y** is not included in the Model text box. The Calc > Make Patterned Data > Simple Set of Numbers command can be helpful in entering the level numbers of a factor.

Rules for Expression Models

1 ∗ indicates an interaction term. For example, A∗B is the interaction of the factors A and B.

2 ( ) indicate nesting. When B is nested within A, type B(A). When C is nested within both A and B, type C(A B). Terms in parentheses are always factors in the model and are listed with blanks between them.

3 Abbreviate a model using a | or ! to indicate crossed factors and a – to remove terms.

Models with many terms take a long time to compute.

Examples of what to type in the Model text box

Two factors crossed: `A B A*B`

Three factors crossed: `A B C A*B A*C B*C A*B*C`

Three factors nested: `A B(A) C(A B)`

Crossed and nested (B nested within A, and both crossed with C): `A B(A) C A*C B*C(A)`

When a term contains both crossing and nesting, put the ∗ (or crossed factor) first, as in C∗B(A), not B(A)∗C

Example of entering level numbers for a data set

Here is an easy way to enter the level numbers for a three-way crossed design with a, b, and c levels of factors A, B, C, with n observations per cell:

1 Choose **Calc > Make Patterned Data > Simple Set of Numbers**, and press <F3> to reset defaults. Enter *A* in **Store patterned data in**. Enter *1* in **From first value**. Enter the number of levels in A in **To last value**. Enter the product of bcn in **List the whole sequence**. Click **OK**.

2 Choose **Calc > Make Patterned Data > Simple Set of Numbers**, and press <F3> to reset defaults. Enter *B* in **Store patterned data in**. Enter *1* in **From first value**. Enter the number of levels in B in **To last value**. Enter the number of levels in A in **List each value**. Enter the product of cn in **List the whole sequence**. Click **OK**.

3 Choose **Calc > Make Patterned Data > Simple Set of Numbers**, and press <F3> to reset defaults. Enter *C* in **Store patterned data in**. Enter *1* in *From first value*. Enter the number of levels in C in *To last value*. Enter the product of ab in **List each value**. Enter the sample size n in **List the whole sequence**. Click **OK**.

Specifying the model terms

You must specify the model terms in the **Model** box. This is an abbreviated form of the statistical model that you may see in textbooks. Because you enter the response variables in **Responses,** in **Model** you enter only the variables or products of variables that correspond to terms in the statistical model. Minitab uses a simplified version of a statistical model as it appears in many textbooks. Here are some examples of statistical models and the terms to enter in **Model**. A, B, and C represent factors.

Case	Statistical model	Terms in model
Factors A, B crossed	$y_{ijk} = \mu + a_i + b_j + ab_{ij} + e_{k(ij)}$	A B A∗ B
Factors A, B, C crossed	$y_{ijkl} = \mu + a_i + b_j + c_k + ab_{ij} + ac_{ik} + bc_{jk} + abc_{ijk} + e_{l(ijk)}$	A B C A∗B A∗ C B∗C A∗ B∗C
3 factors nested (B within A, C within A and B)	$y_{ijkl} = \mu + a_i + b_{j(i)} + c_{k(ij)} + e_{l(ijk)}$	A B(A) C(AB)
Crossed and nested (B nested within A, both crossed with C)	$y_{ijkl} = \mu + a_i + b_{j(i)} + c_k + ac_{ik} + bc_{jk(i)} + e_{l(ijk)}$	A B (A) C A∗C B∗C

In Minitab's models you omit the subscripts, μ, e, and +'s that appear in textbook models. An ∗ is used for an interaction term and parentheses are used for nesting. For example, when B is nested within A, you enter B (A), and when C is nested within both A and B, you enter C (A B). Enter B(A) C(B) for the case of 3 sequentially nested factors. Terms in parentheses are always factors in the model and are listed with blanks between them. Thus, D ∗ F (A B E) is correct but D ∗ F (A ∗ B E) and D (A ∗ B ∗ C) are not. Also, one set of parentheses cannot be used inside another set. Thus, C (A B) is correct but C (A B (A)) is not. An interaction term between a nested factor and the factor it is nested within is invalid.

See Specifying terms involving covariates for details on specifying models with covariates.

Several special rules apply to naming columns. You may omit the quotes around variable names. Because of this, variable names must start with a letter and contain only letters and numbers. Alternatively, you can use C notation (C1, C2, etc.) to denote data columns. You can use special symbols in a variable name, but then you must enclose the name in single quotes.

You can specify multiple responses. In this case, a separate analysis of variance will be performed for each response.

Specifying models for some specialized designs

Some experimental designs can effectively provide information when measurements are difficult or expensive to make or can minimize the effect of unwanted variability on treatment inference. The following is a brief discussion of three commonly used designs that will show you how to specify the model terms in Minitab. To illustrate these designs, two treatment factors (A and B) and their interaction (A∗B) are considered. These designs are not restricted to two factors, however. If your design is balanced, you can use balanced ANOVA to analyze your data. Otherwise, use GLM.

Randomized block design

A *randomized block* design is a commonly used design for minimizing the effect of variability when it is associated with discrete units (e.g. location, operator, plant, batch, time). The usual case is to randomize one replication of each treatment combination within each block. There is usually no intrinsic interest in the blocks and these are considered to be random factors. The usual assumption is that the block by treatment interaction is zero and this interaction becomes the error term for testing treatment effects. If you name the block variable as Block, enter *Block A B A∗B* in **Model** and enter *Block* in **Random Factors**.

Split-plot design

A *split-plot* design is another blocking design, which you can use if you have two or more factors. You might use this design when it is more difficult to randomize one of the factors compared to the other(s). For example, in an agricultural experiment with the factors variety and harvest date, it may be easier to plant each variety in contiguous rows and to randomly assign the harvest dates to smaller sections of the rows. The block, which can be replicated, is termed the *main plot* and within these the smaller plots (variety strips in example) are called *subplots*.

This design is frequently used in industry when it is difficult to randomize the settings on machines. For example, suppose that factors are temperature and material amount, but it is difficult to change the temperature setting. If the blocking factor is operator, observations will be made at different temperatures with each operator, but the temperature setting is held constant until the experiment is run for all material amounts. In this example, the plots under operator constitute the main plots and temperatures constitute the subplots.

There is no single error term for testing all factor effects in a split-plot design. If the levels of factor A form the subplots, then the mean square for Block ∗ A will be the error term for testing factor A. There are two schools of thought for what should be the error term to use for testing B and A ∗ B. If you enter the term Block ∗ B, the expected mean squares show that the mean square for Block ∗ B is the proper term for testing factor B and that the remaining error (which is Block ∗ A ∗ B) will be used for testing A ∗ B. However, it is often assumed that the Block ∗ B and Block ∗ A ∗ B interactions do not exist and these are then lumped together into error [6]. You might also pool the two terms if the mean square for Block ∗ B is small relative to Block ∗ A ∗ B. If you don't pool, enter *Block A Block ∗ A B Block ∗B A ∗ B* in **Model** and what is labeled as Error is really Block ∗ A ∗ B. If you do pool terms, enter *Block A Block ∗ A B A ∗ B* in **Model** and what is labeled as Error is the set of pooled terms. In both cases enter *Block* in **Random Factors**.

Latin square with repeated measures design

A *repeated measures* design is a design where repeated measurements are made on the same subject. There are a number of ways in which treatments can be assigned to subjects. With living subjects especially, systematic differences (due to learning, acclimation, resistance, etc.) between successive observations may be suspected. One common way to assign treatments to subjects is to use a Latin square design. An advantage of this design for a repeated measures experiment is that it ensures a balanced fraction of a complete factorial (i.e. all treatment combinations represented) when subjects are limited and the sequence effect of treatment can be considered to be negligible.

A *Latin square* design is a blocking design with two orthogonal blocking variables. In an agricultural experiment there might be perpendicular gradients that might lead you to choose this design. For a repeated measures experiment, one blocking variable is the group of subjects and the other is time. If the treatment factor B has three levels, b1, b2, and b3, then one of twelve possible Latin square randomizations of the levels of B to subjects groups over time is:

	Time 1	Time 2	Time 3
Group 1	b2	b3	b1
Group 2	b3	b1	b2
Group 3	b1	b2	b3

The subjects receive the treatment levels in the order specified across the row. In this example, group 1 subjects would receive the treatments levels in order b2, b3, b1. The interval between administering treatments should be chosen to minimize carryover effect of the previous treatment.

This design is commonly modified to provide information on one or more additional factors. If each group was assigned a different level of factor A, then information on the A and A ∗ B effects could be made available with minimal effort if an assumption about the sequence effect given to the groups can be made. If the sequence effects are negligible compared to the effects of factor A, then the group effect could be attributed to factor A. If interactions with time are negligible, then partial information on the A ∗ B interaction may be obtained [29]. In the language of repeated measures designs, factor A is called a *between-subjects* factor and factor B a *within-subjects* factor.

Let's consider how to enter the model terms into Minitab. If the group or A factor, subject, and time variables were named A, Subject, and Time, respectively, enter *A Subject(A) Time B A ∗ B* in **Model** and enter *Subject* in **Random Factors**.

It is not necessary to randomize a repeated measures experiments according to a Latin square design. See Example of a repeated measures design for a repeated measures experiment where the fixed factors are arranged in a complete factorial design.

Specifying reduced models

You can fit *reduced* models. For example, suppose you have a three factor design, with factors, A, B, and C. The *full* model would include all one factor terms: A, B, C, all two-factor interactions: A ∗ B, A ∗ C, B ∗ C, and the three-factor interaction: A ∗ B ∗ C. It becomes a reduced model by omitting terms. You might reduce a model if terms are not significant or if you need additional error degrees of freedom and you can assume that certain terms are zero. For this example, the model with terms A B C A ∗ B is a reduced three-factor model.

One rule about specifying reduced models is that they must be hierarchical. That is, for a term to be in the model, all lower order terms contained in it must also be in the model. For example, suppose there is a model with four factors: A, B, C, and D. If the term A ∗ B ∗ C is in the model then the terms A B C A ∗ B A ∗ C B ∗ C must also be in the model, though any terms with D do not have to be in the model. The hierarchical structure applies to nesting as well. If B (A) is in the model, then A must be also.

Because models can be quite long and tedious to type, two shortcuts have been provided. A vertical bar indicates crossed factors, and a minus sign removes terms.

Long form	**Short form**
A B C A ∗ B A ∗ C B ∗ C A ∗ B ∗ C	A | B | C
A B C A ∗ B A ∗ C B ∗ C	A | B | C – A ∗ B ∗ C
A B C B ∗ C E	A B | C E
A B C D A ∗ B A ∗ C A ∗ D B ∗ C B ∗ D C ∗ D A ∗ B ∗ D A ∗ C ∗ D B ∗ C ∗ D	A | B | C | D – A ∗ B ∗ C – A ∗ B ∗ C ∗ D
A B (A) C A ∗ C B ∗ C	A | B (A) | C

In general, all crossings are done for factors separated by bars unless the cross results in an illegal term. For example, in the last example, the potential term A ∗ B (A) is illegal and Minitab automatically omits it. If a factor is nested, you must indicate this when using the vertical bar, as in the last example with the term B (A).

Using patterned data to set up factor levels

Minitab's set patterned data capability can be helpful when entering numeric factor levels. For example, to enter the level values for a three-way crossed design with a, b, and c (a, b, and c represent numbers) levels of factors A, B, C, and n observations per cell, fill out the Calc > Set Patterned Data > Simple Set of Numbers dialog box and execute 3 times, once for each factor, as shown:

		Factor	
Dialog item	**A**	**B**	**C**
From first value	1	1	1
From last value	a	b	c
List each value	bcn	cn	n
List the whole sequence	1	a	ab

Balanced ANOVA – Options

Stat > ANOVA > Balanced ANOVA > Options

Use to fit a restricted model.

Dialog box items

Use the restricted form of the model: Check to fit a restricted model, with mixed interaction terms restricted to sum to zero over the fixed effects. Minitab will fit an unrestricted model if this box is left unchecked. See Restricted and unrestricted form of mixed models.

Balanced ANOVA – Graphs

Stat > ANOVA > Balanced ANOVA > Graphs

Displays residual plots. You do not have to store the residuals in order to produce these plots.

Dialog box items

Residual Plots

Individual plots: Choose to display one or more plots.

Histogram of residuals: Check to display a histogram of the residuals.

Normal plot of residuals: Check to display a normal probability plot of the residuals.

Residuals versus fits: Check to plot the residuals versus the fitted values.

Residuals versus order: Check to plot the residuals versus the order of the data. The row number for each data point is shown on the x-axis–for example, 1 2 3 4... n.

Four in one: Choose to display a layout of a histogram of residuals, a normal plot of residuals, a plot of residuals versus fits, and a plot of residuals versus order.

Residuals versus the variables: Enter one or more columns containing the variables against which you want to plot the residuals. Minitab displays a separate graph for each column.

Balanced ANOVA – Results

Stat > ANOVA > Balanced ANOVA > Results

Use to control the Session window output.

Dialog box items

Display expected mean squares and variance components: Check to display a table that contains expected mean squares, estimated variance components, and the error term (the denominator) used in each F-test. See Expected means squares.

Display means corresponding to the terms: Enter terms for which a table of sample sizes and means will be printed. These terms must be in the model.

Expected mean squares

If you do not specify any factors to be random, Minitab will assume that they are fixed. In this case, the denominator for F-statistics will be the MSE. However, for models which include random terms, the MSE is not always the correct error term. You can examine the expected means squares to determine the error term that was used in the F-test.

When you select **Display expected mean squares and variance components** in the **Results** subdialog box, Minitab will print a table of expected mean squares, estimated variance components, and the error term (the denominator mean squares) used in each F-test. The *expected mean squares* are the expected values of these terms with the specified model. If there is no exact F-test for a term, Minitab solves for the appropriate error term in order to construct an approximate F-test. This test is called a *synthesized test*.

The estimates of variance components are the usual unbiased analysis of variance estimates. They are obtained by setting each calculated mean square equal to its expected mean square, which gives a system of linear equations in the unknown variance components that is then solved. Unfortunately, this method can result in negative estimates, which should be set to zero. Minitab, however, prints the negative estimates because they sometimes indicate that the model being fit is inappropriate for the data. Variance components are not estimated for fixed terms.

Balanced ANOVA – Storage

Stat > ANOVA > Balanced ANOVA > Storage

Stores the fitted values and residuals.

Dialog box items

Fits: Check to store the fitted values for each observation in the data set in the next available columns, using one column for each response.

Residuals: Check to store the residuals using one column for each response.

Example of ANOVA with Two Crossed Factors

An experiment was conducted to test how long it takes to use a new and an older model of calculator. Six engineers each work on both a statistical problem and an engineering problem using each calculator model and the time in minutes to solve the problem is recorded. The engineers can be considered as blocks in the experimental design. There are two factors– type of problem, and calculator model– each with two levels. Because each level of one factor occurs in combination with each level of the other factor, these factors are crossed. The example and data are from Neter, Wasserman, and Kutner [19], page 936.

1 Open the worksheet EXH_AOV.MTW.

2 Choose **Stat > ANOVA > Balanced ANOVA**.

3 In **Responses**, enter *SolveTime*.

4 In **Model**, type *Engineer ProbType* | *Calculator*.

5 In **Random Factors**, enter *Engineer*.

6 Click **Results**. In **Display means corresponding to the terms**, type *ProbType* | *Calculator*. Click **OK** in each dialog box.

Session window output

ANOVA: SolveTime versus Engineer, ProbType, Calculator

```
Factor      Type    Levels  Values
Engineer    random       6  Adams, Dixon, Erickson, Jones, Maynes, Williams
ProbType    fixed        2  Eng, Stat
Calculator  fixed        2  New, Old


Analysis of Variance for SolveTime

Source               DF      SS      MS        F      P
Engineer              5   1.053   0.211     3.13  0.039
ProbType              1  16.667  16.667   247.52  0.000
Calculator            1  72.107  72.107  1070.89  0.000
ProbType*Calculator   1   3.682   3.682    54.68  0.000
Error                15   1.010   0.067
Total                23  94.518


S = 0.259487   R-Sq = 98.93%   R-Sq(adj) = 98.36%


Means

ProbType   N  SolveTime
Eng       12     3.8250
Stat      12     5.4917


Calculator   N  SolveTime
New         12     2.9250
Old         12     6.3917


ProbType  Calculator  N  SolveTime
Eng       New         6     2.4833
Eng       Old         6     5.1667
```

```
Stat      New         6     3.3667
Stat      Old         6     7.6167
```

Interpreting the results

Minitab displays a list of factors, with their type (fixed or random), number of levels, and values. Next displayed is the analysis of variance table. The analysis of variance indicates that there is a significant calculator by problem type interaction, which implies that the decrease in mean compilation time in switching from the old to the new calculator depends upon the problem type.

Because you requested means for all factors and their combinations, the means of each factor level and factor level combinations are also displayed. These show that the mean compilation time decreased in switching from the old to new calculator type.

Example of a Mixed Model ANOVA

A company ran an experiment to see how several conditions affect the thickness of a coating substance that it manufactures. The experiment was run at two different times, in the morning and in the afternoon. Three operators were chosen from a large pool of operators employed by the company. The manufacturing process was run at three settings, 35, 44, and 52. Two determinations of thickness were made by each operator at each time and setting. Thus, the three factors are crossed. One factor, operator, is random; the other two, time and setting, are fixed.

The statistical model is:

$Y_{ijkl} = \mu + T_i + O_j + S_k + TO_{ij} + TS_{ik} + OS_{jk} + TOS_{ijk} + e_{ijkl}$,

where T_i is the time effect, O_j is the operator effect, and S_k is the setting effect, and TO_{ij}, TS_{ik}, OS_{jk}, and TOS_{ijk} are the interaction effects.

Operator, all interactions with operator, and error are random. The random terms are:

O_j TO_{ij} OS_{jk} TOS_{ijk} e_{ijkl}

These terms are all assumed to be normally distributed random variables with mean zero and variances given by

var (O_j) = V(O)   var (TO_{ij}) = V(TO)

var (TOS_{jkl}) = V(TOS)   var (e_{ijkl}) = V(e) = σ**2

These variances are called variance components. The output from expected means squares contains estimates of these variances.

In the unrestricted model, all these random variables are independent. The remaining terms in this model are fixed.

In the restricted model, any term which contains one or more subscripts corresponding to fixed factors is required to sum to zero over each fixed subscript. In the example, this means:

$$\sum_{i} (T_i) = 0 \qquad \sum_{k} (S_k) = 0 \qquad \sum_{i} (TO_{ij}) = 0$$

$$\sum_{k} (TS_{ik}) = 0 \qquad \sum_{k} (OS_{jk}) = 0 \qquad \sum_{i} (TOS_{ijk}) = 0$$

Your choice of model does not affect the sums of squares, degrees of freedom, mean squares, or marginal and cell means. It does affect the expected mean squares, error term for the F-tests, and the estimated variance components.

Step 1: Fit the restricted form of the model

1 Open the worksheet EXH_AOV.MTW.

2 Choose **Stat > ANOVA > Balanced ANOVA**.

3 In **Responses**, enter *Thickness*.

4 In **Model**, type *Time* | *Operator* | *Setting*.

5 In **Random Factors**, enter *Operator*.

6 Click **Options**. Check **Use the restricted form of the mixed model**. Click **OK**.

7 Click **Results**. Check **Display expected mean squares and variance components**.

8 Click **OK** in each dialog box.

Step 2: Fit the unrestricted form of the model

1 Repeat steps 1-8 above except that in 6, uncheck **Use the restricted form of the mixed model**.

Session window output for restricted case

ANOVA: Thickness versus Time, Operator, Setting

```
Factor    Type    Levels  Values
Time      fixed        2  1, 2
Operator  random       3  1, 2, 3
Setting   fixed        3  35, 44, 52


Analysis of Variance for Thickness

Source                 DF       SS      MS       F      P
Time                    1      9.0     9.0    0.29  0.644
Operator                2   1120.9   560.4  165.38  0.000
Setting                 2  15676.4  7838.2   73.18  0.001
Time*Operator           2     62.0    31.0    9.15  0.002
Time*Setting            2    114.5    57.3    2.39  0.208
Operator*Setting        4    428.4   107.1   31.61  0.000
Time*Operator*Setting   4     96.0    24.0    7.08  0.001
Error                  18     61.0     3.4
Total                  35  17568.2


S = 1.84089   R-Sq = 99.65%   R-Sq(adj) = 99.32%


                                            Expected Mean Square
                          Variance   Error  for Each Term (using
   Source                 component   term  restricted model)
1  Time                                  4  (8) + 6 (4) + 18 Q[1]
2  Operator                  46.421      8  (8) + 12 (2)
3  Setting                               6  (8) + 4 (6) + 12 Q[3]
4  Time*Operator              4.602      8  (8) + 6 (4)
5  Time*Setting                          7  (8) + 2 (7) + 6 Q[5]
6  Operator*Setting          25.931      8  (8) + 4 (6)
7  Time*Operator*Setting     10.306      8  (8) + 2 (7)
8  Error                      3.389         (8)
```

Session window output for unrestricted case

ANOVA: Thickness versus Time, Operator, Setting

```
Factor    Type    Levels  Values
Time      fixed        2  1, 2
Operator  random       3  1, 2, 3
Setting   fixed        3  35, 44, 52


Analysis of Variance for Thickness

Source                 DF       SS      MS      F      P
Time                    1      9.0     9.0   0.29  0.644
Operator                2   1120.9   560.4   4.91  0.090 x
Setting                 2  15676.4  7838.2  73.18  0.001
Time*Operator           2     62.0    31.0   1.29  0.369
Time*Setting            2    114.5    57.3   2.39  0.208
Operator*Setting        4    428.4   107.1   4.46  0.088
Time*Operator*Setting   4     96.0    24.0   7.08  0.001
Error                  18     61.0     3.4
Total                  35  17568.2

x Not an exact F-test.


S = 1.84089   R-Sq = 99.65%   R-Sq(adj) = 99.32%
```

```
                               Variance  Error  Expected Mean Square for Each
   Source                     component   term  Term (using unrestricted model)
1  Time                                      4  (8) + 2 (7) + 6 (4) + Q[1,5]
2  Operator                      37.194      *  (8) + 2 (7) + 4 (6) + 6 (4) + 12
                                                (2)
3  Setting                                   6  (8) + 2 (7) + 4 (6) + Q[3,5]
4  Time*Operator                  1.167      7  (8) + 2 (7) + 6 (4)
5  Time*Setting                              7  (8) + 2 (7) + Q[5]
6  Operator*Setting              20.778      7  (8) + 2 (7) + 4 (6)
7  Time*Operator*Setting         10.306      8  (8) + 2 (7)
8  Error                          3.389         (8)


* Synthesized Test.


Error Terms for Synthesized Tests

                                  Synthesis of
Source       Error DF  Error MS  Error MS
2 Operator       3.73     114.1  (4) + (6) - (7)
```

Interpreting the results

The organization of the output is the same for restricted and unrestricted models: a table of factor levels, the analysis of variance table, and as requested, the expected mean squares. The differences in the output are in the expected means squares and the F-tests for some model terms. In this example, the F-test for Operator is synthesized for the unrestricted model because it could not be calculated exactly.

Examine the 3 factor interaction, Time∗Operator∗Setting. The F-test is the same for both forms of the mixed model, giving a p-value of 0.001. This implies that the coating thickness depends upon the combination of time, operator, and setting. Many analysts would go no further than this test. If an interaction is significant, any lower order interactions and main effects involving terms of the significant interaction are not considered meaningful.

Let's examine where these models give different output. The Operator∗Setting F-test is different, because the error terms are Error in the restricted case and Time∗Operator∗Setting in the unrestricted case, giving p-values of < 0.0005 and 0.088, respectively. Likewise, the Time∗Operator differs for the same reason, giving p-values of 0.002 and 0.369, respectively, for the restricted and unrestricted cases, respectively. The estimated variance components for Operator, Time∗Operator, and Operator∗Setting also differ.

Example of a Repeated Measures Design

The following example contains data from Winer [28], p. 546, to illustrate a complex repeated measures model. An experiment was run to see how several factors affect subject accuracy in adjusting dials. Three subjects perform tests conducted at one of two noise levels. At each of three time periods, the subjects monitored three different dials and make adjustments as needed. The response is an accuracy score. The noise, time, and dial factors are crossed, fixed factors. Subject is a random factor, nested within noise. Noise is a between-subjects factor, time and dial are within-subjects factors.

We enter the model terms in a certain order so that the error terms used for the fixed factors are just below the terms for whose effects they test. (With a single random factor, the interaction of a fixed factor with the random factor becomes the error term for that fixed effect.) Because we specified Subject as Subject(Noise) the first time, we don't need to repeat "(Noise)" in the interactions involving Subject. The interaction ETime*Dial*Subject is not entered in the model because there would be zero degrees of freedom left over for error. This is the correct error term for testing ETime*Dial and by not entering ETime*Dial*Subject in the model, it is labeled as Error and we then have the error term that is needed.

1 Open the worksheet EXH_AOV.MTW.

2 Choose **Stat > ANOVA > Balanced ANOVA**.

3 In **Responses**, enter *Score*.

4 In **Model**, enter *Noise Subject(Noise) ETime Noise*ETime ETime*Subject Dial Noise*Dial Dial*Subject ETime*Dial Noise*ETime*Dial*.

5 In **Random Factors (optional)**, enter *Subject*.

6 Click **Options**.

7 Check **Use the restricted form of the mixed model**, then click **OK**.

8 Click **Results**.

9 Check **Display expected mean squares and variance components**. Click **OK** in each dialog box.

Session window output

ANOVA: Score versus Noise, ETime, Dial, Subject

```
Factor          Type    Levels  Values
Noise           fixed        2  1, 2
Subject(Noise)  random       3  1, 2, 3
ETime           fixed        3  1, 2, 3
Dial            fixed        3  1, 2, 3


Analysis of Variance for Score

Source                DF       SS       MS      F      P
Noise                  1   468.17   468.17   0.75  0.435
Subject(Noise)         4  2491.11   622.78  78.39  0.000
ETime                  2  3722.33  1861.17  63.39  0.000
Noise*ETime            2   333.00   166.50   5.67  0.029
ETime*Subject(Noise)   8   234.89    29.36   3.70  0.013
Dial                   2  2370.33  1185.17  89.82  0.000
Noise*Dial             2    50.33    25.17   1.91  0.210
Dial*Subject(Noise)    8   105.56    13.19   1.66  0.184
ETime*Dial             4    10.67     2.67   0.34  0.850
Noise*ETime*Dial       4    11.33     2.83   0.36  0.836
Error                 16   127.11     7.94
Total                 53  9924.83


S = 2.81859   R-Sq = 98.72%   R-Sq(adj) = 95.76%


                                             Expected Mean Square
                           Variance  Error  for Each Term (using
    Source                component   term  restricted model)
 1  Noise                                2  (11) + 9 (2) + 27 Q[1]
 2  Subject(Noise)           68.315     11  (11) + 9 (2)
 3  ETime                                5  (11) + 3 (5) + 18 Q[3]
 4  Noise*ETime                          5  (11) + 3 (5) + 9 Q[4]
 5  ETime*Subject(Noise)      7.139     11  (11) + 3 (5)
 6  Dial                                 8  (11) + 3 (8) + 18 Q[6]
 7  Noise*Dial                           8  (11) + 3 (8) + 9 Q[7]
 8  Dial*Subject(Noise)       1.750     11  (11) + 3 (8)
 9  ETime*Dial                          11  (11) + 6 Q[9]
10  Noise*ETime*Dial                    11  (11) + 3 Q[10]
11  Error                     7.944         (11)
```

Interpreting the results

Minitab displays the table of factor levels, the analysis of variance table, and the expected mean squares. Important information to gain from the expected means squares are the estimated variance components and discovering which error term is used for testing the different model terms.

The term labeled Error is in row 11 of the expected mean squares table. The column labeled "Error term" indicates that term 11 was used to test terms 2, 5, and 8 to 10. Dial*Subject is numbered 8 and was used to test the sixth and seventh terms. You can follow the pattern for other terms.

You can gain some idea about how the design affected the sensitivity of F-tests by viewing the variance components. The variance components used in testing within-subjects factors are smaller (7.139, 1.750, 7.944) than the between-subjects variance (68.315). It is typical that a repeated measures model can detect smaller differences in means within subjects as compared to between subjects.

Of the four interactions among fixed factors, the noise by time interaction was the only one with a low p-value (0.029). This implies that there is significant evidence for judging that a subjects' sensitivity to noise changed over time. Because this interaction is significant, at least at a = 0.05, the noise and time main effects are not examined. There is also significant evidence for a dial effect (p-value < 0.0005). Among random terms, there is significant evidence for time by subject (p-value = 0.013) and subject (p-value < 0.0005) effects.

General Linear Model

Overview of Balanced ANOVA and GLM

Balanced ANOVA and general linear model (GLM) are ANOVA procedures for analyzing data collected with many different experimental designs. Your choice between these procedures depends upon the experimental design and the available options. The experimental design refers to the selection of units or subjects to measure, the assignment of treatments to these units or subjects, and the sequence of measurements taken on the units or subjects. Both procedures can fit univariate models to balanced data with up to 31 factors. Here are some of the other options:

	Balanced ANOVA	**GLM**
Can fit unbalanced data	no	yes
Can specify factors as random and obtain expected means squares	yes	yes
Fits covariates	no	yes
Performs multiple comparisons	no	yes
Fits restricted/unrestricted forms of mixed model	yes	unrestricted only

You can use balanced ANOVA to analyze data from balanced designs. See Balanced designs. You can use GLM to analyze data from any balanced design, though you cannot choose to fit the restricted case of the mixed model, which only balanced ANOVA can fit. See Restricted and unrestricted form of mixed models.

To classify your variables, determine if your factors are:

- crossed or nested
- fixed or random
- covariates

For information on how to specify the model, see Specifying the model terms, Specifying terms involving covariates, Specifying reduced models, and Specifying models for some specialized designs.

For easy entering of repeated factor levels into your worksheet, see Using patterned data to set up factor levels.

General Linear Model

Stat > ANOVA > General Linear Model

Use General Linear Model (GLM) to perform univariate analysis of variance with balanced and unbalanced designs, analysis of covariance, and regression, for each response variable.

Calculations are done using a regression approach. A "full rank" design matrix is formed from the factors and covariates and each response variable is regressed on the columns of the design matrix.

You must specify a hierarchical model. In a hierarchical model, if an interaction term is included, all lower order interactions and main effects that comprise the interaction term must appear in the model.

Factors may be crossed or nested, fixed or random. Covariates may be crossed with each other or with factors, or nested within factors. You can analyze up to 50 response variables with up to 31 factors and 50 covariates at one time. For more information see Overview of Balanced ANOVA and GLM.

Dialog box items

Responses: Select the column(s) containing the response variable(s).

Model: Specify the terms to be included in the model. See Specifying a Model for more information.

Random factors: Specify any columns containing random factors. Do not include model terms that involve other factors.

<Covariates>

<Options>

<Comparisons>

<Graphs>

<Results>

<Storage>

<Factor Plots>

Data – General Linear Model

Set up your worksheet in the same manner as with balanced ANOVA: one column for each response variable, one column for each factor, and one column for each covariate, so that there is one row for each observation. The factor columns may be numeric, text, or date/time. If you wish to change the order in which text categories are processed from their default alphabetical order, you can define your own order. See Ordering Text Categories.

Although models can be unbalanced in GLM, they must be "full rank," that is, there must be enough data to estimate all the terms in your model. For example, suppose you have a two-factor crossed model with one empty cell. Then you can fit the model with terms A B, but not A B A∗B. Minitab will tell you if your model is not full rank. In most cases, eliminating some of the high order interactions in your model (assuming, of course, they are not important) can solve this problem.

Nesting does not need to be balanced. A nested factor must have at least 2 levels at some level of the nesting factor. If factor B is nested within factor A, there can be unequal levels of B within each level of A. In addition, the subscripts used to identify the B levels can differ within each level of A. This means, for example, that the B levels can be (1 2 3 4) in level 1 of A, (5 6 7 8) in level 2 of A, and (9 10 11 12) in level 3 of A. A nested factor must have at least 2 levels at some level of the nested factor.

If any response, factor, or covariate column contains missing data, that entire observation (row) is excluded from all computations. If you want to eliminate missing rows separately for each response, perform GLM separately for each response.

To perform an analysis using general linear model

1 Choose **Stat > ANOVA > General Linear Model**.

2 In **Responses**, enter up to 50 numeric columns containing the response variables.

3 In **Model**, type the model terms you want to fit. See Specifying the model terms.

4 If you like, use any dialog box options, then click **OK**.

Design matrix used by General Linear Model

General Linear Model uses a regression approach to fit the model that you specify. First Minitab creates a design matrix, from the factors and covariates, and the model that you specify. The columns of this matrix are the predictors for the regression.

The design matrix has n rows, where n = number of observations, and one block of columns, often called dummy variables, for each term in the model. There are as many columns in a block as there are degrees of freedom for the term. The first block is for the constant and contains just one column, a column of all ones. The block for a covariate also contains just one column, the covariate column itself.

Suppose A is a factor with 4 levels. Then it has 3 degrees of freedom and its block contains 3 columns, call them A1, A2, A3. Each row is coded as one of the following:

level of A	A1	A2	A3
1	1	0	0
2	0	1	0
3	0	0	1
4	–1	–1	–1

Suppose factor B has 3 levels nested within each level of A. Then its block contains (3 – 1) x 4 = 8 columns, call them B11, B12, B21, B22, B31, B32, B41, B42, coded as follows:

level of A	level of B	B11	B12	B21	B22	B31	B32	B41	B42
1	1	1	0	0	0	0	0	0	0
1	2	0	1	0	0	0	0	0	0
1	3	–1	–1	0	0	0	0	0	0
2	1	0	0	1	0	0	0	0	0
2	2	0	0	0	1	0	0	0	0
2	3	0	0	–1	–1	0	0	0	0
3	1	0	0	0	0	1	0	0	0
3	2	0	0	0	0	0	1	0	0
3	3	0	0	0	0	–1	–1	0	0
4	1	0	0	0	0	0	0	1	0
4	2	0	0	0	0	0	0	0	1
4	3	0	0	0	0	0	0	–1	–1

To calculate the dummy variables for an interaction term, just multiply all the corresponding dummy variables for the factors and/or covariates in the interaction. For example, suppose factor A has 6 levels, C has 3 levels, D has 4 levels, and Z and W are covariates. Then the term A ∗ C ∗ D ∗ Z ∗ W ∗ W has 5 x 2 x 3 x 1 x 1 x 1 = 30 dummy variables. To obtain them, multiply each dummy variable for A by each for C, by each for D, by the covariates Z once and W twice.

Specifying a model

Specify a model in the Model text box using the form **Y = expression**. The **Y** is not included in the Model text box. The Calc > Make Patterned Data > Simple Set of Numbers command can be helpful in entering the level numbers of a factor.

Rules for Expression Models

1. ∗ indicates an interaction term. For example, A∗B is the interaction of the factors A and B.
2. ( ) indicate nesting. When B is nested within A, type B(A). When C is nested within both A and B, type C(A B). Terms in parentheses are always factors in the model and are listed with blanks between them.
3. Abbreviate a model using a | or ! to indicate crossed factors and a – to remove terms.

Models with many terms take a long time to compute.

Examples of what to type in the Model text box

Two factors crossed: `A B A∗B`

Three factors crossed: `A B C A∗B A∗C B∗C A∗B∗C`

Three factors nested: `A B(A) C(A B)`

Crossed and nested (B nested within A, and both crossed with C): `A B(A) C A∗C B∗C(A)`

When a term contains both crossing and nesting, put the ∗ (or crossed factor) first, as in C∗B(A), not B(A)∗C

Example of entering level numbers for a data set

Here is an easy way to enter the level numbers for a three-way crossed design with a, b, and c levels of factors A, B, C, with n observations per cell:

1. Choose **Calc > Make Patterned Data > Simple Set of Numbers**, and press <F3> to reset defaults. Enter *A* in **Store patterned data in**. Enter *1* in **From first value**. Enter the number of levels in A in **To last value**. Enter the product of bcn in **List the whole sequence**. Click **OK**.
2. Choose **Calc > Make Patterned Data > Simple Set of Numbers**, and press <F3> to reset defaults. Enter *B* in **Store patterned data in**. Enter *1* in **From first value**. Enter the number of levels in B in **To last value**. Enter the number of levels in A in **List each value**. Enter the product of cn in **List the whole sequence**. Click **OK**.
3. Choose **Calc > Make Patterned Data > Simple Set of Numbers**, and press <F3> to reset defaults. Enter *C* in **Store patterned data in**. Enter *1* in *From first value*. Enter the number of levels in C in *To last value*. Enter the product of ab in **List each value**. Enter the sample size n in **List the whole sequence**. Click **OK**.

Specifying the model terms

You must specify the model terms in the **Model** box. This is an abbreviated form of the statistical model that you may see in textbooks. Because you enter the response variables in **Responses,** in **Model** you enter only the variables or products of variables that correspond to terms in the statistical model. Minitab uses a simplified version of a statistical model as it appears in many textbooks. Here are some examples of statistical models and the terms to enter in **Model**. A, B, and C represent factors.

Case	Statistical model	Terms in model
Factors A, B crossed	$y_{ijk} = \mu + a_i + b_j + ab_{ij} + e_{k(ij)}$	A B A∗ B
Factors A, B, C crossed	$y_{ijkl} = \mu + a_i + b_j + c_k + ab_{ij} + ac_{ik} + bc_{jk} + abc_{ijk} + e_{l(ijk)}$	A B C A∗B A∗ C B∗C A∗ B∗C
3 factors nested (B within A, C within A and B)	$y_{ijkl} = \mu + a_i + b_{j(i)} + c_{k(ij)} + e_{l(ijk)}$	A B(A) C(AB)
Crossed and nested (B nested within A, both crossed with C)	$y_{ijkl} = \mu + a_i + b_{j(i)} + c_k + ac_{ik} + bc_{jk(i)} + e_{l(ijk)}$	A B (A) C A∗C B∗C

In Minitab's models you omit the subscripts, μ, e, and +'s that appear in textbook models. An ∗ is used for an interaction term and parentheses are used for nesting. For example, when B is nested within A, you enter B (A), and when C is nested within both A and B, you enter C (A B). Enter B(A) C(B) for the case of 3 sequentially nested factors. Terms in parentheses are always factors in the model and are listed with blanks between them. Thus, D ∗ F (A B E) is correct but D ∗ F (A ∗ B E) and D (A ∗ B ∗ C) are not. Also, one set of parentheses cannot be used inside another set. Thus, C (A B) is correct but C (A B (A)) is not. An interaction term between a nested factor and the factor it is nested within is invalid.

See Specifying terms involving covariates for details on specifying models with covariates.

Several special rules apply to naming columns. You may omit the quotes around variable names. Because of this, variable names must start with a letter and contain only letters and numbers. Alternatively, you can use C notation (C1, C2, etc.) to denote data columns. You can use special symbols in a variable name, but then you must enclose the name in single quotes.

You can specify multiple responses. In this case, a separate analysis of variance will be performed for each response.

Specifying models for some specialized designs

Some experimental designs can effectively provide information when measurements are difficult or expensive to make or can minimize the effect of unwanted variability on treatment inference. The following is a brief discussion of three commonly used designs that will show you how to specify the model terms in Minitab. To illustrate these designs, two treatment factors (A and B) and their interaction (A∗B) are considered. These designs are not restricted to two factors, however. If your design is balanced, you can use balanced ANOVA to analyze your data. Otherwise, use GLM.

Randomized block design

A *randomized block* design is a commonly used design for minimizing the effect of variability when it is associated with discrete units (e.g. location, operator, plant, batch, time). The usual case is to randomize one replication of each treatment combination within each block. There is usually no intrinsic interest in the blocks and these are considered to be random factors. The usual assumption is that the block by treatment interaction is zero and this interaction becomes the error term for testing treatment effects. If you name the block variable as Block, enter *Block A B A∗B* in **Model** and enter *Block* in **Random Factors**.

Split-plot design

A *split-plot* design is another blocking design, which you can use if you have two or more factors. You might use this design when it is more difficult to randomize one of the factors compared to the other(s). For example, in an agricultural experiment with the factors variety and harvest date, it may be easier to plant each variety in contiguous rows and to randomly assign the harvest dates to smaller sections of the rows. The block, which can be replicated, is termed the *main plot* and within these the smaller plots (variety strips in example) are called *subplots*.

This design is frequently used in industry when it is difficult to randomize the settings on machines. For example, suppose that factors are temperature and material amount, but it is difficult to change the temperature setting. If the blocking factor is operator, observations will be made at different temperatures with each operator, but the temperature setting is held constant until the experiment is run for all material amounts. In this example, the plots under operator constitute the main plots and temperatures constitute the subplots.

There is no single error term for testing all factor effects in a split-plot design. If the levels of factor A form the subplots, then the mean square for Block ∗ A will be the error term for testing factor A. There are two schools of thought for what should be the error term to use for testing B and A ∗ B. If you enter the term Block ∗ B, the expected mean squares show that the mean square for Block ∗ B is the proper term for testing factor B and that the remaining error (which is Block ∗ A ∗ B) will be used for testing A ∗ B. However, it is often assumed that the Block ∗ B and Block ∗ A ∗ B interactions do not exist and these are then lumped together into error [6]. You might also pool the two terms if the mean square for Block ∗ B is small relative to Block ∗ A ∗ B. If you don't pool, enter *Block A Block ∗ A B Block ∗B A ∗ B* in **Model** and what is labeled as Error is really Block ∗ A ∗ B. If you do pool terms, enter *Block A Block ∗ A B A ∗ B* in **Model** and what is labeled as Error is the set of pooled terms. In both cases enter *Block* in **Random Factors**.

Latin square with repeated measures design

A *repeated measures* design is a design where repeated measurements are made on the same subject. There are a number of ways in which treatments can be assigned to subjects. With living subjects especially, systematic differences (due to learning, acclimation, resistance, etc.) between successive observations may be suspected. One common way to assign treatments to subjects is to use a Latin square design. An advantage of this design for a repeated measures experiment is that it ensures a balanced fraction of a complete factorial (i.e. all treatment combinations represented) when subjects are limited and the sequence effect of treatment can be considered to be negligible.

A *Latin square* design is a blocking design with two orthogonal blocking variables. In an agricultural experiment there might be perpendicular gradients that might lead you to choose this design. For a repeated measures experiment, one blocking variable is the group of subjects and the other is time. If the treatment factor B has three levels, b1, b2, and b3, then one of twelve possible Latin square randomizations of the levels of B to subjects groups over time is:

	Time 1	Time 2	Time 3
Group 1	b2	b3	b1
Group 2	b3	b1	b2
Group 3	b1	b2	b3

The subjects receive the treatment levels in the order specified across the row. In this example, group 1 subjects would receive the treatments levels in order b2, b3, b1. The interval between administering treatments should be chosen to minimize carryover effect of the previous treatment.

This design is commonly modified to provide information on one or more additional factors. If each group was assigned a different level of factor A, then information on the A and A ∗ B effects could be made available with minimal effort if an assumption about the sequence effect given to the groups can be made. If the sequence effects are negligible compared to the effects of factor A, then the group effect could be attributed to factor A. If interactions with time are negligible, then partial information on the A ∗ B interaction may be obtained [29]. In the language of repeated measures designs, factor A is called a *between-subjects* factor and factor B a *within-subjects* factor.

Let's consider how to enter the model terms into Minitab. If the group or A factor, subject, and time variables were named A, Subject, and Time, respectively, enter *A Subject(A) Time B A ∗ B* in **Model** and enter *Subject* in **Random Factors**.

It is not necessary to randomize a repeated measures experiments according to a Latin square design. See Example of a repeated measures design for a repeated measures experiment where the fixed factors are arranged in a complete factorial design.

Specifying reduced models

You can fit *reduced* models. For example, suppose you have a three factor design, with factors, A, B, and C. The *full* model would include all one factor terms: A, B, C, all two-factor interactions: A ∗ B, A ∗ C, B ∗ C, and the three-factor interaction: A ∗ B ∗ C. It becomes a reduced model by omitting terms. You might reduce a model if terms are not significant or if you need additional error degrees of freedom and you can assume that certain terms are zero. For this example, the model with terms A B C A ∗ B is a reduced three-factor model.

One rule about specifying reduced models is that they must be hierarchical. That is, for a term to be in the model, all lower order terms contained in it must also be in the model. For example, suppose there is a model with four factors: A, B, C, and D. If the term A ∗ B ∗ C is in the model then the terms A B C A ∗ B A ∗ C B ∗ C must also be in the model, though any terms with D do not have to be in the model. The hierarchical structure applies to nesting as well. If B (A) is in the model, then A must be also.

Because models can be quite long and tedious to type, two shortcuts have been provided. A vertical bar indicates crossed factors, and a minus sign removes terms.

Long form	**Short form**
A B C A ∗ B A ∗ C B ∗ C A ∗ B ∗ C	A | B | C
A B C A ∗ B A ∗ C B ∗ C	A | B | C – A ∗ B ∗ C
A B C B ∗ C E	A B | C E
A B C D A ∗ B A ∗ C A ∗ D B ∗ C B ∗ D C ∗ D A ∗ B ∗ D A ∗ C ∗ D B ∗ C ∗ D	A | B | C | D – A ∗ B ∗ C – A ∗ B ∗ C ∗ D
A B (A) C A ∗ C B ∗ C	A | B (A) | C

In general, all crossings are done for factors separated by bars unless the cross results in an illegal term. For example, in the last example, the potential term A ∗ B (A) is illegal and Minitab automatically omits it. If a factor is nested, you must indicate this when using the vertical bar, as in the last example with the term B (A).

Using patterned data to set up factor levels

Minitab's set patterned data capability can be helpful when entering numeric factor levels. For example, to enter the level values for a three-way crossed design with a, b, and c (a, b, and c represent numbers) levels of factors A, B, C, and n observations per cell, fill out the Calc > Set Patterned Data > Simple Set of Numbers dialog box and execute 3 times, once for each factor, as shown:

		Factor	
Dialog item	**A**	**B**	**C**
From first value	1	1	1
From last value	a	b	c
List each value	bcn	cn	n
List the whole sequence	1	a	ab

Coefficients in general linear models

General Linear Model (GLM) uses a regression approach to fit your model. First, GLM codes the factor levels as dummy or indicator variables using a 1, 0, – 1, coding scheme. For more information on how Minitab codes the data for a GLM analysis, see Design matrix used by General Linear Model. The dummy variables are then used to calculate the coefficients for all terms. In GLM, the coefficients represent the distance between factor levels and the overall mean.

You can view the ANOVA model equation by displaying the table of coefficients for all terms in the GLM output. In the Results subdialog box, choose **In addition, coefficients for all terms**.

After you conduct the analysis, you will notice that coefficients are listed for all but one of the levels for each factor. This level is the reference level or baseline. All estimated coefficients are interpreted relative to the reference level. In some cases, you may want to know the reference level coefficient to understand how the reference value compares in size and direction to the overall mean.

Suppose you perform a general linear model test with 2 factors. Factor 1 has 3 levels (A, B, and C), and Factor 2 has 2 levels (High and Low). Minitab codes these levels using indicator variables. For factor 1: A = 1, B = 0, and C = – 1. For Factor 2: High = 1 and Low = – 1.

You obtain the following table of coefficients:

```
Term          Coef  SE Coef       T      P
Constant    5.0000   0.1954   25.58  0.000
Factor1
A          -3.0000   0.2764  -10.85  0.000
B          -0.5000   0.2764   -1.81  0.108
Factor2
High       -0.8333   0.1954   -4.26  0.003
```

The ANOVA model is: Response = 5.0 – 3.0 ∗ A – 0.5 ∗ B – 0.833 ∗ High

Notice that the table does not include the coefficients for C (Factor 1) or Low (Factor 2), which are the reference levels for each factor. However, you can easily calculate these values by subtracting the overall mean from each level mean. The constant term is the overall mean. Use Stat > Basic Statistics > Display Descriptive Statistics to obtain the mean for each level. The means are:

Overall	5.0
A (Factor 1)	2.0
B (Factor 1)	4.5
C (Factor 1)	8.5
High (Factor 2)	4.1667
High (Factor 2)	5.8333

The coefficients are calculated as the level mean – overall mean, Thus, the coefficients for each level are:

Level A effect = 2.0 – 5.0 = – 3.0

Level B effect = 4.5 – 5.0 = – 0.5

Level C effect = 8.5 – 5.0 = 3.5 (not given in the coefficients table)

Level High effect = 4.1667 – 5.0 = – 0.8333

Level Low effect = 5.8333 – 5.0 = 0.8333 (not given in the coefficients table)

Tip A quick way to obtain the coefficients not listed in the table is by adding all of the level coefficients for a factor (excluding the intercept) and multiplying by – 1. For example, the coefficient for Level C = – 1 * [(– 3.0) + (– 0.50)] = 3.5.

If you add a covariate or have unequal sample sizes within each group, coefficients are based on weighted means for each factor level rather than the arithmetic mean (sum of the observations divided by n).

General Linear Model – Covariates

Stat > ANOVA > General Linear Model > Covariates

Enter covariates into the model. These are entered into the model first by default.

Dialog box items

Covariates: Enter columns containing the covariates.

Specifying terms involving covariates

You can specify variables to be covariates in GLM. You must specify the covariates in *Covariates*, but you can enter the covariates in *Model***,** though this is not necessary unless you cross or nest the covariates (see table below).

In an unbalanced design or a design involving covariates, GLM's sequential sums of squares (the additional model sums of squares explained by a variable) will depend upon the order in which variables enter the model. If you do not enter the covariates in *Model* when using GLM, they will be fit first, which is what you usually want when a covariate contributes background variability. The subsequent order of fitting is the order of terms in *Model*. The sequential sums of squares for unbalanced terms A B will be different depending upon the order that you enter them in the model. The default adjusted sums of squares (sums of squares with all other terms in the model), however, will be the same, regardless of model order.

GLM allows terms containing covariates crossed with each other and with factors, and covariates nested within factors. Here are some examples of these models, where A is a factor.

Case	Covariates	Terms in model
test homogeneity of slopes (covariate crossed with factor)	X	A X A∗X
same as previous	X	A | X
quadratic in covariate (covariate crossed with itself)	X	A X X∗X
full quadratic in two covariates (covariates crossed)	X Z	A X Z X∗X Z∗Z X∗Z
separate slopes for each level of A (covariate nested within a factor)	X	A X (A)

General Linear Model – Options

Stat > ANOVA > General Linear Model > Options

Allows you to choose a weighted fit and the sums of squares type used in the ANOVA.

Dialog box items

Do a weighted fit, using weights in: Enter a column of weights for a weighted fit. See Weighted regression for more information.

Sum of Squares: Select a sums of squares for calculating F-and p-values.

Adjusted (Type III): Choose if you want sums of squares for terms with other terms in the model

Sequential (Type I): Choose if you want sums of squares with only previous terms in the model

Adjusted vs. sequential sums of squares

Minitab by default uses adjusted (Type III) sums of squares for all GLM calculations. Adjusted sums of squares are the additional sums of squares determined by adding each particular term to the model given the other terms are already in the model. You also have the choice of using sequential (Type I) sums of squares in all GLM calculations. Sequential sums of squares are the sums of squares added by a term with only the previous terms entered in the model. These sums of squares can differ when your design is unbalanced or if you have covariates. Usually, you would probably use adjusted sums of squares. However, there may be cases where you might want to use sequential sums of squares.

General Linear Model – Comparisons

Stat > ANOVA > General Linear Model > Comparisons

Specify terms for comparing the means, as well as the type of multiple comparisons.

Dialog box items

Pairwise comparisons: Choose to obtain pairwise comparison of all mean for designated terms.

Comparisons with a control: Choose to obtain comparisons of means with the mean of a control level.

Terms: Enter the model terms for comparison.

Control levels: Enter the control level if you chose comparisons with a control. (IMPORTANT: For text variables, you must enclose factor levels in double quotes, even if there are no spaces in them.)

Method: Select the multiple comparison method(s). See Multiple comparisons.

Tukey: Choose the Tukey (also call Tukey-Kramer method in unbalanced case) method.

Dunnett: Choose the Dunnett method.

Bonferroni: Choose the Bonferroni method.

Sidak: Choose the Sidak method.

Alternative: Choose one of three possible alternative hypotheses when you choose comparisons with a control. The null hypothesis is equality of treatment and control means.

Less than: Choose the alternative hypothesis of the treatment mean being less than the mean of the control group.

Not equal: Choose the alternative hypothesis of the treatment mean being not equal to the mean of the control group.

Greater than: Choose the alternative hypothesis of the treatment mean being greater than the mean of the control group.

Confidence interval, with confidence level: Check to specify a confidence level and then enter a value for the intervals that is between 0 and 100 (the default is 95%).

Test: Check to select the hypothesis test form of multiple comparison output.

Multiple comparisons of means

Multiple comparisons of means allow you to examine which means are different and to estimate by how much they are different. When you have multiple factors, you can obtain multiple comparisons of means through GLM's Comparisons subdialog box.

There are some common pitfalls to the use of multiple comparisons. If you have a quantitative factor you should probably examine linear and higher order effects rather than performing multiple comparisons (see [12] and Example of using GLM to fit linear and quadratic effects). In addition, performing multiple comparisons for those factors which appear to have the greatest effect or only those with a significant F-test can result in erroneous conclusions (see Which means to compare? below).

You have the following choices when using multiple comparisons:

- Pairwise comparisons or comparisons with a control
- Which means to compare
- The method of comparison
- Display comparisons in confidence interval or hypothesis test form
- The confidence level, if you choose to display confidence intervals
- The alternative, if you choose comparisons with a control

Following are some guidelines for making these choices.

Pairwise comparisons or comparison with a control

Choose **Pairwise Comparisons** when you do not have a control level but you would like to examine which pairs of means are different.

Choose **Comparisons with a Control** when you are comparing treatments to a control. When this method is suitable, it is inefficient to use the all-pairwise approach, because the all-pairwise confidence intervals will be wider and the hypothesis tests less powerful for a given family error rate. If you do not specify a level that represents the control, Minitab will assume that the lowest level of the factors is the control. If you wish to change which level is the control, specify a level that represents the control for each term that you are comparing the means of. If these levels are text or date/time, enclose each with double quotes.

Which means to compare

Choosing which means to compare is an important consideration when using multiple comparisons; a poor choice can result in confidence levels that are not what you think. Issues that should be considered when making this choice might include: 1) should you compare the means for only those terms with a significant F-test or for those sets of means for which differences appear to be large? 2) how deep into the design should you compare means–only within each factor, within each combination of first-level interactions, or across combinations of higher level interactions?

It is probably a good idea to decide which means you will compare before collecting your data. If you compare only those means with differences that appear to be large, which is called data snooping, then you are increasing the likelihood that the results suggest a real difference where no difference exists [9], [19]. Similarly, if you condition the application of multiple comparisons upon achieving a significant F-test, then the error rate of the multiple comparisons can be higher than the error rate in the unconditioned application of multiple comparisons [9], [15]. The multiple comparison methods have protection against false positives already built in.

In practice, however, many people commonly use F-tests to guide the choice of which means to compare. The ANOVA F-tests and multiple comparisons are not entirely separate assessments. For example, if the p-value of an F-test is 0.9, you probably will not find statistically significant differences among means by multiple comparisons.

How deep within the design should you compare means? There is a trade-off: if you compare means at all two-factor combinations and higher orders turn out to be significant, then the means that you compare might be a mix of effects; if you compare means at too deep a level, you lose power because the sample sizes become smaller and the number of comparisons become larger. In practice, you might decide to compare means for factor level combinations for which you believe the interactions are meaningful.

Minitab restricts the terms that you can compare means for to fixed terms or interactions among fixed terms. Nesting is considered to be a form of interaction.

To specify which means to compare, enter terms from the model in the **Terms** box. If you have 2 factors named A and B, entering A B will result in multiple comparisons within each factor. Entering A ∗ B will result in multiple comparisons for all level combination of factors A and B. You can use the notation A | B to indicate interaction for pairwise comparisons but not for comparisons with a control.

The multiple comparison method

You can choose from among three methods for both pairwise comparisons and comparisons with a control. Each method provides simultaneous or joint confidence intervals, meaning that the confidence level applies to the set of intervals computed by each method and not to each one individual interval. By protecting against false positives with multiple comparisons, the intervals are wider than if there were no protection.

The Tukey (also called Tukey-Kramer in the unbalanced case) and Dunnett methods are extensions of the methods used by one-way ANOVA. The Tukey approximation has been proven to be conservative when comparing three means. "Conservative" means that the true error rate is less than the stated one. In comparing larger numbers of means, there is no proof that the Tukey method is conservative for the general linear model. The Dunnett method uses a factor analytic method to approximate the probabilities of the comparisons. Because it uses the factor analytic approximation, the Dunnett method is not generally conservative. The Bonferroni and Sidak methods are conservative methods based upon probability inequalities. The Sidak method is slightly less conservative than the Bonferroni method.

Some characteristics of the multiple comparison methods are summarized below:

Comparison method	Properties
Dunnett	comparison to a control only, not proven to be conservative
Tukey	all pairwise differences only, not proven to be conservative
Bonferroni	most conservative
Sidak	conservative, but slightly less so than Bonferroni

Display of comparisons in confidence interval or hypothesis test form

Minitab presents multiple comparison results in confidence interval and/or hypothesis test form. Both are given by default.

When viewing confidence intervals, you can assess the practical significance of differences among means, in addition to statistical significance. As usual, the null hypothesis of no difference between means is rejected if and only if zero is not contained in the confidence interval. When you request confidence intervals, you can specify family confidence levels for the confidence intervals. The default level is 95%.

Minitab calculates adjusted p-values for hypothesis test statistics. The adjusted p-value for a particular hypothesis within a collection of hypotheses is the smallest family wise a level at which the particular hypothesis would be rejected.

General Linear Model – Graphs

Stat > ANOVA > General Linear Model > Graphs

Displays residual plots. You do not have to store the residuals and fits in order to produce these plots.

Dialog box items

Residuals for Plots: You can specify the type of residual to display on the residual plots.

- **Regular:** Choose to plot the regular or raw residuals.
- **Standardized:** Choose to plot the standardized residuals.
- **Deleted:** Choose to plot the Studentized deleted residuals.

Residual Plots

- **Individual plots:** Choose to display one or more plots.
  - **Histogram of residuals:** Check to display a histogram of the residuals.
  - **Normal plot of residuals:** Check to display a normal probability plot of the residuals.
  - **Residuals versus fits:** Check to plot the residuals versus the fitted values.
  - **Residuals versus order:** Check to plot the residuals versus the order of the data. The row number for each data point is shown on the x-axis–for example, 1 2 3 4... n.

Four in one: Choose to display a layout of a histogram of residuals, a normal plot of residuals, a plot of residuals versus fits, and a plot of residuals versus order.

Residuals versus the variables: Enter one or more columns containing the variables against which you want to plot the residuals. Minitab displays a separate graph for each column.

General Linear Model – Results

Stat > ANOVA > General Linear Model > Results

Control the display of results to the Session window, the display of expected means squares and variance components, and display means of model term levels.

Dialog box items

Display of Results

Display nothing: Choose to display nothing.

Analysis of variance table: Choose to display only the analysis of variance table.

In addition, coefficient for covariate terms and table of unusual observations: Choose to display, in addition to the ANOVA table, a table of covariate term coefficients and a table of unusual observations.

In addition, coefficient for all terms: Choose to display, in addition to the above tables, the coefficients for all terms.

Display expected mean squares and variance components: Check to display the expected mean squares and variance component estimates for random terms.

Display means corresponding to the terms: Enter the model terms for which to display least squares means and their standard errors.

General Linear Model – Storage

Stat > ANOVA > General Linear Model > Storage

Stores the residuals, fitted values, and many other diagnostics for further analysis (see Checking your model).

Dialog box items

Diagnostic Measures

Residuals: Check to store the residuals.

Standardized residuals: Check to store the standardized residuals.

Deleted t residuals: Check to store Studentized residuals.

Hi [leverage]: Check to store leverages.

Cook's distance: Check to store Cook's distance.

DFITS: Check to store DFITS.

Characteristics of Estimated Equation

Coefficients: Check to store the coefficients for a model that corresponds to the design matrix. (If M1 contains the design matrix and C1 the coefficients, then M1 times C1 gives the fitted values.)

Fits: Check to store the fitted values.

Design matrix: Check to store the design matrix corresponding to your model.

General Linear Model – Factorial Plots

Stat > ANOVA > General Linear Model > Factor Plots

Displays plots of the main effects and interactions in your data.

Dialog box items

Main Effects Plot: Display plots of main effects.

Factors: Chose the factors to plot.

Minimum for Y (response) scale: Replace the default minimum value for the Y-axis with one you chose.

Maximum for Y (response) scale: Replace the default maximum value for the Y-axis with one you chose.

Title: Replace the default title with one of your own.

Interactions Plot: Display plots of two-way interactions.

Factors: Choose the factors to include in the plot(s).

Display full interaction plot matrix: By default, Minitab chooses one factor to represent on the X-axis and represents the levels of the other factor with different symbols and lines. Check this option to display each interaction twice, once with each factor represented on the X-axis.

Minimum for Y (response) scale: Replace the default minimum value for the Y-axis with one you chose.

Maximum for Y (response) scale: Replace the default maximum value for the Y-axis with one you chose.

Title: Replace the default title with one of your own.

Example of Using GLM to fit Linear and Quadratic Effects

An experiment is conducted to test the effect of temperature and glass type upon the light output of an oscilloscope. There are three glass types and three temperature levels: 100, 125, and 150 degrees Fahrenheit. These factors are fixed because we are interested in examining the response at those levels. The example and data are from Montgomery [14], page 252.

When a factor is quantitative with three or more levels it is appropriate to partition the sums of squares from that factor into effects of polynomial orders [11]. If there are k levels to the factor, you can partition the sums of squares into k-1 polynomial orders. In this example, the effect due to the quantitative variable temperature can be partitioned into linear and quadratic effects. Similarly, you can partition the interaction. To do this, you must code the quantitative variable with the actual treatment values (that is, code Temperature levels as 100, 125, and 150), use GLM to analyze your data, and declare the quantitative variable to be a covariate.

1 Open the worksheet EXH_AOV.MTW.

2 Choose **Stat > ANOVA > General Linear Model**.

3 In **Responses**, enter *LightOutput*.

4 In **Model**, type *Temperature Temperature ∗Temperature GlassType GlassType ∗Temperature GlassType ∗ Temperature ∗Temperature*.

5 Click **Covariates**. In **Covariates**, enter *Temperature*.

6 Click **OK** in each dialog box.

Session window output

General Linear Model: LightOutput versus GlassType

```
Factor     Type   Levels  Values
GlassType  fixed       3  1, 2, 3


Analysis of Variance for LightOutput, using Adjusted SS for Tests

Source                                DF   Seq SS  Adj SS  Adj MS       F      P
Temperature                            1  1779756  262884  262884  719.21  0.000
Temperature*Temperature                1   190579  190579  190579  521.39  0.000
GlassType                              2   150865   41416   20708   56.65  0.000
GlassType*Temperature                  2   226178   51126   25563   69.94  0.000
GlassType*Temperature*Temperature      2    64374   64374   32187   88.06  0.000
Error                                 18     6579    6579     366
Total                                 26  2418330


S = 19.1185   R-Sq = 99.73%   R-Sq(adj) = 99.61%


Term                                    Coef  SE Coef       T      P
Constant                             -4968.8    191.3  -25.97  0.000
Temperature                           83.867    3.127   26.82  0.000
Temperature*Temperature             -0.28516  0.01249  -22.83  0.000
Temperature*GlassType
           1                         -24.400    4.423   -5.52  0.000
           2                         -27.867    4.423   -6.30  0.000
Temperature*Temperature*GlassType
                        1            0.11236  0.01766    6.36  0.000
                        2            0.12196  0.01766    6.91  0.000


Unusual Observations for LightOutput
```

```
Obs  LightOutput      Fit  SE Fit  Residual  St Resid
 11      1070.00  1035.00   11.04     35.00      2.24 R
 17      1000.00  1035.00   11.04    -35.00     -2.24 R

R denotes an observation with a large standardized residual.
```

Interpreting the results

Minitab first displays a table of factors, with their number of levels, and the level values. The second table gives an analysis of variance table. This is followed by a table of coefficients, and then a table of unusual observations.

The Analysis of Variance table gives, for each term in the model, the degrees of freedom, the sequential sums of squares (Seq SS), the adjusted (partial) sums of squares (Adj SS), the adjusted means squares (Adj MS), the F-statistic from the adjusted means squares, and its p-value. The sequential sums of squares is the added sums of squares given that prior terms are in the model. These values depend upon the model order. The adjusted sums of squares are the sums of squares given that all other terms are in the model. These values do not depend upon the model order. If you had selected sequential sums of squares in the Options subdialog box, Minitab would use these values for mean squares and F-tests.

In the example, all p-values were printed as 0.000, meaning that they are less than 0.0005. This indicates significant evidence of effects if your level of significance, a, is greater than 0.0005. The significant interaction effects of glass type with both linear and quadratic temperature terms implies that the coefficients of second order regression models of the effect of temperature upon light output depends upon the glass type.

The next table gives the estimated coefficients for the covariate, Temperature, and the interactions of Temperature with GlassType, their standard errors, t-statistics, and p-values. Following the table of coefficients is a table of unusual values. Observations with large standardized residuals or large leverage values are flagged. In our example, two values have standardized residuals whose absolute values are greater than 2.

Example of Using GLM and Multiple Comparisons with an Unbalanced Nested Design

Four chemical companies produce insecticides that can be used to kill mosquitoes, but the composition of the insecticides differs from company to company. An experiment is conducted to test the efficacy of the insecticides by placing 400 mosquitoes inside a glass container treated with a single insecticide and counting the live mosquitoes 4 hours later. Three replications are performed for each product. The goal is to compare the product effectiveness of the different companies. The factors are fixed because you are interested in comparing the particular brands. The factors are nested because each insecticide for each company is unique. The example and data are from Milliken and Johnson [12], page 414. You use GLM to analyze your data because the design is unbalanced and you will use multiple comparisons to compare the mean response for the company brands.

1 Open the worksheet EXH_AOV.MTW.

2 Choose **Stat > ANOVA > General Linear Model**.

3 In **Responses**, enter *NMosquito*.

4 In **Model**, type *Company Product(Company)*.

5 Click **Comparisons**. Under **Pairwise Comparisons**, enter *Company* in **Terms**.

6 Under **Method**, check **Tukey**. Click **OK** in each dialog box.

Session window output

General Linear Model: NMosquito versus Company, Product

```
Factor            Type   Levels  Values
Company           fixed       4  A, B, C, D
Product(Company)  fixed      11  A1, A2, A3, B1, B2, C1, C2, D1, D2, D3, D4


Analysis of Variance for NMosquito, using Adjusted SS for Tests

Source            DF   Seq SS   Adj SS  Adj MS       F      P
Company            3  22813.3  22813.3  7604.4  132.78  0.000
Product(Company)   7   1500.6   1500.6   214.4    3.74  0.008
Error             22   1260.0   1260.0    57.3
Total             32  25573.9


S = 7.56787   R-Sq = 95.07%   R-Sq(adj) = 92.83%
```

```
Tukey 95.0% Simultaneous Confidence Intervals
Response Variable NMosquito
All Pairwise Comparisons among Levels of Company
Company = A  subtracted from:

Company   Lower  Center   Upper  --------+---------+---------+--------
B         -2.92    8.17   19.25                             (---*----)
C        -52.25  -41.17  -30.08         (----*---)
D        -61.69  -52.42  -43.14     (---*---)
                                 --------+---------+---------+--------
                                       -50       -25         0


Company = B  subtracted from:

Company   Lower  Center   Upper  --------+---------+---------+--------
C        -61.48  -49.33  -37.19     (----*----)
D        -71.10  -60.58  -50.07  (---*---)
                                 --------+---------+---------+--------
                                       -50       -25         0


Company = C  subtracted from:

Company   Lower  Center    Upper  --------+---------+---------+--------
D        -21.77  -11.25  -0.7347                     (----*---)
                                  --------+---------+---------+--------
                                        -50       -25         0



Tukey Simultaneous Tests
Response Variable NMosquito
All Pairwise Comparisons among Levels of Company
Company = A  subtracted from:

         Difference       SE of           Adjusted
Company    of Means  Difference  T-Value   P-Value
B              8.17       3.989     2.05    0.2016
C            -41.17       3.989   -10.32    0.0000
D            -52.42       3.337   -15.71    0.0000


Company = B  subtracted from:

         Difference       SE of           Adjusted
Company    of Means  Difference  T-Value   P-Value
C            -49.33       4.369   -11.29    0.0000
D            -60.58       3.784   -16.01    0.0000


Company = C  subtracted from:

         Difference       SE of           Adjusted
Company    of Means  Difference  T-Value   P-Value
D            -11.25       3.784   -2.973    0.0329
```

Interpreting the results

Minitab displays a factor level table, an ANOVA table, multiple comparison confidence intervals for pairwise differences between companies, and the corresponding multiple comparison hypothesis tests. The ANOVA F-tests indicate that there is significant evidence for company effects.

Examine the multiple comparison confidence intervals. There are three sets: 1) for the company A mean subtracted from the company B, C, and D means; 2) for the company B mean subtracted from the company C and D means; and 3) for the company C mean subtracted from the company D mean. The first interval, for the company B mean minus the company A mean, contains zero is in the confidence interval. Thus, there is no significant evidence at $\alpha = 0.05$ for differences in means. However, there is evidence that all other pairs of means are different, because the confidence intervals for the differences in means do not contain zero. An advantage of confidence intervals is that you can see the magnitude of the differences between the means.

Examine the multiple comparison hypothesis tests. These are laid out in the same way as the confidence intervals. You can see at a glance the mean pairs for which there is significant evidence of differences. The adjusted p-values are small for all but one comparison, that of company A to company B. An advantage of hypothesis tests is that you can see what a level would be required for significant evidence of differences.

Fully Nested ANOVA

Fully Nested ANOVA

Stat > ANOVA > Fully Nested ANOVA

Use to perform fully nested (hierarchical) analysis of variance and to estimate variance components for each response variable. All factors are implicitly assumed to be random. Minitab uses sequential (Type I) sums of squares for all calculations.

You can analyze up to 50 response variables with up to 9 factors at one time.

If your design is not hierarchically nested or if you have fixed factors, use either Balanced ANOVA or GLM. Use GLM if you want to use adjusted sums of squares for a fully nested model.

Dialog box items

Responses: Enter the columns containing your response variables.

Factors: Enter the columns containing the factors in hierarchical order.

Data – Fully Nested ANOVA

Set up your worksheet in the same manner as with Balanced ANOVA or GLM: one column for each response variable and one column for each factor, so that there is one row for each observation. The factor columns may be numeric, text, or date/time. If you wish to change the order in which text categories are processed from their default alphabetical order, you can define your own order. See Ordering Text Categories.

Nesting does not need to be balanced. A nested factor must have at least 2 levels at some level of the nesting factor. If factor B is nested within factor A, there can be unequal levels of B within each level of A. In addition, the subscripts used to identify the B levels can differ within each level of A.

If any response or factor column contains missing data, that entire observation (row) is excluded from all computations. If an observation is missing for one response variable, that row is eliminated for all responses. If you want to eliminate missing rows separately for each response, perform a fully nested ANOVA separately for each response.

You can analyze up to 50 response variables with up to 9 factors at one time.

To perform an analysis using fully nested ANOVA

1 Choose **Stat > ANOVA > Fully Nested ANOVA**.

2 In **Responses**, enter up to 50 numeric columns containing the response variables.

3 In **Factors**, type in the factors in hierarchical order. See Fully Nested or Hierarchical Models.

4 Click **OK**.

Fully Nested or Hierarchical Models

Minitab fits a fully nested or hierarchical model with the nesting performed according to the order of factors in the **Factors** box. If you enter factors A B C, then the model terms will be A B(A) C(B). You do not need to specify these terms in model form as you would for Balanced ANOVA or GLM.

Minitab uses sequential (Type I) sums of squares for all calculations of fully nested ANOVA. This usually makes sense for a hierarchical model. General Linear Models (GLM) offers the choice of sequential or adjusted (Type III) sums of squares and uses the adjusted sums of squares by default. These sums of squares can differ when your design is unbalanced. Use GLM if you want to use adjusted sums of squares for calculations.

Example of Fully Nested ANOVA

You are an engineer trying to understand the sources of variability in the manufacture of glass jars. The process of making the glass requires mixing materials in small furnaces for which the temperature setting is to be 475° F. Your company has a number of plants where the jars are made, so you select four as a random sample. You conduct an experiment and measure furnace temperature for four operators over four different shifts. You take two batch measurements during each shift. Because your design is fully nested, you use Fully Nested ANOVA to analyze your data.

1 Open the worksheet FURNTEMP.MTW.

2 Choose **Stat > ANOVA > Fully Nested ANOVA**.

3 In **Responses**, enter *Temp*.

4 In **Factors**, enter *Plant - Batch*. Click **OK**.

Session window output

Nested ANOVA: Temp versus Plant, Operator, Shift, Batch

```
Analysis of Variance for Temp

Source      DF         SS        MS      F      P
Plant        3   731.5156  243.8385  5.854  0.011
Operator    12   499.8125   41.6510  1.303  0.248
Shift       48  1534.9167   31.9774  2.578  0.000
Batch      128  1588.0000   12.4062
Total      191  4354.2448


Variance Components

                          % of
Source     Var Comp.  Total  StDev
Plant          4.212  17.59  2.052
Operator       0.806   3.37  0.898
Shift          6.524  27.24  2.554
Batch         12.406  51.80  3.522
Total         23.948         4.894


Expected Mean Squares

1  Plant       1.00(4) +  3.00(3) + 12.00(2) + 48.00(1)
2  Operator    1.00(4) +  3.00(3) + 12.00(2)
3  Shift       1.00(4) +  3.00(3)
4  Batch       1.00(4)
```

Interpreting the results

Minitab displays three tables of output: 1) the ANOVA table, 2) the estimated variance components, and 3) the expected means squares. There are four sequentially nested sources of variability in this experiment: plant, operator, shift, and batch. The ANOVA table indicates that there is significant evidence for plant and shift effects at α = 0.05 (F-test p-values < 0.05). There is no significant evidence for an operator effect. The variance component estimates indicate that the variability attributable to batches, shifts, and plants was 52, 27, and 18 percent, respectively, of the total variability.

If a variance component estimate is less than zero, Minitab displays what the estimate is, but sets the estimate to zero in calculating the percent of total variability.

Balanced MANOVA

Balanced MANOVA

Stat > ANOVA > Balanced MANOVA

Use balanced MANOVA to perform multivariate analysis of variance (MANOVA) for balanced designs. You can take advantage of the data covariance structure to simultaneously test the equality of means from different responses.

Your design must be balanced, with the exception of one-way designs. *Balanced* means that all treatment combinations (cells) must have the same number of observations. Use General MANOVA to analyze either balanced and unbalanced MANOVA designs or if you have covariates. You cannot designate factors to be random with general MANOVA, unlike for balanced ANOVA, though you can work around this restriction by supplying error terms to test the model terms.

Factors may be crossed or nested, fixed or random.

Dialog box items

Responses: Enter up to 50 numeric columns containing the response variables

Model: Type the model terms that you want to fit.

Random Factors: Enter which factors are random factors.

<Options>

<Graphs>

<Results>

<Storage>

Data – Balanced MANOVA

You need one column for each response variable and one column for each factor, with each row representing an observation. Regardless of whether factors are crossed or nested, use the same form for the data. Factor columns may be numeric, text, or date/time. If you wish to change the order in which text categories are processed from their default alphabetical order, you can define your own order. See Ordering Text Categories. You may include up to 50 response variables and up to 31 factors at one time.

Balanced data are required except for one-way designs. The requirement for balanced data extends to nested factors as well. Suppose A has 3 levels, and B is nested within A. If B has 4 levels within the first level of A, B must have 4 levels within the second and third levels of A. Minitab will tell you if you have unbalanced nesting. In addition, the subscripts used to indicate the 4 levels of B within each level of A must be the same. Thus, the four levels of B cannot be (1 2 3 4) in level 1 of A, (5 6 7 8) in level 2 of A, and (9 10 11 12) in level 3 of A. You can use general MANOVA if you have different levels of B within the levels of A.

If any response or factor column specified contains missing data, that entire observation (row) is excluded from all computations. The requirement that data be balanced must be preserved after missing data are omitted.

To perform a balanced MANOVA

1 Choose **Stat > ANOVA > Balanced MANOVA**.

2 In **Responses**, enter up to 50 numeric columns containing the response variables.

3 In **Model**, type the model terms that you want to fit. See Overview of Balanced ANOVA and GLM.

4 If you like, use any dialog box options, then click **OK**.

Balanced MANOVA – Options

Stat > ANOVA > Balanced MANOVA > Options

Dialog box items

Use the restricted form of the model: Check to use the restricted form of the mixed models (both fixed and random effects). The restricted model forces mixed interaction effects to sum to zero over the fixed effects. By default, Minitab fits the unrestricted model.

Balanced MANOVA – Graphs

Stat > ANOVA > Balanced MANOVA > Graphs

Displays residual plots. You do not have to store the residuals and fits in order to produce these plots.

Dialog box items

Residual Plots

Individual plots: Choose to display one or more plots.

Histogram of residuals: Check to display a histogram of the residuals.

Normal plot or residuals: Check to display a normal probability plot of the residuals.

Residuals versus fits: Check to plot the residuals versus the fitted values.

Residuals versus order: Check to plot the residuals versus the order of the data. The row number for each data point is shown on the x-axis--for example, 1 2 3 4... n.

Four in one: Choose to display a layout of a histogram of residuals, a normal plot of residuals, a plot of residuals versus fits, and a plot of residuals versus order.

Residuals versus the variables: Enter one or more columns containing the variables against which you want to plot the residuals. Minitab displays a separate graph for each column.

Balanced MANOVA – Results

Stat > ANOVA > Balanced MANOVA > Results

You can control the Session window output.

Dialog box items

Display of Results

Matrices (hypothesis, error, partial correlations): Check to display the hypothesis matrix H, the error matrix E, and a matrix of partial correlations. See MANOVA tests.

Eigen analysis: Check to display the eigenvalues and eigenvalues for the matrix E**-1 H.

Univariate analysis of variance: Check to perform a univariate analysis of variance for each response variable.

Expected mean squares for univariate analysis: Check to display the expected mean squares when you have requested univariate analysis of variance.

Display means corresponding to the terms: Display a table of means corresponding to specified terms from the model. For example, if you specify A B D A ∗ B ∗ D, four table of means will be printed, one for each main effect, A, B, D, and one for the three-way interaction, A ∗ B ∗ D.

Custom multivariate tests for the following terms: Perform 4 multivariate tests for model terms that you specify. See Specifying terms to test. Default tests are performed for all model terms.

Error: Designate an error term for the four multivariate tests. It must be a single term that is in the model. If you do not specify an error term, Minitab uses the error associated with mean squares error, as in the univariate case.

Specifying terms to test – Balanced MANOVA

In the **Results** subdialog box, you can specify model terms in **Custom multivariate test for the following terms** and designate an error term in **Error** and Minitab will perform four multivariate tests for those terms. This option is probably less useful for balanced MANOVA than it is for general MANOVA; because you can specify factors to be random with balanced MANOVA, Minitab will use the correct error terms. This option exists for special purpose tests.

If you specify an error term, it must be a single term that is in the model. This error term is used for all requested tests. If you do not specify an error term, Minitab determines an appropriate error term.

MANOVA tests – Balanced MANOVA

Minitab automatically performs four multivariate tests–Wilks' test, Lawley-Hotelling test, Pillai's test, and Roy's largest root test–for each term in the model and for specially requested terms (see Specifying terms to test). All four tests are based on two SSCP (sums of squares and cross products) matrices: H, the hypothesis matrix and E, the error matrix. There is one H associated with each term. E is the matrix associated with the error for the test. These matrices are displayed when you request the hypothesis matrices and are labeled by SSCP Matrix.

The test statistics can be expressed in terms of either H and/or E or the eigenvalues of E**-1 H. You can request to have these eigenvalues printed. (If the eigenvalues are repeated, corresponding eigenvectors are not unique and in this case, the eigenvectors Minitab prints and those in books or other software may not agree. The MANOVA tests, however, are always unique.)

You can also display the matrix of partial correlations, which are the correlations among the residuals, or alternatively, the correlations among the responses conditioned on the model. The formula for this matrix is W**-.5 E W**-.5, where E is the error matrix and W has the diagonal of E as its diagonal and 0's off the diagonal.

Hotelling's T-Squared Test

Hotelling's T-squared test to compare the mean vectors of two groups is a special case of MANOVA, using one factor that has two levels. Minitab's MANOVA option can be used to do this test. The usual T-squared test statistic can be calculated from Minitab's output using the relationship T-squared = (N-2) U, where N is the total number of observations and U is the Lawley-Hotelling trace. S, the pooled covariance matrix, is E / (N-2), where E is the error matrix.

Balanced MANOVA – Storage

Stat > ANOVA > Balanced MANOVA > Storage

Store fits and residuals for each response. If you fit a full model, fits are cell means. If you fit a reduced model, fits are least squares estimates.

Dialog box items

Fits: Check to store the fitted values for each observation in the data set in the next available columns, using one column for each response.

Residuals: Check to store the residuals using one column for each response.

Testing the equality of means from multiple responses

Balanced MANOVA and general MANOVA are procedures for testing the equality of vectors of means from multiple responses. Your choice between these two procedures depends upon the experimental design and the available options. Both procedures can fit MANOVA models to balanced data with up to 31 factors.

- Balanced MANOVA is used to perform multivariate analysis of variance with balanced designs. See Balanced designs. You can also specify factors to be random and obtain expected means squares. Use general MANOVA with unbalanced designs.
- General MANOVA is used to perform multivariate analysis of variance with either balanced or unbalanced designs that can also include covariates. You cannot specify factors to be random as you can for balanced MANOVA, although you can work around this restriction by specifying the error term for testing different model terms.

The table below summarizes the differences between Balanced and General MANOVA:

	Balanced	**General**
	MANOVA	**MANOVA**
Can fit unbalanced data	no	yes
Can specify factors as random and obtain expected means squares	yes	no
Can fit covariates	no	yes
Can fit restricted and unrestricted forms of a mixed model	yes	no; unrestricted only

Example of Balanced MANOVA

You perform a study in order to determine optimum conditions for extruding plastic film. You measure three responses–tear resistance, gloss, and opacity–five times at each combination of two factors–rate of extrusion and amount of an additive–each set at low and high levels. The data and example are from Johnson and Wichern [5], page 266. You use Balanced MANOVA to test the equality of means because the design is balanced.

1 Open the file EXH_MVAR.MTW.

2 Choose **Stat > ANOVA > Balanced MANOVA**.

3 In **Responses**, enter *Tear Gloss Opacity*.

4 In **Model**, enter *Extrusion* | *Additive*.

5 Click **Results**. Under **Display of Results**, check **Matrices (hypothesis, error, partial correlations)** and **Eigen analysis**.

6 Click **OK** in each dialog box.

Session window output

ANOVA: Tear, Gloss, Opacity versus Extrusion, Additive

```
MANOVA for Extrusion
s = 1    m = 0.5    n = 6.0

                       Test             DF
Criterion         Statistic      F  Num  Denom      P
Wilks'              0.38186  7.554    3     14  0.003
Lawley-Hotelling    1.61877  7.554    3     14  0.003
Pillai's            0.61814  7.554    3     14  0.003
Roy's               1.61877


SSCP Matrix for Extrusion

           Tear   Gloss  Opacity
Tear      1.740  -1.505   0.8555
Gloss    -1.505   1.301  -0.7395
```

```
Opacity   0.855  -0.739   0.4205


SSCP Matrix for Error

           Tear    Gloss  Opacity
Tear      1.764   0.0200   -3.070
Gloss     0.020   2.6280   -0.552
Opacity  -3.070  -0.5520   64.924


Partial Correlations for the Error SSCP Matrix

             Tear     Gloss   Opacity
Tear      1.00000   0.00929  -0.28687
Gloss     0.00929   1.00000  -0.04226
Opacity  -0.28687  -0.04226   1.00000


EIGEN Analysis for Extrusion


Eigenvalue  1.619  0.00000  0.00000
Proportion  1.000  0.00000  0.00000
Cumulative  1.000  1.00000  1.00000


Eigenvector        1       2        3
Tear          0.6541  0.4315   0.0604
Gloss        -0.3385  0.5163   0.0012

Opacity       0.0359  0.0302  -0.1209


MANOVA for Additive
s = 1    m = 0.5    n = 6.0

                       Test             DF
Criterion         Statistic      F  Num  Denom      P
Wilks'              0.52303  4.256    3     14  0.025
Lawley-Hotelling    0.91192  4.256    3     14  0.025
Pillai's            0.47697  4.256    3     14  0.025
Roy's               0.91192


SSCP Matrix for Additive

           Tear   Gloss  Opacity
Tear     0.7605  0.6825    1.931
Gloss    0.6825  0.6125    1.732
Opacity  1.9305  1.7325    4.901


EIGEN Analysis for Additive


Eigenvalue  0.9119  0.00000  0.00000
Proportion  1.0000  0.00000  0.00000
Cumulative  1.0000  1.00000  1.00000


Eigenvector        1        2        3
Tear         -0.6330   0.4480  -0.1276
Gloss        -0.3214  -0.4992  -0.1694
Opacity      -0.0684   0.0000   0.1102


MANOVA for Extrusion*Additive
s = 1    m = 0.5    n = 6.0

                       Test             DF
```

```
Criterion         Statistic      F  Num  Denom      P
Wilks'              0.77711  1.339    3     14  0.302
Lawley-Hotelling    0.28683  1.339    3     14  0.302
Pillai's            0.22289  1.339    3     14  0.302
Roy's               0.28683


SSCP Matrix for Extrusion*Additive

             Tear    Gloss  Opacity
Tear     0.000500  0.01650  0.04450
Gloss    0.016500  0.54450  1.46850
Opacity  0.044500  1.46850  3.96050


EIGEN Analysis for Extrusion*Additive


Eigenvalue  0.2868  0.00000  0.00000

Proportion  1.0000  0.00000  0.00000
Cumulative  1.0000  1.00000  1.00000


Eigenvector        1        2        3
Tear         -0.1364   0.1806   0.7527
Gloss        -0.5376  -0.3028  -0.0228
Opacity      -0.0683   0.1102  -0.0000
```

Interpreting the results

By default, Minitab displays a table of the four multivariate tests (Wilks', Lawley-Hotelling, Pillai's, and Roy's) for each term in the model. The values s, m, and n are used in the calculations of the F-statistics for Wilks', Lawley-Hotelling, and Pillai's tests. The F-statistic is exact if s = 1 or 2, otherwise it is approximate [6]. Because you requested the display of additional matrices (hypothesis, error, and partial correlations) and an eigen analysis, this information is also displayed. The output is shown only for one model term, Extrusion, and not for the terms Additive or Extrusion*Additive.

Examine the p-values for the Wilks', Lawley-Hotelling, and Pillai's test statistic to judge whether there is significant evidence for model effects. These values are 0.003 for the model term Extrusion, indicating that there is significant evidence for Extrusion main effects at a levels greater than 0.003. The corresponding p-values for Additive and for Additive*Extrusion are 0.025 and 0.302, respectively (not shown), indicating that there is no significant evidence for interaction, but there is significant evidence for Extrusion and Additive main effects at a levels of 0.05 or 0.10.

You can use the SSCP matrices to assess the partitioning of variability in a similar way as you would look at univariate sums of squares. The matrix labeled as SSCP Matrix for Extrusion is the hypothesis sums of squares and cross-products matrix, or H, for the three response with model term Extrusion. The diagonal elements of this matrix, 1.740, 1.301, and 0.4205, are the univariate ANOVA sums of squares for the model term Extrusion when the response variables are Tear, Gloss, and Opacity, respectfully. The off-diagonal elements of this matrix are the cross products.

The matrix labeled as SSCP Matrix for Error is the error sums of squares and cross-products matrix, or E. The diagonal elements of this matrix, 1.764, 2.6280, and 64.924, are the univariate ANOVA error sums of squares when the response variables are Tear, Gloss, and Opacity, respectfully. The off-diagonal elements of this matrix are the cross products. This matrix is displayed once, after the SSCP matrix for the first model term.

You can use the matrix of partial correlations, labeled as Partial Correlations for the Error SSCP Matrix, to assess how related the response variables are. These are the correlations among the residuals or, equivalently, the correlations among the responses conditioned on the model. Examine the off-diagonal elements. The partial correlations between Tear and Gloss of 0.00929 and between Gloss and Opacity of -0.04226 are small. The partial correlation of -0.28687 between Tear and Opacity is not large. Because the correlation structure is weak, you might be satisfied with performing univariate ANOVA for these three responses. This matrix is displayed once, after the SSCP matrix for error.

You can use the eigen analysis to assess how the response means differ among the levels of the different model terms. The eigen analysis is of E-1 H, where E is the error SCCP matrix and H is the response variable SCCP matrix. These are the eigenvalues that are used to calculate the four MANOVA tests.

Place the highest importance on the eigenvectors that correspond to high eigenvalues. In the example, the second and third eigenvalues are zero and therefore the corresponding eigenvectors are meaningless. For both factors, Extrusion and Additive, the first eigenvectors contain similar information The first eigenvector for Extrusion is 0.6541, -0.3385, 0.0359 and for Additive it is -0.6630, -0.3214, -0.0684 (not shown). The highest absolute value within these eigenvectors is for the response Tear, the second highest is for Gloss, and the value for Opacity is small. This implies that the Tear means have the largest differences between the two factor levels of either Extrusion or Additive, the Gloss means have the next largest differences, and the Opacity means have small differences.

General MANOVA

General MANOVA

Stat > ANOVA > General MANOVA

Use general MANOVA to perform multivariate analysis of variance (MANOVA) with balanced and unbalanced designs, or if you have covariates. This procedure takes advantage of the data covariance structure to simultaneously test the equality of means from different responses.

Calculations are done using a regression approach. A "full rank" design matrix is formed from the factors and covariates and each response variable is regressed on the columns of the design matrix.

Factors may be crossed or nested, but they cannot be declared as random; it is possible to work around this restriction by specifying the error term to test model terms (See Specifying terms to test). Covariates may be crossed with each other or with factors, or nested within factors. You can analyze up to 50 response variables with up to 31 factors and 50 covariates at one time.

Dialog box items

Responses: Enter up to 50 numeric columns containing the response variables.

Model: Type the model terms that you want to fit.

<Covariates>

<Options>

<Graphs>

<Results>

<Storage>

Data – General MANOVA

Set up your worksheet in the same manner as with balanced MANOVA: one column for each response variable, one column for each factor, and one column for each covariate, so that there is one row of the worksheet for each observation. The factor columns may be numeric, text, or date/time. If you wish to change the order in which text categories are processed from their default alphabetical order, you can define your own order. (See Ordering Text Categories.) You may include up to 50 response variables and up to 31 factors at one time.

Although models can be unbalanced in general MANOVA, they must be "full rank." That is, there must be enough data to estimate all the terms in your model. For example, suppose you have a two-factor crossed model with one empty cell. Then you can fit the model with terms A B, but not A B A∗B. Minitab will tell you if your model is not full rank. In most cases, eliminating some of the high order interactions in your model (assuming, of course, they are not important) can solve non-full rank problems.

Nesting does not need to be balanced. If factor B is nested within factor A, there can be unequal levels of B within each level of A. In addition, the subscripts used to identify the B levels can differ within each level of A.

If any response, factor, or covariate column contains missing data, that entire observation (row) is excluded from all computations. If an observation is missing for one response variable, that row is eliminated for all responses.

To perform a general MANOVA

1 Choose **Stat > ANOVA > General MANOVA**.

2 In **Responses**, enter up to 50 numeric columns containing the response variables.

3 In **Model**, type the model terms that you want to fit. See Overview of Balanced ANOVA and GLM.

4 If you like, use any dialog box options, then click **OK**.

General MANOVA – Covariates

Stat > ANOVA > General MANOVA > Covariates

Enter covariates into the model.

Dialog box items

Covariates: Enter up to 50 columns containing the covariates.

General MANOVA – Options

Stat > ANOVA > General MANOVA > Options

Allows you to perform a weighted regression.

Dialog box items

Do a weighted fit, using weights in: Enter a column containing weights to perform weighted regression.

General MANOVA – Graphs

Stat > ANOVA > General MANOVA > Graphs

Displays residual plots. You do not have to store the residuals and fits in order to produce these plots.

Dialog box items

Residuals for Plots: You can specify the type of residual to display on the residual plots.

Regular: Choose to plot the regular or raw residuals.

Standardized: Choose to plot the standardized residuals.

Deleted: Choose to plot the Studentized deleted residuals.

Residual Plots

Individual plots: Choose to display one or more plots.

Histogram of residuals: Check to display a histogram of the residuals.

Normal plot of residuals: Check to display a normal probability plot of the residuals.

Residuals versus fits: Check to plot the residuals versus the fitted values.

Residuals versus order: Check to plot the residuals versus the order of the data. The row number for each data point is shown on the x-axis--for example, 1 2 3 4... n.

Four in one: Choose to display a layout of a histogram of residuals, a normal plot of residuals, a plot of residuals versus fits, and a plot of residuals versus order.

Residuals versus the variables: Enter one or more columns containing the variables against which you want to plot the residuals. Minitab displays a separate graph for each column.

General MANOVA – Results

Stat > ANOVA > General MANOVA > Results

Display certain MANOVA and ANOVA output, display means, and customize MANOVA tests.

Dialog box items

Display of Results

Matrices (hypothesis, error, partial correlations): Check to display the hypothesis matrix H, the error matrix E, and a matrix of partial correlations. See MANOVA tests.

Eigen analysis: Check to display the eigenvalues and eigenvalues for the matrix E**-1 H.

Univariate analysis of variance: Check to perform a univariate analysis of variance for each response variable.

Display least squares means corresponding to the terms: Enter terms for which to display a table of means. For example, if you specify A B D A∗B∗D, four table of means will be displayed, one for each main effect, A, B, D, and one for the three-way interaction, A∗B∗D.

Custom multivariate tests for the following terms: Enter terms for which to perform 4 multivariate tests. See Specifying terms to test. By default the tests are performed for all model terms.

Error: Enter an error term for the four multivariate tests. It must be a single term that is in the model. If you do not specify an error term, Minitab uses the error associated with mean squares error, as in the univariate case.

Specifying terms to test

In the Results subdialog box, you can specify model terms in **Custom multivariate test for the following terms** and designate the error term in **Error**. Minitab will perform four multivariate tests (see MANOVA tests) for those terms. This option is most useful when you have factors that you consider as random factors. Model terms that are random or that are interactions with random terms may need a different error term than general MANOVA supplies. You can determine the appropriate error term by entering one response variable with General Linear Model, choose to display the expected mean square, and determine which error term was used for each model terms.

If you specify an error term, it must be a single term that is in the model. This error term is used for all requested tests. If you have different error terms for certain model terms, enter each separately and exercise the general MANOVA dialog for each one. If you do not specify an error term, Minitab uses MSE.

MANOVA tests – General MANOVA

The MANOVA tests with general MANOVA are similar to those performed for balanced MANOVA. See MANOVA tests for Balanced Designs for details.

However, with general MANOVA, there are two SSCP matrices associated with each term in the model, the sequential SSCP matrix and the adjusted SSCP matrix. These matrices are analogous to the sequential SS and adjusted SS in univariate General Linear Model. In fact, the univariate SS's are along the diagonal of the corresponding SSCP matrix. If you do not specify an error term in **Error** when you enter terms in **Custom multivariate tests for the following terms**, then the adjusted SSCP matrix is used for H and the SSCP matrix associated with MSE is used for E. If you do specify an error term, the sequential SSCP matrices associated with H and E are used. Using sequential SSCP matrices guarantees that H and E are statistically independent.

General MANOVA – Storage

Stat > ANOVA > General MANOVA > Storage

Stores the residuals, fitted values, and many other diagnostics for further analysis (see Checking your model).

Dialog box items

Storage

Coefficients: Check to store the coefficients for a model that corresponds to the design matrix. (If M1 contains the design matrix and C1 the coefficients, then M1 times C1 gives the fitted values.)

Fits: Check to store the fitted values.

Residuals: Check to store the residuals.

Standardized residuals: Check to store the standardized residuals.

Deleted t residuals: Check to store Studentized residuals.

Hi [leverage]: Check to store leverages.

Cook's distance: Check to store Cook's distance.

DFITS: Check to store DFITS.

Design matrix: Check to store the design matrix corresponding to your model.

Testing the equality of means from multiple responses

Balanced MANOVA and general MANOVA are procedures for testing the equality of vectors of means from multiple responses. Your choice between these two procedures depends upon the experimental design and the available options. Both procedures can fit MANOVA models to balanced data with up to 31 factors.

- Balanced MANOVA is used to perform multivariate analysis of variance with balanced designs. See Balanced designs. You can also specify factors to be random and obtain expected means squares. Use general MANOVA with unbalanced designs.
- General MANOVA is used to perform multivariate analysis of variance with either balanced or unbalanced designs that can also include covariates. You cannot specify factors to be random as you can for balanced MANOVA, although you can work around this restriction by specifying the error term for testing different model terms.

The table below summarizes the differences between Balanced and General MANOVA:

	Balanced MANOVA	General MANOVA
Can fit unbalanced data	no	yes
Can specify factors as random and obtain expected means squares	yes	no
Can fit covariates	no	yes
Can fit restricted and unrestricted forms of a mixed model	yes	no; unrestricted only

Test for Equal Variances

Test for Equal Variances

Stat > ANOVA > Test for Equal Variances

Use variance test to perform hypothesis tests for equality or homogeneity of variance using Bartlett's and Levene's tests. An F Test replaces Bartlett's test when you have just two levels.

Many statistical procedures, including analysis of variance, assume that although different samples may come from populations with different means, they have the same variance. The effect of unequal variances upon inferences depends in part upon whether your model includes fixed or random effects, disparities in sample sizes, and the choice of multiple comparison procedure. The ANOVA F-test is only slightly affected by inequality of variance if the model contains fixed factors only and has equal or nearly equal sample sizes. F-tests involving random effects may be substantially affected, however [19]. Use the variance test procedure to test the validity of the equal variance assumption.

Dialog box items

Response: Enter the column containing the response variable.

Factors: Enter the columns containing the factors in the model.

Confidence level: Enter a value from 0 to 100 for the level of confidence desired for the confidence intervals displayed on the graph. The default level is 95. Minitab uses the Bonferroni method to calculate the simultaneous confidence intervals.

Title: Type the desired text in this box to replace the default title with your own custom title.

<Storage>

Data – Test for Equal Variances

Set up your worksheet with one column for the response variable and one column for each factor, so that there is one row for each observation. Your response data must be in one column. You may have up to 9 factors. Factor columns may be numeric, text, or date/time, and may contain any value. If there are many cells (factors and levels), the print in the output chart can get very small.

Rows where the response column contains missing data (∗) are automatically omitted from the calculations. When one or more factor columns contain missing data, Minitab displays the chart and Bartlett's test results. When you have missing data in a factor column, Minitab displays the Levene's test results only when two or more cells have multiple observations and one of those cells has three or more observations.

Data limitations include the following:

1 If none of the cells have multiple observations, nothing is calculated. In addition, there must be at least one nonzero standard deviation

2 The F-test for 2 levels requires both cells to have multiple observations.

3 Bartlett's test requires two or more cells to have multiple observations.

4 Levene's test requires two or more cells to have multiple observations, but one cell must have three or more.

Bartlett's versus Levene's tests

Minitab calculates and displays a test statistic and p-value for both Bartlett's test and Levene's test where the null hypothesis is of equal variances versus the alternative of not all variances being equal. If there are only two levels, an F-test is performed in place of Bartlett's test.

- Use Bartlett's test when the data come from normal distributions; Bartlett's test is not robust to departures from normality.
- Use Levene's test when the data come from continuous, but not necessarily normal, distributions. This method considers the distances of the observations from their sample median rather than their sample mean, makes the test more robust for smaller samples.

To perform a test for equal variances

1 Choose **Stat > ANOVA > Test for Equal Variances**.

2 In **Response**, enter the column containing the response.

3 In **Factors**, enter up to nine columns containing the factor levels.

4 If you like, use any dialog box options, then click **OK**.

Test for Equal Variances – Storage

Stat > ANOVA > Test for Equal Variances > Storage

Allows for the storage of the cell standard deviations, cell variances, and confidence limits for the standard deviations.

Dialog box items

Storage

Standard deviations: Check to store the standard deviation of each cell (each level of each factor).

Variances: Check to store the variance of each cell.

Upper confidence limits for sigmas: Check to store the upper confidence limits for the standard deviations.

Lower confidence limits for sigmas: Check to store the lower confidence limits for the standard deviations.

Example of Performing a Test for Equal Variance

You study conditions conducive to potato rot by injecting potatoes with bacteria that cause rotting and subjecting them to different temperature and oxygen regimes. Before performing analysis of variance, you check the equal variance assumption using the test for equal variances.

1 Open the worksheet EXH_AOV.MTW.

2 Choose **Stat > ANOVA > Test for Equal Variances**.

3 In **Response**, enter *Rot*.

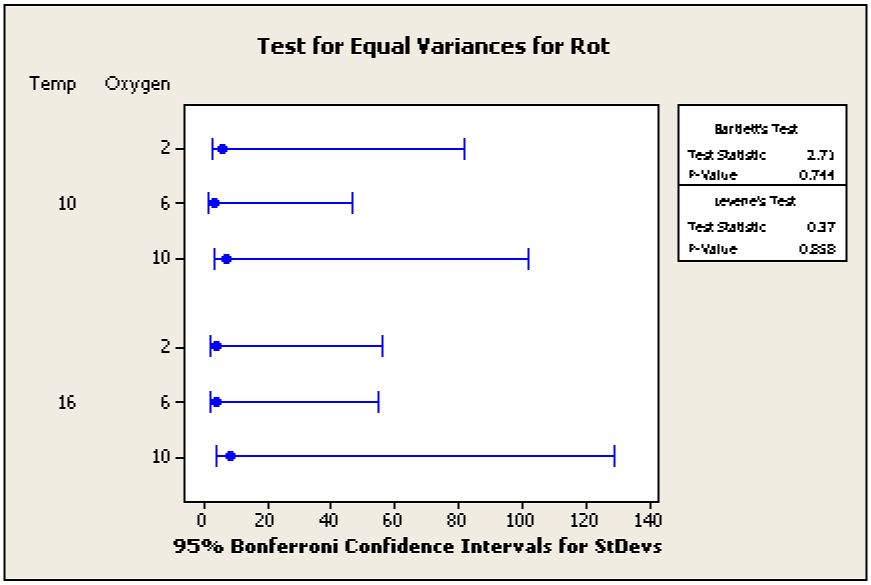
4 In **Factors**, enter *Temp Oxygen*. Click **OK**.

Session window output

Test for Equal Variances: Rot versus Temp, Oxygen

```
95% Bonferroni confidence intervals for standard deviations

Temp  Oxygen  N    Lower    StDev    Upper
  10       2  3  2.26029  5.29150   81.890
  10       6  3  1.28146  3.00000   46.427
  10      10  3  2.80104  6.55744  101.481
  16       2  3  1.54013  3.60555   55.799
  16       6  3  1.50012  3.51188   54.349
  16      10  3  3.55677  8.32666  128.862


Bartlett's Test (normal distribution)
Test statistic = 2.71, p-value = 0.744


Levene's Test (any continuous distribution)
Test statistic = 0.37, p-value = 0.858


Test for Equal Variances: Rot versus Temp, Oxygen
```

Graph window output



Interpreting the results

The test for equal variances generates a plot that displays Bonferroni 95% confidence intervals for the response standard deviation at each level. Bartlett's and Levene's test results are displayed in both the Session window and in the graph. Note that the 95% confidence level applies to the family of intervals and the asymmetry of the intervals is due to the skewness of the chi-square distribution.

For the potato rot example, the p-values of 0.744 and 0.858 are greater than reasonable choices of α, so you fail to reject the null hypothesis of the variances being equal. That is, these data do not provide enough evidence to claim that the populations have unequal variances.

Interval Plot

Interval Plot

Gallery

Data

One Y, Simple

To display a simple interval plot

Example, one y - simple

One Y, With Groups

To display an interval plot with groups

Example, one y - with groups

Multiple Y's, Simple

To display a simple interval plot with multiple y's

Example multiple y's - simple

Multiple Y's, With Groups

To display an interval plot with multiple y's and groups

Example, multiple y's - with groups

Main Effects Plot

Main Effects Plot

Stat > ANOVA > Main Effects Plot

Use Main Effects Plot to plot data means when you have multiple factors. The points in the plot are the means of the response variable at the various levels of each factor, with a reference line drawn at the grand mean of the response data. Use the main effects plot for comparing magnitudes of main effects.

Use Factorial Plots to generate main effects plots specifically for two-level factorial designs.

Dialog box items

Responses: Enter the columns containing the response data. You can include up to 50 responses.

Factors: Enter the columns containing the factor levels. You can include up to 9 factors.

<Options>

Data – Main Effects Plot

Set up your worksheet with one column for the response variable and one column for each factor, so that each row in the response and factor columns represents one observation. It is not required that your data be balanced.

The factor columns may be numeric, text, or date/time and may contain any values. If you wish to change the order in which text levels are processed, you can define your own order. See Ordering Text Categories. You may have up to 9 factors.

Missing values are automatically omitted from calculations.

To perform a main effects plot

1 Choose **Stat > ANOVA > Main Effects Plot**.

2 In **Responses**, enter the columns containing the response data.

3 In **Factors**, enter the columns containing the factor levels. You can enter up to 9 factors.

4 If you like, use any dialog box options, then click **OK**.

Main Effects Plot – Options

Stat > ANOVA > Main Effects Plot > Options

Allows you control the y scale minima and maxima and to add a title to the main effects plot.

Minimum for Y (response) scale: Enter either a single scale minimum for all responses or one scale minimum for each response.

Maximum for Y (response) scale: Enter either a single scale minimum for all responses or one scale minimum for each response.

Title: To replace the default title with your own custom title, type the desired text in this box.

Example of Main Effects Plot

You grow six varieties of alfalfa on plots within four different fields and you weigh the yield of the cuttings. You are interested in comparing yields from the different varieties and consider the fields to be blocks. You want to preview the data and examine yield by variety and field using the main effects plot.

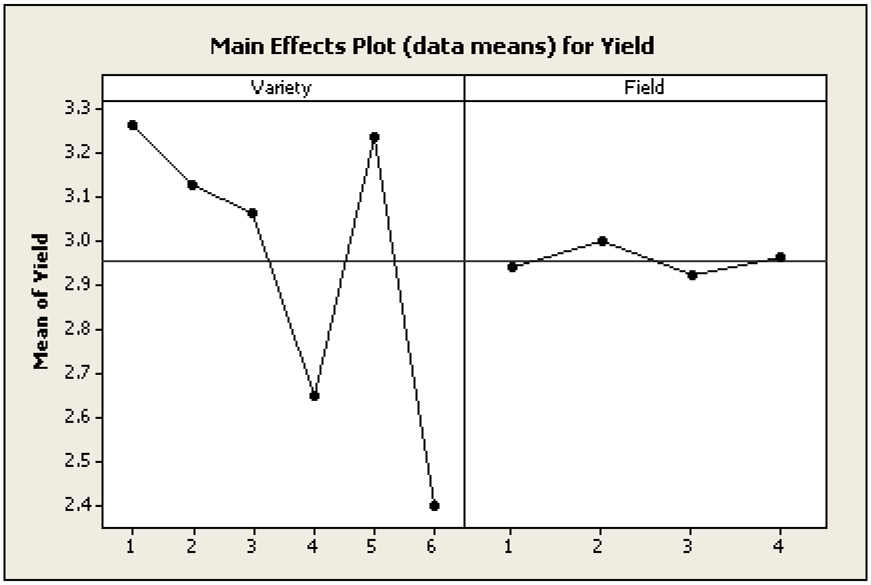
1 Open the worksheet ALFALFA.MTW.

2 Choose **Stat > ANOVA > Main Effects Plot**.

3 In **Responses**, enter *Yield*.

4 In **Factors**, enter *Variety Field*. Click **OK**.

Graph window output



Interpreting the results

The main effects plot displays the response means for each factor level in sorted order if the factors are numeric or date/time or in alphabetical order if text, unless value ordering has been assigned (see Ordering Text Categories). A horizontal line is drawn at the grand mean. The effects are the differences between the means and the reference line. In the example, the variety effects upon yield are large compared to the effects of field (the blocking variable).

Interactions Plot

Interactions Plot

Stat > ANOVA > Interactions Plot

Interactions Plot creates a single interaction plot for two factors, or a matrix of interaction plots for three to nine factors. An interactions plot is a plot of means for each level of a factor with the level of a second factor held constant. Interactions plots are useful for judging the presence of interaction.

Interaction is present when the response at a factor level depends upon the level(s) of other factors. Parallel lines in an interactions plot indicate no interaction. The greater the departure of the lines from the parallel state, the higher the degree of interaction. To use interactions plot, data must be available from all combinations of levels.

Use Interactions plots for factorial designs to generate interaction plots specifically for 2-level factorial designs, such as those generated by Fractional Factorial Design, Central Composite Design, and Box-Behnken Design.

Dialog box items

Responses: Enter the columns containing the response data. You can include up to 50 responses.

Factors: Enter the columns containing the factor levels. You can include up to 9 factors.

Display full interaction plot matrix: Check to display the full interaction matrix when more than two factors are specified instead of displaying only the upper right portion of the matrix. In the full matrix, the transpose of each plot in the upper right displays in the lower left portion of the matrix. The full matrix takes longer to display than the half matrix.

<Options>

Data – Interactions Plot

Set up your worksheet with one column for the response variable and one column for each factor, so that each row in the response and factor columns represents one observation. Your data is not required to be balanced.

The factor columns may be numeric, text, or date/time and may contain any values. If you wish to change the order in which text levels are processed, you can define your own order. See Ordering Text Categories. You may have from 2 through 9 factors.

Missing data are automatically omitted from calculations.

To display an interactions plot

1 Choose **Stat > ANOVA > Interactions Plot**.

2 In **Responses**, enter the columns containing the response data.

3 In **Factors**, enter from 2 to 9 columns containing the factor levels. If you have two factors, the x-variable will be the second factor that you enter.

4 If you like, use any of the dialog box options, then click **OK**.

Interactions Plot – Options

Stat > ANOVA > Interactions Plot > Options

Allows you control the y scale minima and maxima and to add a title to the interaction plot.

Minimum for Y (response) scale: Enter either a single scale minimum for all responses or one scale minimum for each response.

Maximum for Y (response) scale: Enter either a single scale minimum for all responses or one scale minimum for each response.

Title: To replace the default title with your own custom title, type the desired text in this box.

Example of an Interactions Plot with Two Factors

You conduct an experiment to test the effect of temperature and glass type upon the light output of an oscilloscope (example and data from [13], page 252). There are three glass types and three temperatures, 100, 125, and 150 degrees Fahrenheit. You choose interactions plot to visually assess interaction in the data. You enter the quantitative variable second because you want this variable as the x variable in the plot.

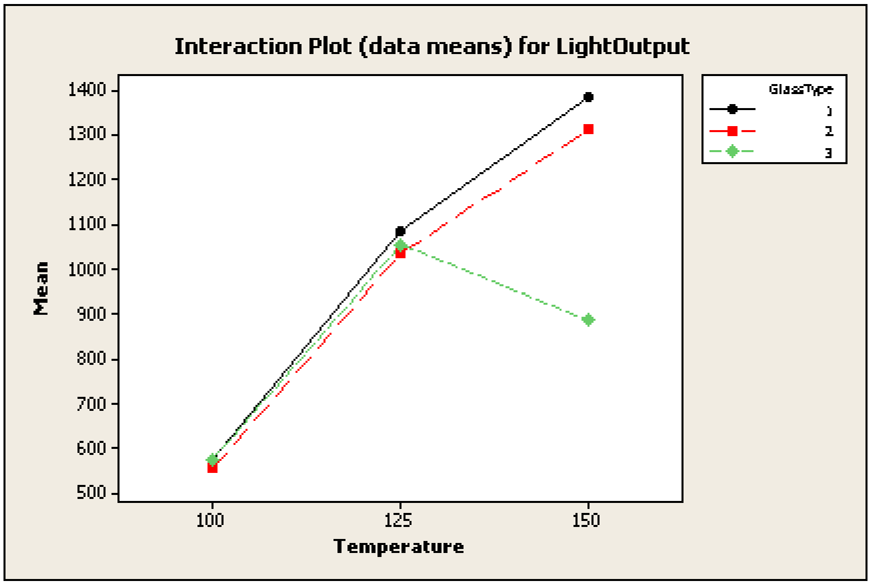
1 Open the worksheet EXH_AOV.

2 Choose **Stat > ANOVA > Interactions Plot**.

3 In **Responses**, enter *LightOutput*.

4 In **Factors**, enter *GlassType Temperature*. Click **OK**.

Graph window output



Interpreting the results

This interaction plot shows the mean light output versus the temperature for each of the three glass types. The legend shows which symbols and lines are assigned to the glass types. The means of the factor levels are plotted in sorted order if numeric or date/time or in alphabetical order if text, unless value ordering has been assigned (see Ordering Text Categories).

This plot shows apparent interaction because the lines are not parallel, implying that the effect of temperature upon light output depends upon the glass type. We test this using the General Linear Model.

Example of an Interactions Plot with more than Two Factors

Plywood is made by cutting thin layers of wood from logs as they are spun on their axis. Considerable force is required to turn a log hard enough so that a sharp blade can cut off a layer. Chucks are inserted into the ends of the log to apply the torque necessary to turn the log. You conduct an experiment to study factors that affect torque. These factors are diameter of the logs, penetration distance of the chuck into the log, and the temperature of the log. You wish to preview the data to check for the presence of interaction.

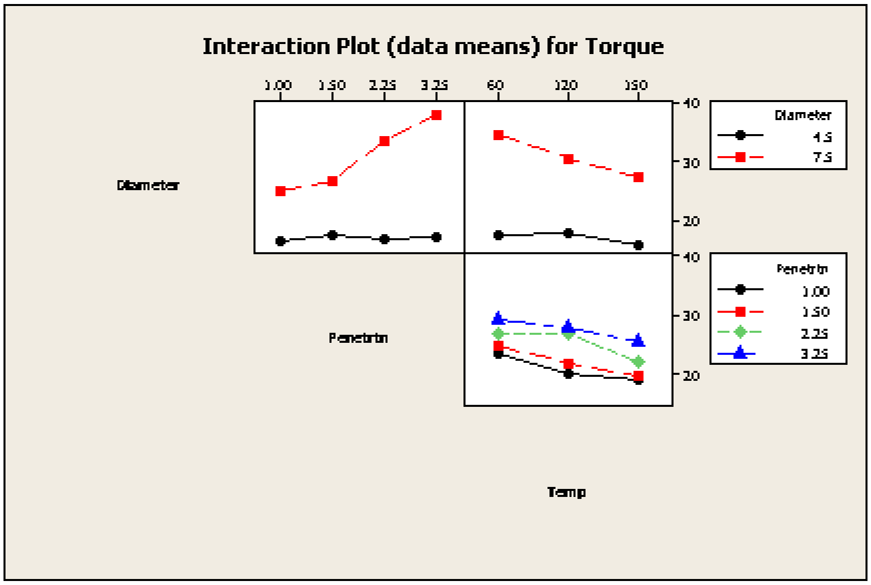
1 Open the worksheet PLYWOOD.MTW.

2 Choose **Stat > ANOVA > Interactions Plot**.

3 In **Responses**, enter *Torque*.

4 In **Factors**, enter *Diameter-Temp*. Click **OK**.

Graph window output



Interpreting the results

An interaction plot with three or more factors show separate two-way interaction plots for all two-factor combinations. In this example, the plot in the middle of the top row shows the mean torque versus the penetration levels for both levels of diameter, 4.5 and 7.5, averaged over all levels of temperature. There are analogous interactions plots for diameter by temperature (upper right) and penetration by temperature (second row).

For this example, the diameter by penetration and the diameter by temperature plots show nonparallel lines, indicating interaction. The presence of penetration by temperature interaction is not so easy to judge. This interaction might best be judged in conjunction with a model-fitting procedure, such as GLM.

Multivariate Analysis

Overview

Multivariate Analysis Overview

Use Minitab's multivariate analysis procedures to analyze your data when you have made multiple measurements on items or subjects. You can choose to:

- Analyze the data covariance structure to understand it or to reduce the data dimension
- Assign observations to groups
- Explore relationships among categorical variables

Because Minitab does not compare tests of significance for multivariate procedures, interpreting the results is somewhat subjective. However, you can make informed conclusions if you are familiar with your data.

Analysis of the data structure

Minitab offers two procedures for analyzing the data covariance structure:

- Principal Components helps you to understand the covariance structure in the original variables and/or to create a smaller number of variables using this structure.
- Factor Analysis, like principal components, summarizes the data covariance structure in a smaller number of dimensions. The emphasis in factor analysis is the identification of underlying "factors" that might explain the dimensions associated with large data variability.

Grouping observations

Minitab offers three cluster analysis methods and discriminant analysis for grouping observations:

- Cluster Observations groups or clusters observations that are "close" to each other when the groups are initially unknown. This method is a good choice when no outside information about grouping exists. The choice of final grouping is usually made according to what makes sense for your data after viewing clustering statistics.
- Cluster Variables groups or clusters variables that are "close" to each other when the groups are initially unknown. The procedure is similar to clustering of observations. You may want to cluster variables to reduce their number.
- Cluster K-Means, like clustering of observations, groups observations that are "close" to each other. K-means clustering works best when sufficient information is available to make good starting cluster designations.
- Discriminant Analysis classifies observations into two or more groups if you have a sample with known groups. You can use discriminant analysis to investigate how the predictors contribute to the groupings.

Correspondence Analysis

Minitab offers two methods of correspondence analysis to explore the relationships among categorical variables:

- Simple Correspondence Analysis explores relationships in a 2-way classification. You can use this procedure with 3-way and 4-way tables because Minitab can collapse them into 2-way tables. Simple correspondence analysis decomposes a contingency table similar to how principal components analysis decomposes multivariate continuous data. Simple correspondence analysis performs an eigen analysis of data, breaks down variability into underlying dimensions, and associates variability with rows and/or columns.
- Multiple Correspondence Analysis extends simple correspondence analysis to the case of 3 or more categorical variables. Multiple correspondence analysis performs a simple correspondence analysis on an indicator variables matrix in which each column corresponds to a level of a categorical variable. Rather than a 2-way table, the multi-way table is collapsed into 1 dimension.

Multivariate

Stat > Multivariate

Allows you to perform a principal components analysis, factor analysis, cluster analysis, discriminant analysis, and correspondence analysis.

Select one of the following options:

Principal Components – performs principal components analysis

Factor Analysis – performs factor analysis

Cluster Observations – performs agglomerative hierarchical clustering of observations

Cluster Variables – performs agglomerative hierarchical clustering of variables

Cluster K-Means – performs K-means non-hierarchical clustering of observations

Discriminant Analysis – performs linear and quadratic discriminant analysis

Simple Correspondence Analysis – performs simple correspondence analysis on a two-way contingency table

Multiple Correspondence Analysis – performs multiple correspondence analysis on three or more categorical variables

Minitab offers the following additional multivariate analysis options:

Balanced MANOVA

General MANOVA

Multivariate control charts

Examples of Multivariate Analysis

The following examples illustrate how to use the various multivariate analysis techniques available. Choose an example below:

Principal Components Analysis

Factor Analysis

Cluster Observations

Cluster Variables

Cluster K-Means

Discriminant Analysis

Simple Correspondence Analysis

Multiple Correspondence Analysis

References – Multivariate Analysis

[1] T.W. Anderson (1984). *An Introduction to Multivariate Statistical Analysis*, Second Edition. John Wiley & Sons.

[2] W. Dillon and M. Goldstein (1984). *Multivariate Analysis: Methods and Applications*. John Wiley & Sons.

[3] S.E. Fienberg (1987). *The Analysis of Cross-Classified Categorical Data*. The MIT Press.

[4] M. J. Greenacre (1993). *Correspondence Analysis in Practice.* Academic Press, Harcourt, Brace & Company.

[5] H. Harmon (1976). *Modern Factor Analysis*, Third Edition. University of Chicago Press.

[6] R. Johnson and D. Wichern (1992). *Applied Multivariate Statistical Methods*, Third Edition. Prentice Hall.

[7] K. Joreskog (1977). "Factor Analysis by Least Squares and Maximum Likelihood Methods," *Statistical Methods for Digital Computers,* ed. K. Enslein, A. Ralston and H. Wilf, John Wiley & Sons.

[8] J. K. Kihlberg, E. A. Narragon, and B. J. Campbell. (1964). *Automobile crash injury in relation to car size*. Cornell Aero. Lab. Report No. VJ-1823-R11.

[9] G.N. Lance and W.T. Williams (1967). "A General Theory of Classificatory Sorting Strategies, I. Hierarchical systems," *Computer Journal*, 9, 373–380

[10] G. W. Milligan (1980). "An Examination of the Effect of Six Types of Error Pertubation on Fifteen Clustering Algorithms," *Psychometrika*, 45, 325-342.

[11] S.J. Press and S. Wilson (1978). "Choosing Between Logistic Regression and Discriminant Analysis," *Journal of the American Statistical Association*, 73, 699-705.

[12] A. C. Rencher (1995). *Methods of Multivariate Analysis*, John Wiley & Sons.

Principal Components

Principal Components Analysis

Stat > Multivariate > Principal Components

Use principal component analysis to help you to understand the underlying data structure and/or form a smaller number of uncorrelated variables (for example, to avoid multicollinearity in regression).

An overview of principal component analysis can be found in most books on multivariate analysis, such as [5].

Dialog box items

Variables: Choose the columns containing the variables to be included in the analysis.

Number of components to compute: Enter the number of principal components to be extracted. If you do not specify the number of components and there are p variables selected, then p principal components will be extracted. If p is large, you may want just the first few.

Type of Matrix

Correlation: Choose to calculate the principal components using the correlation matrix. Use the correlation matrix if it makes sense to standardize variables (the usual choice when variables are measured by different scales).

Covariance: Choose to calculate the principal components using the covariance matrix. Use the covariance matrix if you do not wish to standardize variables.

<Graphs>

<Storage>

Data – principal components analysis

Set up your worksheet so that each row contains measurements on a single item or subject. You must have two or more numeric columns, with each column representing a different measurement (response). If a missing value exists in any column, Minitab ignores the whole row. Missing values are excluded from the calculation of the correlation or covariance matrix.

To perform Principal Component Analysis

1 Choose **Stat > Multivariate > Principal Components**.

2 In **Variables**, enter the columns containing the measurement data.

3 If you like, use any dialog box options, then click **OK**.

Nonuniqueness of Coefficients

The coefficients are unique (except for a change in sign) if the eigenvalues are distinct and not zero. If an eigenvalue is repeated, then the "space spanned" by all the principal component vectors corresponding to the same eigenvalue is unique, but the individual vectors are not. Therefore, the coefficients that Minitab prints and those in a book or another program may not agree, though the eigenvalues (variances) will always be the same.

If the covariance matrix has rank $r < p$, where p is the number of variables, then there will be $p - r$ eigenvalues equal to zero. Eigenvectors corresponding to these eigenvalues may not be unique. This can happen if the number of observations is less than p or if there is multicollinearity.

Principal Components Analysis – Graphs

Stat > Multivariate > Principal Components> Graphs

Displays plots for judging the importance of the different principal components and for examining the scores of the first two principal components.

Dialog box items

Scree plot: Check to display a Scree plot (eigenvalue profile plot). Minitab plots the eigenvalue associated with a principal component versus the number of the component. Use this plot to judge the relative magnitude of eigenvalues.

Score plot for first 2 components: Check to plot the scores for the second principal component (y-axis) versus the scores for the first principal component (x-axis). To create plots for other components, store the scores and use Graph > Scatterplot.

Loading plot for first 2 components: Check to plot the loadings for the second component (y-axis) versus the loadings for the first component (x-axis). A line is drawn from each loading to the (0, 0) point.

Principal Components Analysis – Storage

Stat > Multivariate > Principal Components > Storage

Stores the coefficients and scores.

Dialog box items

Coefficients: Enter the storage columns for the coefficients of the principal components. The number of columns specified must be less than or equal to the number of principal components calculated.

Scores: Enter the storage columns for the principal components scores. Scores are linear combinations of your data using the coefficients. The number of columns specified must be less than or equal to the number of principal components calculated.

Example of Principal Components Analysis

You record the following characteristics for 14 census tracts: total population (Pop), median years of schooling (School), total employment (Employ), employment in health services (Health), and median home value (Home). The data were obtained from [6], Table 8.2.

You perform principal components analysis to understand the underlying data structure. You use the correlation matrix to standardize the measurements because they are not measured with the same scale.

1 Open the worksheet EXH_MVAR.MTW.

2 Choose **Stat > Multivariate > Principal Components**.

3 In **Variables**, enter *Pop-Home*.

4 Under **Type of Matrix**, choose **Correlation**.

5 Click **Graphs** and check **Scree plot**.

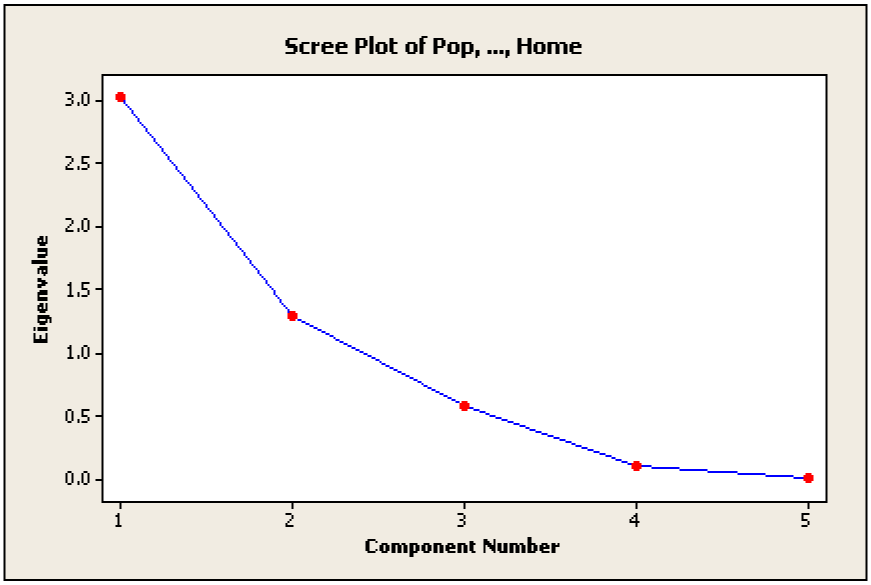
6 Click **OK** in each dialog box.

Session window output

Principal Component Analysis: Pop, School, Employ, Health, Home

```
Eigenanalysis of the Correlation Matrix

Eigenvalue  3.0289  1.2911  0.5725  0.0954  0.0121
Proportion   0.606   0.258   0.114   0.019   0.002
Cumulative   0.606   0.864   0.978   0.998   1.000


Variable      PC1     PC2     PC3     PC4     PC5
Pop        -0.558  -0.131   0.008   0.551  -0.606
School     -0.313  -0.629  -0.549  -0.453   0.007
Employ     -0.568  -0.004   0.117   0.268   0.769
Health     -0.487   0.310   0.455  -0.648  -0.201
Home        0.174  -0.701   0.691   0.015   0.014
```

Graph window output



Interpreting the results

The first principal component has variance (eigenvalue) 3.0289 and accounts for 60.6% of the total variance. The coefficients listed under PC1 show how to calculate the principal component scores:

PC1 = –.558 Pop – .313 School – .568 Employ – .487 Health + .174 Home

It should be noted that the interpretation of the principal components is subjective, however, obvious patterns emerge quite often. For instance, one could think of the first principal component as representing an overall population size, level of schooling, employment level, and employment in health services effect, because the coefficients of these terms have the same sign and are not close to zero.

The second principal component has variance 1.2911 and accounts for 25.8% of the data variability. It is calculated from the original data using the coefficients listed under PC2. This component could be thought of as contrasting level of schooling and home value with health employment to some extent.

Together, the first two and the first three principal components represent 86.4% and 97.8%, respectfully, of the total variability. Thus, most of the data structure can be captured in two or three underlying dimensions. The remaining principal components account for a very small proportion of the variability and are probably unimportant. The Scree plot provides this information visually.

Factor Analysis

Factor Analysis

Stat > Multivariate > Factor Analysis

Use factor analysis, like principal components analysis, to summarize the data covariance structure in a few dimensions of the data. However, the emphasis in factor analysis is the identification of underlying "factors" that might explain the dimensions associated with large data variability.

Dialog box items

Variables: Choose the columns containing the variables you want to use in the analysis. If you want to use a stored correlation or covariance matrix, or the loadings from a previous analysis instead of the raw data, click <Options>.

Number of factors to extract: Enter number of factors to extract (required if you use maximum likelihood as your method of extraction). If you don't specify a number with a principal components extraction, Minitab sets it equal to the number of variables in the data set. If you choose too many factors, Minitab will issue a warning in the Session window.

Method of Extraction:

Principal components: Choose to use the principal components method of factor extraction.

Maximum likelihood: Choose to use maximum likelihood for the initial solution.

Type of Rotation: Controls orthogonal rotations.

None: Choose not to rotate the initial solution.

Equimax: Choose to perform an equimax rotation of the initial solution (gamma = number of factors / 2).

Varimax: Choose to perform a varimax rotation of the initial solution (gamma = 1).

Quartimax: Choose to perform a quartimax rotation of the initial solution (gamma = 0).

Orthomax with gamma: Choose to perform an orthomax rotation of the initial solution, then enter value for gamma between 0 and 1.

<Options>

<Graphs>

<Storage>

<Results>

Data – Factor Analysis

You can have three types of input data:

- Columns of raw data
- A matrix of correlations or covariances
- Columns containing factor loadings

The typical case is to use raw data. Set up your worksheet so that a row contains measurements on a single item or subject. You must have two or more numeric columns, with each column representing a different measurement (response). Minitab automatically omits rows with missing data from the analysis.

Usually the factor analysis procedure calculates the correlation or covariance matrix from which the loadings are calculated. However, you can enter a matrix as input data. You can also enter both raw data and a matrix of correlations or covariances. If you do, Minitab uses the matrix to calculate the loadings. Minitab then uses these loadings and the raw data to calculate storage values and generate graphs. See To perform factor analysis with a correlation or covariance matrix.

If you store initial factor loadings, you can later input these initial loadings to examine the effect of different rotations. You can also use stored loadings to predict factor scores of new data. See To perform factor analysis with stored loadings.

To perform factor analysis with a correlation or covariance matrix

You can choose to calculate the factor loadings and coefficients from a stored correlation or covariance matrix rather than the raw data. In this case, the raw data will be ignored. (Please note that this means scores can not be calculated.)

If it makes sense to standardize variables (usual choice when variables are measured by different scales), enter a correlation matrix; if you do not wish to standardize, enter a covariance matrix.

1 Choose **Stat > Multivariate > Factor Analysis**.

2 Click **Options**.

3 Under **Matrix to Factor**, choose **Correlation** or **Covariance**.

4 Under **Source of Matrix**, choose **Use matrix** and enter the matrix. Click **OK**.

To perform factor analysis with raw data

There are three ways that you might carry out a factor analysis in Minitab. The usual way, described below, is to enter columns containing your measurement variables, but you can also use a matrix as input (See To perform factor analysis with a correlation or covariance matrix) or use stored loadings as input (See To perform factor analysis with stored loadings).

1 Choose **Stat > Multivariate > Factor Analysis**.

2 In **Variables**, enter the columns containing the measurement data.

3 If you like, use any dialog box options, then click **OK**.

To perform factor analysis with stored loadings

If you store initial factor loadings from an earlier analysis, you can input these initial loadings to examine the effect of different rotations. You can also use stored loadings to predict factor scores of new data.

1 Click **Options** in the Factor Analysis dialog box.

2 Under **Loadings for Initial Solution**, choose **Use loadings**. Enter the columns containing the loadings. Click **OK**.

3 Do one of the following, and then click **OK**:

- To examine the effect of a different rotation method, choose an option under **Type of Rotation**. See Rotating the factor loadings for a discussion of the various rotations>Main.
- To predict factor scores with new data, in **Variables**, enter the columns containing the new data.

Factor analysis in practice

The goal of factor analysis is to find a small number of factors, or unobservable variables, that explains most of the data variability and yet makes contextual sense. You need to decide how many factors to use, and find loadings that make the most sense for your data.

Number of factors

The choice of the number of factors is often based upon the proportion of variance explained by the factors, subject matter knowledge, and reasonableness of the solution [6]. Initially, try using the principal components extraction method without specifying the number of components. Examine the proportion of variability explained by different factors and narrow down your choice of how many factors to use. A Scree plot may be useful here in visually assessing the importance of factors. Once you have narrowed this choice, examine the fits of the different factor analyses. Communality values, the proportion of variability of each variable explained by the factors, may be especially useful in comparing fits. You may decide to add a factor if it contributes to the fit of certain variables. Try the maximum likelihood method of extraction as well.

Rotation

Once you have selected the number of factors, you will probably want to try different rotations. Johnson and Wichern [6] suggest the varimax rotation. A similar result from different methods can lend credence to the solution you have selected. At this point you may wish to interpret the factors using your knowledge of the data. For more information see Rotating the factor loadings.

Rotating the factor loadings

There are four methods to orthogonally rotate the initial factor loadings found by either principal components or maximum likelihood extraction. An orthogonal rotation simply rotates the axes to give you a different perspective. The methods are equimax, varimax, quartimax, and orthomax. Minitab rotates the loadings in order to minimize a simplicity criterion [5]. A parameter, gamma, within this criterion is determined by the rotation method. If you use a method with a low value of gamma, the rotation will tend to simplify the rows of the loadings; if you use a method with a high value of gamma, the rotation will tend to simplify the columns of the loadings. The table below summarizes the rotation methods.

Rotation method	**Goal is ...**	**Gamma**
equimax	to rotate the loadings so that a variable loads high on one factor but low on others	number of factors / 2
varimax	to maximize the variance of the squared loadings	1
quartimax	simple loadings	0
orthomax	user determined, based on the given value of gamma	0-1

Factor Analysis – Options

Stat > Multivariate > Factor Analysis > Options

Allows you to specify the matrix type and source, and the loadings to use for the initial extraction.

Dialog box items

Matrix to Factor

Correlation: Choose to calculate the factors using the correlation matrix. Use the correlation matrix if it makes sense to standardize variables (the usual choice when variables are measured by different scales).

Covariance: Choose to calculate the factors using the covariance matrix. Use the covariance matrix if you do not wish to standardize variables. The covariance matrix cannot be used with a maximum likelihood estimation.

Source of Matrix:

Compute from variables: Choose to use the correlation or covariance matrix of the measurement data.

Use matrix: Choose to use a stored matrix for calculating the loadings and coefficients. (Note: Scores can not be calculated if this option is chosen.) See To perform factor analysis with a correlation or covariance matrix.

Loadings for Initial Solution

Compute from variables: Choose to compute loadings from the raw data.

Use loadings: Choose to use loadings which were previously calculated, then specify the columns containing the loadings. You must specify one column for each factor calculated. See To perform factor analysis with stored loadings.

Maximum Likelihood Extraction

Use initial communality estimates in: Choose the column containing data to be used as the initial values for the communalities. The column should contain one value for each variable.

Max iterations: Enter the maximum number of iterations allowed for a solution (default is 25).

Convergence: Enter the criterion for convergence (occurs when the uniqueness values do not change very much). This number is the size of the smallest change (default is 0.005).

Factor Analysis – Graphs

Stat > Multivariate > Factor Analysis > Graphs

Displays a Scree plot, and score and loading plots for the first two factors.

To create simple loading plots for other factors, store the loadings and use Graph > Scatterplot. If you want to connect the loading point to the zero point, add a zero to the bottom of each column of loadings in the Data window, then add lines connecting the loading points to the zero point with the graph editor. See graph editing overview.

Dialog box items

Scree plot: Check to display a Scree plot (eigenvalue profile plot). Minitab plots the eigenvalue associated with a factor versus the number of the factor.

Score plot for first 2 factors: Check to plot the scores for the second factor (y-axis) versus the scores for the first factor (x-axis). Scores are linear combinations of your data using the coefficients. To create plots for other factors, store the scores and use Graph > Scatterplot. (Note: Scores must be calculated from raw data, therefore this graph can not be generated if the **Use matrix** option is selected. See <Options>.)

Loading plot for first 2 factors: Check to plot the loadings for the second factor (y-axis) versus the loadings for the first factor (x-axis). A line is drawn from each loading to the (0, 0) point.

Factor Analysis – Storage

Stat > Multivariate > Factor Analysis > Storage

Allows you to store factor loadings, factor score coefficients, factor or standard scores, rotation matrix, residual matrix, eigenvalues, and eigenvectors. You can then use this information for further analysis.

Dialog box items

Storage

Loadings: Enter storage columns for the factor loadings. You must enter one column for each factor. If a rotation was specified, Minitab stores the values for the rotated factor loadings These can be input using <Options> and specifying the columns under Loadings for initial solutions.

Coefficients: Enter storage columns for the factor score coefficients. You must enter one column for each factor.

Scores: Enter storage columns for the scores. You must enter one column for each factor. Minitab calculates factor scores by multiplying factor score coefficients and your data after they have been centered by subtracting means. (Note: Scores must be calculated from raw data, therefore the **Use matrix** option must not be selected. See <Options>.)

Rotation matrix: Enter a location to store the matrix used to rotate the initial loadings. You may enter a matrix name or number (for example, M3). The rotation matrix is the matrix used to rotate the initial loadings. If L is the matrix of initial loadings and M is the rotation matrix, LM is the matrix of rotated loadings.

Residual matrix: Enter a location to store the residual matrix. The residual matrix for the initial and rotated solutions are the same. You may enter a matrix name or number (for example, M3). The residual matrix is (A-LL'), where A is the correlation or covariance matrix and L is a matrix of loadings. The residual matrix is the same for initial or rotated solutions.

Eigenvalues: Enter a column to store the eigenvalues of the matrix that was factored. The eigenvalues are stored in numerical order from largest to smallest. To store eigenvalues, you must do the initial extraction using principal components. You can plot the eigenvalues to obtain a Scree plot.

Eigenvector matrix: Enter a matrix to store the eigenvectors of the matrix that was factored. Each vector is stored as a column of the matrix, in the same order as the eigenvalues.

Factor analysis storage

To store loadings, factor score coefficients, or factor scores, enter a column name or column number for each factor that has been extracted. The number of storage columns specified must be equal in number to the number of factors calculated. If a rotation was specified, Minitab stores the values for the rotated solution. Minitab calculates factor scores by multiplying factor score coefficients and your data after they have been centered by subtracting means.

You can also store the rotation matrix and residual matrix. Enter a matrix name or matrix number. The rotation matrix is the matrix used to rotate the initial loadings. If L is the matrix of initial loadings and M is the rotation matrix that you store, LM is the matrix of rotated loadings. The residual matrix is (A-LL), where A is the correlation or covariance matrix and L is a matrix of loadings. The residual matrix is the same for initial and rotated solutions.

You can also store the eigenvalues and eigenvectors of the correlation or covariance matrix (depending on which is factored) if you chose the initial factor extraction via principal components. Enter a single column name or number for storing eigenvalues, which are stored from largest to smallest. Enter a matrix name or number to store the eigenvectors in an order corresponding to the sorted eigenvalues.

Factor Analysis – Results

Stat > Multivariate > Factor Analysis > Results

Controls the display of Session window results.

Dialog box items

Display of Results:

Do not display: Choose to suppress the display of results. All requested storage is done.

Loadings only: Choose to display loadings (and sorted loadings if requested) for the final solution.

Loadings and factor score coefficients: Choose to display factor loadings and scores.

All and MLE iterations: Choose to display the factor loadings, factor scores and information on the iterations if a maximum likelihood estimation was used.

Sort loading: Check to sort the loadings in the Session window (within a factor if the maximum absolute loading occurs there).

Zero loading less than: Check to enter a value. Loadings less than this value will be displayed as zero.

Example of Factor Analysis, Using Maximum Likelihood and a Rotation

Two factors were chosen as the number to represent the census tract data of the Example of Factor Analysis Using Principal Components. You perform a maximum likelihood extraction and varimax rotation to interpret the factors.

1 Open the worksheet EXH_MVAR.MTW.

2 Choose **Stat > Multivariate > Factor Analysis**.

3 In **Variables**, enter *Pop-Home*.

4 In **Number of factors to extract**, enter *2*.

5 Under **Method of Extraction**, choose **Maximum likelihood**.

6 Under **Type of Rotation**, choose **Varimax**.

7 Click **Graphs** and check **Loading plot for first 2 factors**.

8 Click **Results** and check **Sort loadings**. Click **OK** in each dialog box.

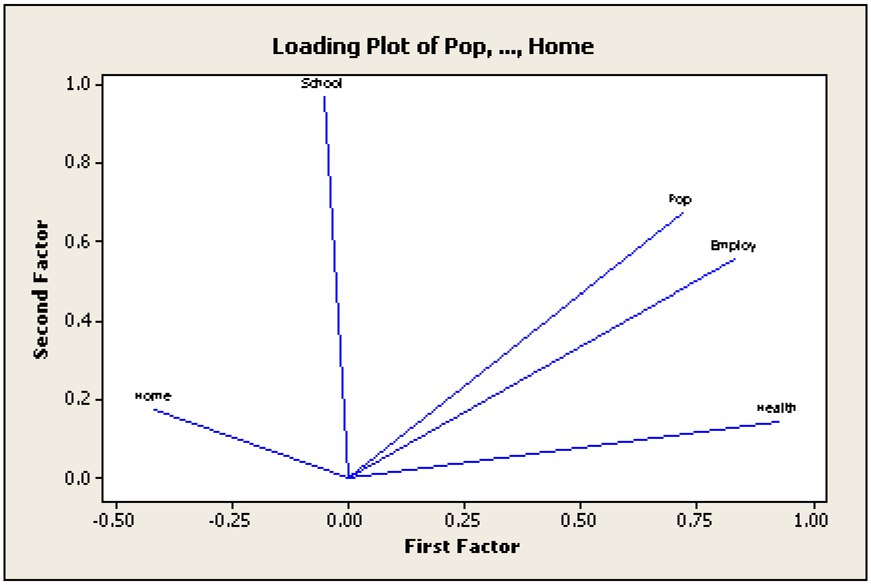
Session window output

Factor Analysis: Pop, School, Employ, Health, Home

```
Maximum Likelihood Factor Analysis of the Correlation Matrix

* NOTE * Heywood case

Unrotated Factor Loadings and Communalities

Variable  Factor1  Factor2  Communality
Pop         0.971    0.160        0.968
School      0.494    0.833        0.938
Employ      1.000    0.000        1.000
Health      0.848   -0.395        0.875
Home       -0.249    0.375        0.202
```

```
Variance   2.9678   1.0159       3.9837
% Var       0.594    0.203        0.797


Rotated Factor Loadings and Communalities
Varimax Rotation

Variable  Factor1  Factor2  Communality
Pop         0.718    0.673        0.968
School     -0.052    0.967        0.938
Employ      0.831    0.556        1.000
Health      0.924    0.143        0.875
Home       -0.415    0.173        0.202

Variance   2.2354   1.7483       3.9837
% Var       0.447    0.350        0.797


Sorted Rotated Factor Loadings and Communalities

Variable  Factor1  Factor2  Communality
Health      0.924    0.143        0.875
Employ      0.831    0.556        1.000
Pop         0.718    0.673        0.968
Home       -0.415    0.173        0.202
School     -0.052    0.967        0.938

Variance   2.2354   1.7483       3.9837
% Var       0.447    0.350        0.797


Factor Score Coefficients


Variable  Factor1  Factor2
Pop        -0.165    0.246
School     -0.528    0.789
Employ      1.150    0.080
Health      0.116   -0.173
Home       -0.018    0.027
```

Graph window output



Interpreting the results

The results indicates that this is a Heywood case. There are three tables of loadings and communalities: unrotated, rotated, and sorted and rotated. The unrotated factors explain 79.7% of the data variability (see last line under Communality) and the communality values indicate that all variables but Home are well represented by these two factors (communalities are 0.202 for Home, 0.875-1.0 for other variables). The percent of total variability represented by the factors does not change with rotation, but after rotating, these factors are more evenly balanced in the percent of variability that they represent, being 44.7% and 35.0%, respectfully.

Sorting is done by the maximum absolute loading for any factor. Variables that have their highest absolute loading on factor 1 are printed first, in sorted order. Variables with their highest absolute loadings on factor 2 are printed next, in sorted order, and so on. Factor 1 has large positive loadings on Health (0.924), Employ (0.831), and Pop (0.718), and a -0.415 loading on Home while the loading on School is small. Factor 2 has a large positive loading on School of 0.967 and loadings of 0.556 and 0.673, respectively, on Employ and Pop, and small loadings on Health and Home.

You can view the rotated loadings graphically in the loadings plot. What stands out for factor 1 are the high loadings on the variables Pop, Employ, and Health and the negative loading on Home. School has a high positive loading for factor 2 and somewhat lower values for Pop and Employ.

Let's give a possible interpretation to the factors. The first factor positively loads on population size and on two variables, Employ and Health, that generally increase with population size. It negatively loads on home value, but this may be largely influenced by one point. We might consider factor 1 to be a "health care - population size" factor. The second factor might be considered to be a "education - population size" factor. Both Health and School are correlated with Pop and Employ, but not much with each other.

In addition, Minitab displays a table of factor score coefficients. These show you how the factors are calculated. Minitab calculates factor scores by multiplying factor score coefficients and your data after they have been centered by subtracting means.

You might repeat this factor analysis with three factors to see if it makes more sense for your data.

Example of Factor Analysis, Using Principal Components

You record the following characteristics of 14 census tracts: total population (Pop), median years of schooling (School), total employment (Employ), employment in health services (Health), and median home value (Home) (data from [6], Table 8.2). You would like to investigate what "factors" might explain most of the variability. As the first step in your factor analysis, you use the principal components extraction method and examine an eigenvalues (scree) plot in order to help you to decide upon the number of factors.

1 Open the worksheet EXH_MVAR.MTW.

2 Choose **Stat > Multivariate > Factor Analysis**.

3 In **Variables**, enter *Pop-Home*.

4 Click **Graphs** and check **Scree plot**. Click **OK** in each dialog box.

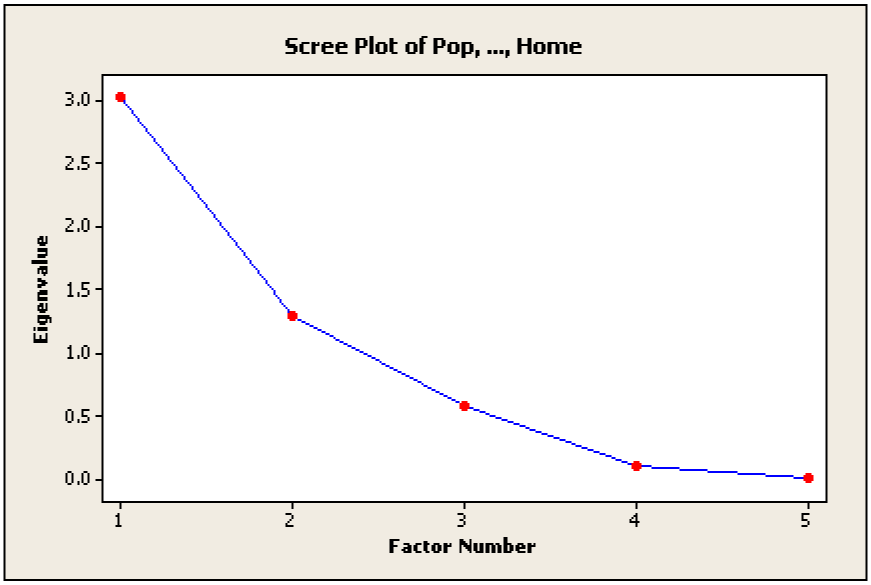
Session window output

Factor Analysis: Pop, School, Employ, Health, Home

```
Principal Component Factor Analysis of the Correlation Matrix


Unrotated Factor Loadings and Communalities

Variable  Factor1  Factor2  Factor3  Factor4  Factor5  Communality
Pop        -0.972   -0.149    0.006    0.170   -0.067        1.000
School     -0.545   -0.715   -0.415   -0.140    0.001        1.000
Employ     -0.989   -0.005    0.089    0.083    0.085        1.000
Health     -0.847    0.352    0.344   -0.200   -0.022        1.000
Home        0.303   -0.797    0.523    0.005    0.002        1.000

Variance   3.0289   1.2911   0.5725   0.0954   0.0121       5.0000
% Var       0.606    0.258    0.114    0.019    0.002        1.000


Sorted Unrotated Factor Loadings and Communalities

Variable  Factor1  Factor2  Factor3  Factor4  Factor5  Communality
Employ     -0.989   -0.005    0.089    0.083    0.085        1.000
Pop        -0.972   -0.149    0.006    0.170   -0.067        1.000
Health     -0.847    0.352    0.344   -0.200   -0.022        1.000
Home        0.303   -0.797    0.523    0.005    0.002        1.000
```

```
School     -0.545   -0.715   -0.415   -0.140    0.001        1.000

Variance   3.0289   1.2911   0.5725   0.0954   0.0121       5.0000
% Var       0.606    0.258    0.114    0.019    0.002        1.000


Factor Score Coefficients

Variable  Factor1  Factor2  Factor3  Factor4  Factor5
Pop        -0.321   -0.116    0.011    1.782   -5.511
School     -0.180   -0.553   -0.726   -1.466    0.060
Employ     -0.327   -0.004    0.155    0.868    6.988
Health     -0.280    0.272    0.601   -2.098   -1.829
Home        0.100   -0.617    0.914    0.049    0.129
```

Graph window output



Interpreting the results

Five factors describe these data perfectly, but the goal is to reduce the number of factors needed to explain the variability in the data. Examine the Session window results line of % Var or the eigenvalues plot. The proportion of variability explained by the last two factors is minimal (0.019 and 0.002, respectively) and they can be eliminated as being important. The first two factors together represent 86% of the variability while three factors explain 98% of the variability. The question is whether to use two or three factors. The next step might be to perform separate factor analyses with two and three factors and examine the communalities to see how individual variables are represented. If there were one or more variables not well represented by the more parsimonious two factor model, you might select a model with three or more factors.

See the example below for a rotation of loadings extracted by the maximum likelihood method with a selection of two factors.

Cluster Observations

Cluster Observations

Stat > Multivariate > Cluster Observations

Use clustering of observations to classify observations into groups when the groups are initially not known.

This procedure uses an agglomerative hierarchical method that begins with all observations being separate, each forming its own cluster. In the first step, the two observations closest together are joined. In the next step, either a third observation joins the first two, or two other observations join together into a different cluster. This process will continue until all clusters are joined into one, however this single cluster is not useful for classification purposes. Therefore you must decide how many groups are logical for your data and classify accordingly. See Determining the final cluster grouping for more information.

Dialog box items

Variables or distance matrix: Enter either the columns containing measurement data or a stored distance matrix on which to perform the hierarchical clustering of observations.

Linkage Method: Choose the linkage method that will determine how the distance between two clusters is defined.

Distance Measure: Choose the distance measure to use if you selected columns as input variables.

Standardize variables: Check to standardize all variables by subtracting the means and dividing by the standard deviation before the distance matrix is calculated–a good idea if variables are in different units and you wish to minimize the effect of scale differences. If you standardize, cluster centroids and distance measures are in standardized variable space.

Specify Final Partition by

Number of Clusters: Choose to determine the final partition by a specified number of clusters. Enter this number in the box. See Determining the final cluster grouping.

Similarity Level: Choose to determine the final partition by the specified level of similarity. Enter this value in the box. See Determining the final cluster grouping.

Show Dendrogram: Check to display the dendrogram or tree diagram, showing the amalgamation steps. Use <Customize> to change the default display of the dendrogram.

<Customize>

<Storage>

Data – Cluster Observations

You can have two types of input data: columns of raw data or a matrix of distances.

Typically, you would use raw data. Each row contains measurements on a single item or subject. You must have two or more numeric columns, with each column representing a different measurement. You must delete rows with missing data from the worksheet before using this procedure.

If you store an n x n distance matrix, where n is the number of observations, you can use this matrix as input data. The (i, j) entry in this matrix is the distance between observations i and j. If you use the distance matrix as input, statistics on the final partition are not available.

To perform clustering of observations

1 Choose **Stat > Multivariate > Cluster Observations**.

2 In **Variables or distance matrix**, enter either columns containing the raw (measurement) data or a matrix of distances.

3 If you like, use any dialog box options, then click **OK**.

Cluster Observations – Dendrogram – Customize

Stat > Multivariate > Cluster Observations > Show Dendrogram > Customize

Allows you to add a title and control y-axis labeling and displaying for the dendrogram.

Double-click the dendrogram after you create it to specify the line type, color, and size for the cluster groups. See Graph Editing Overview.

Dialog box items

Title: To display a title above the dendrogram, type the desired text in this box.

Case labels: Enter a column of case labels. This column must be the same length as the data column.

Label Y Axis with

Similarity: Choose to display similarities on the y-axis.

Distance: Choose to display distances on the y-axis.

Show Dendrogram in

One graph: Choose to display the dendrogram in a single graph window.

Maximum number of observations per graph (without splitting a group): Choose to display a specified number of observation per graph and enter an integer greater than or equal to 1.

Deciding Which Distance Measures and Linkage Methods to Use – Cluster Observations

Distance Measures

If you do not supply a distance matrix, Minitab's first step is to calculate an n x n distance matrix, D, where n is the number of observations. The matrix entries, d(i, j), in row i and column j, is the distance between observations i and j.

Minitab provides five different methods to measure distance. You might choose the distance measure according to properties of your data.

- The Euclidean method is a standard mathematical measure of distance (square root of the sum of squared differences).
- The Pearson method is a square root of the sum of square distances divided by variances. This method is for standardizing.
- Manhattan distance is the sum of absolute distances, so that outliers receive less weight than they would if the Euclidean method were used.
- The squared Euclidean and squared Pearson methods use the square of the Euclidean and Pearson methods, respectfully. Therefore, the distances that are large under the Euclidean and Pearson methods will be even larger under the squared Euclidean and squared Pearson methods.

Tip If you choose Average, Centroid, Median, or Ward as the linkage method, it is generally recommended [9] that you use one of the squared distance measures.

Linkage methods

The linkage method that you choose determines how the distance between two clusters is defined. At each amalgamation stage, the two closest clusters are joined. At the beginning, when each observation constitutes a cluster, the distance between clusters is simply the inter-observation distance. Subsequently, after observations are joined together, a linkage rule is necessary for calculating inter-cluster distances when there are multiple observations in a cluster.

You may wish to try several linkage methods and compare results. Depending on the characteristics of your data, some methods may provide "better" results than others.

- With single linkage, or "nearest neighbor," the distance between two clusters is the minimum distance between an observation in one cluster and an observation in the other cluster. Single linkage is a good choice when clusters are clearly separated. When observations lie close together, single linkage tends to identify long chain-like clusters that can have a relatively large distance separating observations at either end of the chain [6].
- With average linkage, the distance between two clusters is the mean distance between an observation in one cluster and an observation in the other cluster. Whereas the single or complete linkage methods group clusters based upon single pair distances, average linkage uses a more central measure of location.
- With centroid linkage, the distance between two clusters is the distance between the cluster centroids or means. Like average linkage, this method is another averaging technique.
- With complete linkage, or "furthest neighbor," the distance between two clusters is the maximum distance between an observation in one cluster and an observation in the other cluster. This method ensures that all observations in a cluster are within a maximum distance and tends to produce clusters with similar diameters. The results can be sensitive to outliers [10].
- With median linkage, the distance between two clusters is the median distance between an observation in one cluster and an observation in the other cluster. This is another averaging technique, but uses the median rather than the mean, thus downweighting the influence of outliers.
- With McQuitty's linkage, when two clusters are be joined, the distance of the new cluster to any other cluster is calculated as the average of the distances of the soon to be joined clusters to that other cluster. For example, if clusters 1 and 3 are to be joined into a new cluster, say 1*, then the distance from 1* to cluster 4 is the average of the distances from 1 to 4 and 3 to 4. Here, distance depends on a combination of clusters rather than individual observations in the clusters.
- With Ward's linkage, the distance between two clusters is the sum of squared deviations from points to centroids. The objective of Ward's linkage is to minimize the within-cluster sum of squares. It tends to produce clusters with similar numbers of observations, but it is sensitive to outliers [10]. In Ward's linkage, it is possible for the distance between two clusters to be larger than dmax, the maximum value in the original distance matrix. If this happens, the similarity will be negative.

Determining the Final Grouping of Clusters

The final grouping of clusters (also called the final partition) is the grouping of clusters which will, hopefully, identify groups whose observations or variables share common characteristics. The decision about final grouping is also called cutting the dendrogram. The complete dendrogram (tree diagram) is a graphical depiction of the amalgamation of observations or variables into one cluster. Cutting the dendrogram is akin to drawing a line across the dendrogram to specify the final grouping.

How do you know where to cut the dendrogram? You might first execute cluster analysis without specifying a final partition. Examine the similarity and distance levels in the Session window results and in the dendrogram. You can view the similarity levels by placing your mouse pointer over a horizontal line in the dendrogram. The similarity level at any step is the percent of the minimum distance at that step relative to the maximum inter-observation distance in the data. The pattern of how similarity or distance values change from step to step can help you to choose the final grouping. The step where the values change abruptly may identify a good point for cutting the dendrogram, if this makes sense for your data.

After choosing where you wish to make your partition, rerun the clustering procedure, using either **Number of clusters** or **Similarity level** to give you either a set number of groups or a similarity level for cutting the dendrogram. Examine the resulting clusters in the final partition to see if the grouping seems logical. Looking at dendrograms for different final groupings can also help you to decide which one makes the most sense for your data.

Note For some data sets, average, centroid, median and Ward's methods may not produce a hierarchical dendrogram. That is, the amalgamation distances do not always increase with each step. In the dendrogram, such a step will produce a join that goes downward rather than upward.

Cluster Observations – Storage

Stat > Multivariate > Cluster Observations > Storage

Allows you to store cluster membership for each observation, the distance between each observation and each cluster centroid, and the distance matrix.

Dialog box items

Cluster membership column: Enter a single column to store for cluster membership for each observation. This column can then be used as a categorical variable in other Minitab commands.

Distance between observations and cluster centroids (Give a column for each cluster group): Enter storage column(s) for the distance between each observation and each cluster centroid. The number of columns specified must equal the number of cluster in the final partition. The distances stored are always Euclidean distances.

Distance matrix: Enter a storage matrix (M) for the N x N distance matrix, where N is the number of observations. The stored distance matrix can then be used in subsequent commands.

Example of Cluster Observations

You make measurements on five nutritional characteristics (protein, carbohydrate, and fat content, calories, and percent of the daily allowance of Vitamin A) of 12 breakfast cereal brands. The example and data are from p. 623 of [6]. The goal is to group cereal brands with similar characteristics. You use clustering of observations with the complete linkage method, squared Euclidean distance, and you choose standardization because the variables have different units. You also request a dendrogram and assign different line types and colors to each cluster.

1 Open the worksheet CEREAL.MTW.

2 Choose **Stat > Multivariate > Cluster Observations**.

3 In **Variables or distance matrix**, enter *Protein-VitaminA*.

4 From **Linkage Method**, choose **Complete** and from **Distance Measure** choose **Squared Euclidean**.

5 Check **Standardize variables**.

6 Under **Specify Final Partition by**, choose **Number of clusters** and enter *4*.

7 Check **Show dendrogram**.

8 Click **Customize**. In **Title**, enter *Dendrogram for Cereal Data*.

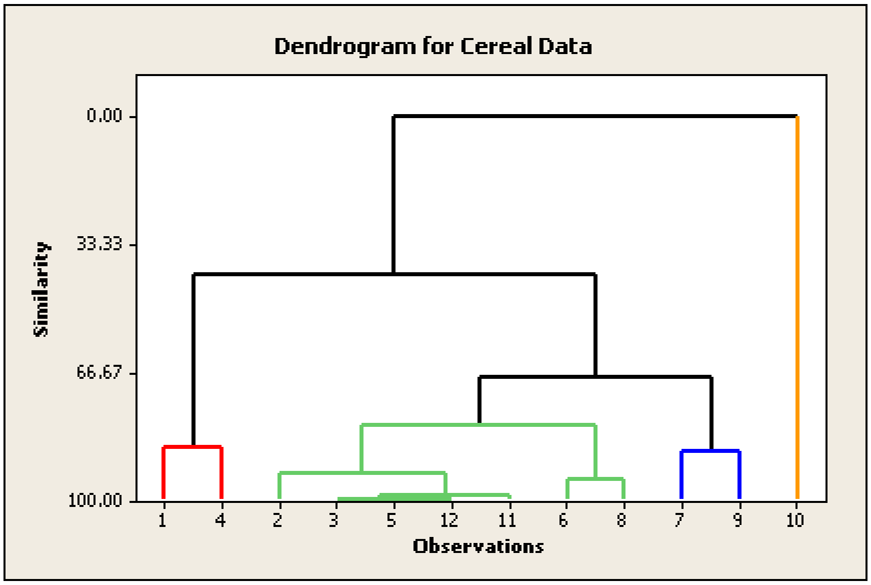
9 Click **OK** in each dialog box.

Session window output

Cluster Analysis of Observations: Protein, Carbo, Fat, Calories, VitaminA

```
Standardized Variables, Squared Euclidean Distance, Complete Linkage
Amalgamation Steps


                                                              Number
        Number                                                of obs.
            of  Similarity  Distance  Clusters        New     in new
Step  clusters       level     level   joined     cluster    cluster
   1        11     100.000    0.0000  5      12         5          2
   2        10      99.822    0.0640  3       5         3          3
   3         9      98.792    0.4347  3     11         3        4
   4         8      94.684    1.9131  6      8         6        2
   5         7      93.406    2.3730  2      3         2        5
   6         6      87.329    4.5597  7      9         7        2
   7         5      86.189    4.9701  1      4         1        2
   8         4      80.601    6.9810  2      6         2        7
   9         3      68.079   11.4873  2      7         2        9
  10         2      41.409   21.0850  1      2         1       11
  11         1       0.000   35.9870  1     10         1       12


Final Partition
Number of clusters: 4


                         Within   Average   Maximum
                        cluster  distance  distance
             Number of   sum of      from      from
          observations  squares  centroid  centroid
Cluster1             2  2.48505   1.11469   1.11469
Cluster2             7  8.99868   1.04259   1.76922
Cluster3             2  2.27987   1.06768   1.06768
Cluster4             1  0.00000   0.00000   0.00000


Cluster Centroids

Variable  Cluster1   Cluster2  Cluster3  Cluster4  Grand centroid
Protein    1.92825  -0.333458  -0.20297  -1.11636       0.0000000
Carbo     -0.75867   0.541908   0.12645  -2.52890       0.0000000
Fat        0.33850  -0.096715   0.33850  -0.67700       0.0000000
Calories   0.28031   0.280306   0.28031  -3.08337      -0.0000000
VitaminA  -0.63971  -0.255883   2.04707  -1.02353      -0.0000000


Distances Between Cluster Centroids

          Cluster1  Cluster2  Cluster3  Cluster4
Cluster1   0.00000   2.67275   3.54180   4.98961

Cluster2   2.67275   0.00000   2.38382   4.72050
Cluster3   3.54180   2.38382   0.00000   5.44603
Cluster4   4.98961   4.72050   5.44603   0.00000
```

Graph window output



Interpreting the results

Minitab displays the amalgamation steps in the Session window. At each step, two clusters are joined. The table shows which clusters were joined, the distance between them, the corresponding similarity level, the identification number of the new cluster (this number is always the smaller of the two numbers of the clusters joined), the number of observations in the new cluster, and the number of clusters. Amalgamation continues until there is just one cluster.

The amalgamation steps show that the similarity level decreases by increments of about 6 or less until it decreases by about 13 at the step from four clusters to three. This indicates that four clusters are reasonably sufficient for the final partition. If this grouping makes intuitive sense for the data, then it is probably a good choice.

When you specify the final partition, Minitab displays three additional tables. The first table summarizes each cluster by the number of observations, the within cluster sum of squares, the average distance from observation to the cluster centroid, and the maximum distance of observation to the cluster centroid. In general, a cluster with a small sum of squares is more compact than one with a large sum of squares. The centroid is the vector of variable means for the observations in that cluster and is used as a cluster midpoint. The second table displays the centroids for the individual clusters while the third table gives distances between cluster centroids.

The dendrogram displays the information in the amalgamation table in the form of a tree diagram. In our example, cereals 1 and 4 make up the first cluster; cereals 2, 3, 5, 12, 11, 6, and 8 make up the second; cereals 7 and 9 make up the third; cereal 10 makes up the fourth.

Cluster Variables

Cluster Variables

Stat > Multivariate > Cluster Variables

Use Clustering of Variables to classify variables into groups when the groups are initially not known. One reason to cluster variables may be to reduce their number. This technique may give new variables that are more intuitively understood than those found using principal components.

This procedure is an agglomerative hierarchical method that begins with all variables separate, each forming its own cluster. In the first step, the two variables closest together are joined. In the next step, either a third variable joins the first two, or two other variables join together into a different cluster. This process will continue until all clusters are joined into one, but you must decide how many groups are logical for your data. See Determining the final grouping.

Dialog box items

Variables or distance matrix: Enter either the columns containing measurement data or a distance matrix on which to perform the hierarchical clustering of variables.

Linkage Method: Choose the linkage method that will determine how the distance between two clusters is defined.

Distance Measure: If you selected columns as input variables, choose the desired distance measure.

Correlation: Choose to use the correlation distance measure.

Absolute correlation: Choose to use the absolute correlation distance measure.

Specify Final Partition by

Number of clusters: Choose to determine the final partition by a specified number of clusters. Enter this number in the box.

Similarity level: Choose to determine the final partition by the specified level of similarity. Enter a value between 0 and 100 in the box.

Show dendrogram: Check to display the dendrogram (tree diagram), showing the amalgamation steps. Use <Customize> to change the default display of the dendrogram.

<Customize>

<Storage>

Data – Cluster Variables

You can have two types of input data to cluster variables: columns of raw data or a matrix of distances.

Typically, you use raw data. Each row contains measurements on a single item or subject. You must have two or more numeric columns, with each column representing a different measurement. Delete rows with missing data from the worksheet before using this procedure.

If you store a p x p distance matrix, where p is the number of variables, you can use the matrix as input data. The (i, j) entry in the matrix is the distance between observations i and j. If you use the distance matrix as input, final partition statistics are not available.

To perform clustering of variables

1. Choose **Stat > Multivariate > Cluster Variables**.
2. In **Variables or distance matrix**, enter either columns containing the raw (measurement) data or a matrix of distances.
3. If you like, use any dialog box options, then click **OK**.

Clustering variables in practice

You must make similar decisions to cluster variables as you would to cluster observations. Follow the guidelines in Determining the final grouping to help you determine groupings. However, if the purpose behind clustering of variables is data reduction, you may decide to use your knowledge of the data to a greater degree in determining the final clusters of variables.

Deciding Which Distance Measures and Linkage Methods to Use – Cluster Variables

Distance Measures

You can use correlations or absolute correlations for distance measures. With the correlation method, the (i,j) entry of the distance matrix is $d_{ij} = 1 - \rho_{ij}$ and for the absolute correlation method, $d_{ij} = 1 - |\rho_{ij}|$, where ρ_{ij} is the (Pearson product moment) correlation between variables i and j. Thus, the correlation method will give distances between 0 and 1 for positive correlations, and between 1 and 2 for negative correlations. The absolute correlation method will always give distances between 0 and 1.

- If it makes sense to consider negatively correlated data to be farther apart than positively correlated data, then use the correlation method.
- If you think that the strength of the relationship is important in considering distance and not the sign, then use the absolute correlation method.

Linkage methods

The linkage method that you choose determines how the distance between two clusters is defined. At each amalgamation stage, the two closest clusters are joined. At the beginning, when each variables constitutes a cluster, the distance between clusters is simply the inter-variables distance. Subsequently, after observations are joined together, a linkage rule is necessary for calculating inter-cluster distances when there are multiple variables in a cluster.

You may wish to try several linkage methods and compare results. Depending on the characteristics of your data, some methods may provide "better" results than others.

- With single linkage, or "nearest neighbor," the distance between two clusters is the minimum distance between a variable in one cluster and a variable in the other cluster. Single linkage is a good choice when clusters are clearly separated. When variables lie close together, single linkage tends to identify long chain-like clusters that can have a relatively large distance separating variables at either end of the chain [6].
- With average linkage, the distance between two clusters is the mean distance between a variable in one cluster and a variable in the other cluster. Whereas the single or complete linkage methods group clusters based upon single pair distances, average linkage uses a more central measure of location.
- With centroid linkage, the distance between two clusters is the distance between the cluster centroids or means. Like average linkage, this method is another averaging technique.
- With complete linkage, or "furthest neighbor," the distance between two clusters is the maximum distance between a variable in one cluster and a variable in the other cluster. This method ensures that all variables in a cluster are within a maximum distance and tends to produce clusters with similar diameters. The results can be sensitive to outliers [10].
- With median linkage, the distance between two clusters is the median distance between a variable in one cluster and a variable in the other cluster. This is another averaging technique, but uses the median rather than the mean, thus downweighting the influence of outliers.
- With McQuitty's linkage, when two clusters are be joined, the distance of the new cluster to any other cluster is calculated as the average of the distances of the soon to be joined clusters to that other cluster. For example, if clusters 1 and 3 are to be joined into a new cluster, say 1*, then the distance from 1* to cluster 4 is the average of the distances from 1 to 4 and 3 to 4. Here, distance depends on a combination of clusters rather than individual variables in the clusters.
- With Ward's linkage, the distance between two clusters is the sum of squared deviations from points to centroids. The objective of Ward's linkage is to minimize the within-cluster sum of squares. It tends to produce clusters with similar numbers of variables, but it is sensitive to outliers [10]. In Ward's linkage, it is possible for the distance between two clusters to be larger than dmax, the maximum value in the original distance matrix. If this happens, the similarity will be negative.

Cluster Variables – Dendrogram – Customize

Stat > Multivariate > Cluster Variables > Show Dendrogram > Customize

Allows you to add a title and control y-axis labeling and displaying for the dendrogram.

Double-click the dendrogram after you create it to specify the line type, color, and size for the cluster groups. See Graph Editing Overview.

Dialog box items

Title: To display a title above the dendrogram, type the desired text in this box.

Label Y Axis with

Similarity: Choose to display similarities on the y-axis.

Distance: Choose to display distances on the y-axis.

Show Dendrogram in

One graph: Choose to display the dendrogram in a single graph window.

Maximum number of variables per graph (without splitting a group): Choose to display a specified number of variables per graph and enter an integer greater than or equal to 1.

Cluster Variables – Storage

Stat > Multivariate > Cluster Variables > Storage

Allows you to store the distance matrix.

Dialog box items

Distance matrix: Specify a storage matrix (M) for the P x P distance matrix (D), where P is the number of variables. The stored distance matrix can then be used in subsequent commands.

Example of Cluster Variables

You conduct a study to determine the long-term effect of a change in environment on blood pressure. The subjects are 39 Peruvian males over 21 years of age who had migrated from the Andes mountains to larger towns at lower elevations. You recorded their age (Age), years since migration (Years), weight in kg (Weight), height in mm (Height), skin fold of the chin, forearm, and calf in mm (Chin, Forearm, Calf), pulse rate in beats per minute (Pulse), and systolic and diastolic blood pressure (Systol, Diastol).

Your goal is to reduce the number of variables by combining variables with similar characteristics. You use clustering of variables with the default correlation distance measure, average linkage and a dendrogram.

1 Open the worksheet PERU.MTW.

2 Choose **Stat > Multivariate > Cluster Variables**.

3 In **Variables or distance matrix**, enter *Age-Diastol*.

4 For **Linkage Method**, choose **Average**.

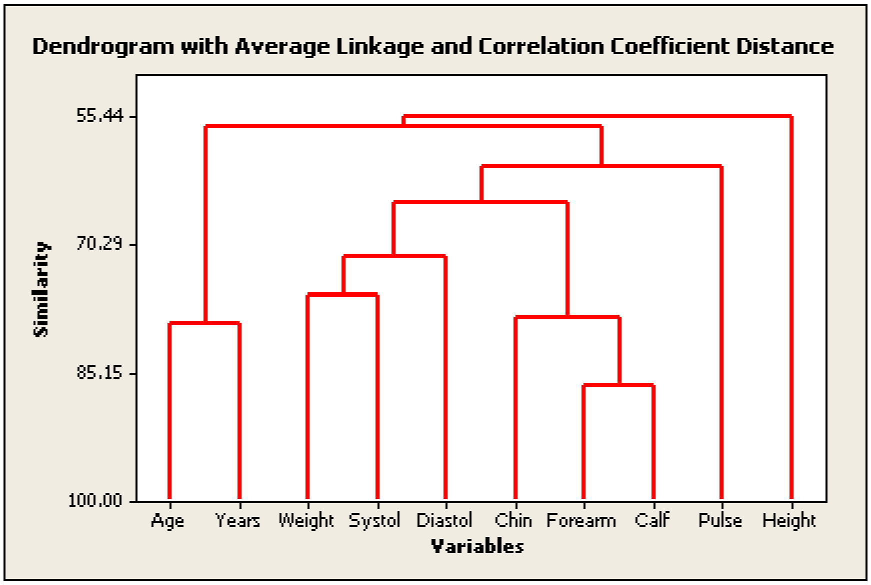
5 Check **Show dendrogram**. Click **OK**.

Session window output

Cluster Analysis of Variables: Age, Years, Weight, Height, Chin, Forearm, ...

```
Correlation Coefficient Distance, Average Linkage
Amalgamation Steps


                                                          Number
        Number                                           of obs.
            of  Similarity  Distance  Clusters      New   in new
Step  clusters       level     level   joined   cluster  cluster
   1         9     86.7763  0.264474  6      7        6        2
   2         8     79.4106  0.411787  1      2        1        2
   3         7     78.8470  0.423059  5      6        5        3
   4         6     76.0682  0.478636  3      9        3        2
   5         5     71.7422  0.565156  3     10        3        3
   6         4     65.5459  0.689082  3      5        3        6
   7          3      61.3391   0.773218   3        8          3         7
   8          2      56.5958   0.868085   1        3          1         9
   9          1      55.4390   0.891221   1        4          1        10
```

Graph window output



Interpreting the results

Minitab displays the amalgamation steps in the Session window. At each step, two clusters are joined. The table shows which clusters were joined, the distance between them, the corresponding similarity level, the identification number of the new cluster (this is always the smaller of the two numbers of the clusters joined), the number of variables in the new cluster and the number of clusters. Amalgamation continues until there is just one cluster.

If you had requested a final partition you would also receive a list of which variables are in each cluster.

The dendrogram displays the information printed in the amalgamation table in the form of a tree diagram. Dendrogram suggest variables which might be combined, perhaps by averaging or totaling. In this example, the chin, forearm, and calf skin fold measurements are similar and you decide to combine those. The age and year since migration variables are similar, but you will investigate this relationship. If subjects tend to migrate at a certain age, then these variables could contain similar information and be combined. Weight and the two blood pressure measurements are similar. You decide to keep weight as a separate variable but you will combine the blood pressure measurements into one.

Cluster K-Means

Cluster K-Means

Stat > Multivariate > Cluster K-Means

Use K-means clustering of observations, like clustering of observations, to classify observations into groups when the groups are initially unknown. This procedure uses non-hierarchical clustering of observations according to MacQueen's algorithm [6]. K-means clustering works best when sufficient information is available to make good starting cluster designations.

Dialog box items

Variables: Enter the columns containing measurement data on which to perform the K-means non-hierarchical clustering of observations.

Specify Partition by: Allows you to specify the initial partition for the K-means algorithm.

> **Number of clusters:** Choose to specify the number of clusters to form. If you enter the number 5, for example, Minitab uses the first 5 observations as initial cluster centroids. Each observation is assigned to the cluster whose centroid it is closest to. Minitab recalculates the cluster centroids each time a cluster gains or loses an observation.
>
> **Initial partition column:** Choose to specify a column containing cluster membership to begin the partition process.

Standardize variables: Check to standardize all variables by subtracting the means and dividing by the standard deviation before the distance matrix is calculated. This is a good idea if the variables are in different units and you wish to minimize the effect of scale differences. If you standardize, cluster centroids and distance measures are in standardized variable space before the distance matrix is calculated.

<Storage>

Data – Cluster K Means

You must use raw data as input to K-means clustering of observations. Each row contains measurements on a single item or subject. You must have two or more numeric columns, with each column representing a different measurement. You must delete rows with missing data from the worksheet before using this procedure.

To initialize the clustering process using a data column, you must have a column that contains a cluster membership value for each observation. The initialization column must contain positive, consecutive integers or zeros (it should not contain all zeros). Initially, each observation is assigned to the cluster identified by the corresponding value in this column. An initialization of zero means that an observation is initially unassigned to a group. The number of distinct positive integers in the initial partition column equals the number of clusters in the final partition.

To perform K-means clustering of observations

1 Choose **Stat > Multivariate > Cluster K-Means**.

2 In **Variables**, enter the columns containing the measurement data.

3 If you like, use any dialog box options, then click **OK**.

Initializing Cluster K-Means Process

K-means clustering begins with a grouping of observations into a predefined number of clusters.

1 Minitab evaluates each observation, moving it into the nearest cluster. The nearest cluster is the one which has the smallest Euclidean distance between the observation and the centroid of the cluster.

2 When a cluster changes, by losing or gaining an observation, Minitab recalculates the cluster centroid.

3 This process is repeated until no more observations can be moved into a different cluster. At this point, all observations are in their nearest cluster according to the criterion listed above.

Unlike hierarchical clustering of observations, it is possible for two observations to be split into separate clusters after they are joined together.

K-means procedures work best when you provide good starting points for clusters [10]. There are two ways to initialize the clustering process: specifying a number of clusters or supplying an initial partition column that contains group codes.

You may be able to initialize the process when you do not have complete information to initially partition the data. Suppose you know that the final partition should consist of three groups, and that observations 2, 5, and 9 belong in each of those groups, respectively. Proceeding from here depends upon whether you specify the number of clusters or supply an initial partition column.

- If you specify the number of clusters, you must rearrange your data in the Data window to move observations 2, 5 and 9 to the top of the worksheet, and then specify 3 for Number of clusters.
- If you enter an initial partition column, you do not need to rearrange your data in the Data window. In the initial partition worksheet column, enter group numbers 1, 2, and 3, for observations 2, 5, and 9, respectively, and enter 0 for the other observations.

The final partition will depend to some extent on the initial partition that Minitab uses. You might try different initial partitions. According to Milligan [10], K-means procedures may not perform as well when the initializations are done arbitrarily. However, if you provide good starting points, K-means clustering may be quite robust.

To initialize the process by specifying the number of clusters

1 Choose **Stat > Multivariate > Cluster K-Means**.

2 In **Variables**, enter the columns containing the measurement data.

3 Under **Specify Partition by**, choose **Number of clusters** and enter a number, k, in the box. Minitab will use the first k observations as initial cluster seeds, or starting locations.

4 Click **OK**.

To initialize the process using a data column

1 Choose **Stat > Multivariate > Cluster K-Means**.

2 In **Variables**, enter the columns containing the measurement data.

3 Under **Specify Partition by**, choose **Initial partition column**. Enter the column containing the initial cluster membership for each observation.

4 Click **OK**.

Cluster K-Means – Storage

Stat > Multivariate > Cluster K-Means > Storage

Allows cluster membership for each observation and the distance between each observation and each cluster centroid.

Dialog box items

Cluster membership column: Enter a single storage column for final cluster membership for each observation. This column can then be used as a categorical variable in other Minitab commands, such as Discriminant Analysis or Plot.

Distance between observations and cluster centroids (Give a column for each cluster group): Enter storage columns for the distance between each observation and each cluster centroid. The number of columns specified must equal the number of clusters specified for the initial partition. The distances stored are Euclidean distances.

Example of Cluster K-Means

You live-trap, anesthetize, and measure one hundred forty-three black bears. The measurements are total length and head length (Length, Head.L), total weight and head weight (Weight, Weight.H), and neck girth and chest girth (Neck.G, Chest.G). You wish to classify these 143 bears as small, medium-sized, or large bears. You know that the second, seventy-eighth, and fifteenth bears in the sample are typical of the three respective categories. First, you create an initial partition column with the three seed bears designated as 1 = small, 2 = medium-sized, 3 = large, and with the remaining bears as 0 (unknown) to indicate initial cluster membership. Then you perform K-means clustering and store the cluster membership in a column named BearSize.

1 Open the worksheet BEARS.MTW.

2 To create the initial partition column, choose **Calc > Make Patterned Data > Simple Set of Numbers**.

3 In **Store patterned data in**, enter *Initial* for the storage column name.

4 In both **From first value** and **From last value**, enter *0*.

5 In **List each value**, type *143*. Click **OK**.

6 Go to the Data window and type *1*, *2*, and *3* in the second, seventy-eighth, and fifteenth rows, respectively, of the column named *Initial*.

7 Choose **Stat > Multivariate > Cluster K-Means**.

8 In **Variables**, enter **'Head.L'–Weight**.

9 Under **Specify Partition by**, choose **Initial partition column** and enter *Initial*.

10 Check **Standardize variables**.

11 Click **Storage**. In **Cluster membership column**, enter *BearSize*.

12 Click **OK** in each dialog box.

Session window output

K-means Cluster Analysis: Head.L, Head.W, Neck.G, Length, Chest.G, Weight

```
Standardized Variables


Final Partition


Number of clusters: 3


                         Within   Average   Maximum
                        cluster  distance  distance
             Number of   sum of      from      from
          observations  squares  centroid  centroid
Cluster1            41   63.075     1.125     2.488
Cluster2            67   78.947     0.997     2.048
Cluster3            35   65.149     1.311     2.449


Cluster Centroids

                                              Grand
Variable  Cluster1  Cluster2  Cluster3  centroid
Head.L     -1.0673    0.0126    1.2261   -0.0000
Head.W     -0.9943   -0.0155    1.1943    0.0000
Neck.G     -1.0244   -0.1293    1.4476   -0.0000
Length     -1.1399    0.0614    1.2177    0.0000
Chest.G    -1.0570   -0.0810    1.3932   -0.0000
Weight     -0.9460   -0.2033    1.4974   -0.0000


Distances Between Cluster Centroids

          Cluster1  Cluster2  Cluster3
Cluster1    0.0000    2.4233    5.8045
Cluster2    2.4233    0.0000    3.4388
Cluster3    5.8045    3.4388    0.0000
```

Interpreting the results

K-means clustering classified the 143 bears as 41 small bears, 67 medium-size bears, and 35 large bears. Minitab displays, in the first table, the number of observations in each cluster, the within cluster sum of squares, the average distance from observation to the cluster centroid, and the maximum distance of observation to the cluster centroid. In general, a cluster with a small sum of squares is more compact than one with a large sum of squares. The centroid is the vector of variable means for the observations in that cluster and is used as a cluster midpoint.

The centroids for the individual clusters are displayed in the second table while the third table gives distances between cluster centroids.

The column BearSize contains the cluster designations.

Discriminant Analysis

Discriminant Analysis

Stat > Multivariate > Discriminant Analysis

Use discriminant analysis to classify observations into two or more groups if you have a sample with known groups. Discriminant analysis can also used to investigate how variables contribute to group separation.

Minitab offers both linear and quadratic discriminant analysis. With linear discriminant analysis, all groups are assumed to have the same covariance matrix. Quadratic discrimination does not make this assumption but its properties are not as well understood.

In the case of classifying new observations into one of two categories, logistic regression may be superior to discriminant analysis [3], [11].

Dialog box items

Groups: Choose the column containing the group codes. There may be up to 20 groups.

Predictors: Choose the column(s) containing the measurement variables or predictors.

Discriminant Function

Linear: Choose to perform linear discriminant analysis. All groups are assumed to have the same covariance matrix.

Quadratic: Choose to perform quadratic discriminant analysis. No assumption is made about the covariance matrix; its properties are not as well understood.

Use cross validation: Check to perform the discrimination using cross-validation. This technique is used to compensate for an optimistic apparent error rate.

Storage

Linear discriminant function: Enter storage columns for the coefficients from the linear discriminant function, using one column for each group. The constant is stored at the top of each column.

Fits: Check to store the fitted values. The fitted value for an observation is the group into which it is classified.

Fits from cross validation: Check to store the fitted values if discrimination was done using cross-validation.

<Options>

Data – Discriminant Analysis

Set up your worksheet so that a row of data contains information about a single item or subject. You must have one or more numeric columns containing measurement data, or predictors, and a single grouping column containing up to 20 groups. The column of group codes may be numeric, text, or date/time. If you wish to change the order in which text groups are processed from their default alphabetized order, you can define your own order. (See Ordering Text Categories). Minitab automatically omits observations with missing measurements or group codes from the calculations.

If a high degree of multicollinearity exists (i.e., if one or more predictors is highly correlated with another) or one or more of the predictors is essential constant, discriminant analysis calculations cannot be done and Minitab displays a message to that effect.

To perform linear discriminant analysis

1 Choose **Stat > Multivariate > Discriminant Analysis**.

2 In **Groups**, enter the column containing the group codes.

3 In **Predictors**, enter the column or columns containing the measurement data.

4 If you like, use any dialog box options, then click **OK**.

Linear discriminant analysis

An observation is classified into a group if the squared distance (also called the Mahalanobis distance) of observation to the group center (mean) is the minimum. An assumption is made that covariance matrices are equal for all groups. There is a unique part of the squared distance formula for each group and that is called the linear discriminant function for that group. For any observation, the group with the smallest squared distance has the largest linear discriminant function and the observation is then classified into this group.

Linear discriminant analysis has the property of symmetric squared distance: the linear discriminant function of group i evaluated with the mean of group j is equal to the linear discriminant function of group j evaluated with the mean of group i.

We have described the simplest case, no priors and equal covariance matrices. If you consider Mahalanobis distance a reasonable way to measure the distance of an observation to a group, then you do not need to make any assumptions about the underlying distribution of your data.

See Prior Probabilities for more information.

Quadratic discriminant analysis

There is no assumption with quadratic discriminant analysis that the groups have equal covariance matrices. As with linear discriminant analysis, an observation is classified into the group that has the smallest squared distance. However, the squared distance does not simplify into a linear function, hence the name quadratic discriminant analysis.

Unlike linear distance, quadratic distance is not symmetric. In other words, the quadratic discriminant function of group i evaluated with the mean of group j is not equal to the quadratic discriminant function of group j evaluated with the mean of group i. On the results, quadratic distance is called the generalized squared distance. If the determinant of the sample group covariance matrix is less than one, the generalized squared distance can be negative.

Cross-Validation

Cross-validation is one technique that is used to compensate for an optimistic apparent error rate. The apparent error rate is the percent of misclassified observations. This number tends to be optimistic because the data being classified are the same data used to build the classification function.

The cross-validation routine works by omitting each observation one at a time, recalculating the classification function using the remaining data, and then classifying the omitted observation. The computation time takes approximately four times longer with this procedure. When cross-validation is performed, Minitab displays an additional summary table.

Another technique that you can use to calculate a more realistic error rate is to split your data into two parts. Use one part to create the discriminant function, and the other part as a validation set. Predict group membership for the validation set and calculate the error rate as the percent of these data that are misclassified.

Prior Probabilities

Sometimes items or subjects from different groups are encountered according to different probabilities. If you know or can estimate these probabilities a priori, discriminant analysis can use these so-called prior probabilities in calculating the posterior probabilities, or probabilities of assigning observations to groups given the data. With the assumption that the data have a normal distribution, the linear discriminant function is increased by $\ln(p_i)$, where p_i is the prior probability of group i. Because observations are assigned to groups according to the smallest generalized distance, or equivalently the largest linear discriminant function, the effect is to increase the posterior probabilities for a group with a high prior probability.

Now suppose we have priors and suppose $f_i(x)$ is the joint density for the data in group i (with the population parameters replaced by the sample estimates).

The posterior probability is the probability of group i given the data and is calculated by

$$\frac{p_i f_i(x)}{\sum_i p_i f_i(x)}$$

The largest posterior probability is equivalent to the largest value of $\ln[p_i f_i(x)]$.

If $f_i(x)$ is the normal distribution, then

$$\ln[p_i f_i(x)] = -0.5\,[d_i^2(x) - 2\ln p_i] - \text{(a constant)}$$

The term in square brackets is called the generalized squared distance of x to group i and is denoted by $d_i^2(x)$. Notice,

$$d_i^2(x) = -2\,[\mathbf{m}_i'\,\mathbf{S}_p^{-1}\,\mathbf{x} - 0.5\,\mathbf{m}_i'\,\mathbf{S}_p^{-1}\,\mathbf{m}_i + \ln p_i] + \mathbf{x}'\,\mathbf{S}_p^{-1}\,\mathbf{x}$$

The term in square brackets is the linear discriminant function. The only difference from the non-prior case is a change in the constant term. Notice, the largest posterior is equivalent to the smallest generalized distance, which is equivalent to the largest linear discriminant function.

Predicting group membership for new observations

Generally, discriminant analysis is used to calculate the discriminant functions from observations with known groups. When new observations are made, you can use the discriminant function to predict which group that they belong to. You can do this by either calculating (using **Calc > Calculator**) the values of the discriminant function for the observation(s) and then assigning it to the group with the highest function value or by using Minitab's discriminant procedure. See To predict group membership for new observations.

To predict group membership for new observations

1 Choose **Stat > Multivariate > Discriminant Analysis**.

2 In **Groups**, enter the column containing the group codes from the original sample.

3 In **Predictors**, enter the column(s) containing the measurement data of the original sample.

4 Click **Options**. In **Predict group membership for**, enter constants or columns representing one or more observations. The number of constants or columns must be equivalent to the number of predictors.

5 If you like, use any dialog box options, and click **OK**.

Discriminant Analysis – Options

Stat > Multivariate > Discriminant Analysis > Options

Allows you to specify prior probabilities, predict group membership for new observations, and control the display of Session window output.

Dialog box items

Prior probabilities: Enter prior probabilities. You may type in the probabilities or specify constants (K) that contain stored values. There should be one value for each group. The first value will be assigned to the group with the smallest code, the second to the group with the second smallest code, etc. If the probabilities do not sum to one, Minitab normalizes them. See Prior Probabilities.

Predict group membership for: Enter values for predicting group membership for new observations.

Display of Results:

Do not display: Choose to suppress all results. Requested storage is done.

Classification matrix: Choose to display only the classification matrix.

Above plus ldf, distances, and misclassification summary: Choose to display the classification matrix, the squared distance between group centers, linear discriminant function, and a summary of misclassified observations.

Above plus mean, std. dev., and covariance summary: Choose to display the classification matrix, the squared distance between group centers, linear discriminant function, a summary of misclassified observations, means, standard deviations, and covariance matrices, for each group and pooled.

Above plus complete classification summary: Choose to display the classification matrix, the squared distance between group centers, linear discriminant function, a summary of misclassified observations, means, standard deviations, covariance matrices, for each group and pooled, and a summary of how all observations were classified. Minitab notes misclassified observations with two asterisks beside the observation number.

Example of Discriminant Analysis

In order to regulate catches of salmon stocks, it is desirable to identify fish as being of Alaskan or Canadian origin. Fifty fish from each place of origin were caught and growth ring diameters of scales were measured for the time when they lived in freshwater and for the subsequent time when they lived in saltwater. The goal is to be able to identify newly-caught fish as being from Alaskan or Canadian stocks. The example and data are from [6], page 519-520.

1 Open the worksheet EXH_MVAR.MTW.

2 Choose **Stat > Multivariate > Discriminant Analysis**.

3 In **Groups**, enter *SalmonOrigin*.

4 In **Predictors**, enter *Freshwater Marine*. Click **OK**.

Session window output

Discriminant Analysis: SalmonOrigin versus Freshwater, Marine

```
Linear Method for Response: SalmonOrigin


Predictors: Freshwater, Marine


Group    Alaska    Canada
Count        50        50


Summary of classification

                  True Group
Put into Group  Alaska  Canada
Alaska              44       1
Canada               6      49
Total N             50      50
N correct           44      49
Proportion       0.880   0.980

N = 100           N Correct = 93           Proportion Correct = 0.930
```

```
Squared Distance Between Groups

         Alaska   Canada
Alaska  0.00000  8.29187
Canada  8.29187  0.00000


Linear Discriminant Function for Groups

             Alaska  Canada
Constant    -100.68  -95.14
Freshwater     0.37    0.50
Marine         0.38    0.33


Summary of Misclassified Observations

                                                   Squared
Observation    True Group  Pred Group   Group  Distance  Probability
          1**      Alaska      Canada  Alaska     3.544        0.428
                                       Canada     2.960        0.572
          2**      Alaska      Canada  Alaska    8.1131        0.019
                                       Canada    0.2729        0.981

         12**      Alaska      Canada  Alaska    4.7470        0.118
                                       Canada    0.7270        0.882
         13**      Alaska      Canada  Alaska    4.7470        0.118
                                       Canada    0.7270        0.882
         30**      Alaska      Canada  Alaska     3.230        0.289
                                       Canada     1.429        0.711
         32**      Alaska      Canada  Alaska     2.271        0.464
                                       Canada     1.985        0.536
         71**      Canada      Alaska  Alaska     2.045        0.948
                                       Canada     7.849        0.052
```

Interpreting the results

As shown in the Summary of Classification table, the discriminant analysis correctly identified 93 of 100 fish, though the probability of correctly classifying an Alaskan fish was lower (44/50 or 88%) than was the probability of correctly classifying a Canadian fish (49/50 or 98%). To identify newly-caught fish, you could compute the linear discriminant functions associated with Alaskan and Canadian fish and identify the new fish as being of a particular origin depending upon which discriminant function value is higher. You can either do this by using **Calc > Calculator** using stored or output values, or performing discriminant analysis again and predicting group membership for new observations.

The Summary of Misclassified Observations table shows the squared distances from each misclassified point to group centroids and the posterior probabilities. The squared distance value is that value from observation to the group centroid, or mean vector. The probability value is the posterior probability. Observations are assigned to the group with the highest posterior probability.

Simple Correspondence Analysis

Simple Correspondence Analysis

Stat > Multivariate > Simple Correspondence Analysis

Simple correspondence analysis helps you to explore relationships in a two-way classification. Simple correspondence analysis can also operate on three-way and four-way tables because they can be collapsed into two-way tables. This procedure decomposes a contingency table in a manner similar to how principal components analysis decomposes multivariate continuous data. An eigen analysis of the data is performed, and the variability is broken down into underlying dimensions and associated with rows and/or columns.

Dialog box items

Input Data

Categorical variables: Choose to enter the data as categorical variables. If you do not use the Combine subdialog box, enter two worksheet columns. The first is for the row categories; the second is for the column categories. Minitab then forms a contingency table from the input data.

Columns of a contingency table: Choose to enter the data as columns of a contingency table. Each worksheet column you enter will be used as one column of the contingency table. All values in the contingency columns must be positive integers or zero.

Row names: Enter a column that contains names for the rows of the contingency table. The name column must be a text column whose length matches the number of rows in the contingency table. Minitab prints the first 8 characters of the names in tables, but prints the full name on graphs. If you do not enter names here, the rows will be named Row1, Row2, etc.

Column names: Enter a column that contains names for the columns of the contingency table. The name column must be a text column whose length matches the number of columns in the contingency table. Minitab prints the first 8 characters of the names in tables, but prints the full name on graphs. If you do not enter names here, the columns will be named Column1, Column2, etc.

Number of components: Enter the number of components to calculate. The minimum number of components is one. The maximum number of components for a contingency table with r rows and c columns is the smaller of (r-1) or (c-1), which is equivalent to the dimension of the subspace onto which you project the profiles. The default number of components is 2.

<Combine>

<Supp Data>

<Results>

<Graphs>

<Storage>

Data – Simple Correspondence Analysis

Worksheet data may be arranged in two ways: raw or contingency table form. See Arrangement of Input Data. Worksheet data arrangement determines acceptable data values.

- If your data are in raw form, you can have two, three, or four classification columns with each row representing one observation. All columns must be the same length. The data represent categories and may be numeric, text, or date/time. If the categories in a column are text data, the levels are used in the order of first occurrence, i.e., the first level becomes the first row (column) of the table, the next distinct level becomes the second row (column) of the table, and so on. If you wish to change the order in which text categories are processed from their default alphabetized order, you can define your own order. See Ordering Text Categories. You must delete missing data before using this procedure. Because simple correspondence analysis works with a two-way classification, the standard approach is to use two worksheet columns. However, you can obtain a two-way classification with three or four variables by crossing variables within the simple correspondence analysis procedure. See Crossing variables to create a two-way table.
- If your data are in contingency table form, worksheet columns must contain integer frequencies of your category combinations. You must delete any rows or columns with missing data or combine them with other rows or columns. Unlike the χ^2 test for association procedure, there is no set limit on the number of contingency table columns. You could use simple correspondence analysis to obtain χ^2 statistics for large tables.

Supplementary data

When performing a simple correspondence analysis, you have a main classification set of data on which you perform your analysis. However, you may also have additional or supplementary data in the same form as the main set, because you can see how these supplementary data are "scored" using the results from the main set. These supplementary data may be further information from the same study, information from other studies, or target profiles [4]. Minitab does not include these data when calculating the components, but you can obtain a profile and display supplementary data in graphs.

You can have row supplementary data or column supplementary data. Row supplementary data constitutes an additional row(s) of the contingency table, while column supplementary data constitutes an additional column(s) of the contingency table. Supplementary data must be entered in contingency table form. Therefore, each worksheet column of these data must contain c entries (where c is the number of contingency table columns) or r entries (where r is the number of contingency table rows).

To perform a simple correspondence analysis

1. Choose **Stat > Multivariate > Simple Correspondence Analysis**.
2. How you enter your data depends on the form of the data and the number of categorical variables.
   - If you have two categorical variables, do one of the following:
     – For raw data, enter the columns containing the raw data in **Categorical variables**.
     – For contingency table data, enter the columns containing the data in **Columns of a contingency table**.
   - If you have three or four categorical variables, you must cross some variables before entering data as shown above. See Crossing variables to create a two-way table.
3. If you like, use any dialog box options, then click **OK**.

Crossing variables to create a two-way table

Crossing variables allows you to use simple correspondence analysis to analyze three-way and four-way contingency tables. You can cross the first two variables to form rows and/or the last two variables to form columns. You must enter three categorical variables to perform one cross, and four categorical variables to perform two crosses.

The following example illustrates row crossing. Column crossing is similar. Suppose you have two variables. The row variable, Sex, has two levels: male and female; the column variable, Age, has three levels; young, middle aged, old.

Crossing Sex with Age will create 2 x 3 = 6 rows, ordered as follows:

male / young
male / middle aged
male / old
female / young
female / middle aged
female / old

Simple Correspondence Analysis – Combine

Stat > Multivariate > Simple Correspondence Analysis > Combine

Crossing variables allows you to use simple correspondence analysis to analyze three-way and four-way contingency tables. You can cross the first two variables to form rows and/or the last two variables to form columns. You must enter three categorical variables to perform one cross, and four categorical variables to perform two crosses.

In order to cross columns, you must choose **Categorical variables** for **Input Data** rather than **Columns of a contingency table** in the main dialog box. If you want to cross for either just the rows or for just the columns of the contingency table, you must enter three worksheet columns in the **Categorical variables** text box. If you want to cross both the rows and the columns of the table, you must specify four worksheet columns in this text box.

Dialog box items

Define Rows of the Contingency Table Using:

First variable: Choose to use the first input column to form the rows of the contingency table. Thus, the rows of the contingency table are not formed by crossing variables.

First 2 variables crossed: Choose to cross the categories in the first two input columns to form the rows of the contingency table. For example, if the first variable is Sex (with 2 levels, male and female) and the second variable is Age (with 3 levels, young, middle aged, old), then there will be 2 x 3 = 6 rows, ordered as follows:

males / young
males / middle aged
males / old
females / young
females / middle aged
females / old

Define Columns of Contingency Table Using:

Last variable: Choose to use the last input column to form the columns of the contingency table.

Last 2 variables crossed: Choose to cross the categories in the last two input columns to form the columns of the contingency table.

Simple Correspondence Analysis – Supplementary Data

Stat > Multivariate > Simple Correspondence Analysis > Supp Data

When performing a simple correspondence analysis, you have a main classification set of data on which you perform your analysis. However, you may also have additional or supplementary data in the same form as the main set, because you can see how these supplementary data are "scored" using the results from the main set. See What are Supplementary Data?

Dialog box items

Supplementary Rows: Enter one or more columns containing additional rows of the contingency table.

Supplementary Columns: Enter one or more columns containing additional columns of the contingency table.

Row names: Enter a column containing text names for the supplementary rows.

Column names: Enter a column containing text names for the supplementary columns.

What are Supplementary Data?

When performing a simple correspondence analysis, you have a main classification set of data on which you perform your analysis. However, you may also have additional or supplementary data in the same form as the main set, because you can see how these supplementary data are "scored" using the results from the main set. These supplementary data may be further information from the same study, information from other studies, or target profiles [4]. Minitab does not include these data when calculating the components, but you can obtain a profile and display supplementary data in graphs.

You can have row supplementary data or column supplementary data. Row supplementary data constitutes an additional row(s) of the contingency table, while column supplementary data constitutes an additional column(s) of the contingency table. Supplementary data must be entered in contingency table form. Therefore, each worksheet column of these data must contain c entries (where c is the number of contingency table columns) or r entries (where r is the number of contingency table rows).

Simple Correspondence Analysis – Results

Stat > Multivariate > Simple Correspondence Analysis > Results

Allows you to control displayed output.

Dialog box items

Contingency table: Check to display the contingency table.

Row profiles: Check to display a table of row profiles and row masses.

Columns profiles: Check to display a table of column profiles and column masses.

Expected frequencies: Check to display a table of the expected frequency in each cell of the contingency table.

Observed - expected frequencies: Check to display a table of the observed minus the expected frequency in each cell of the contingency table.

Chi-square values: Check to display a table of the χ^2 value in each cell of the contingency table.

Inertias: Check to display the table of the relative inertia in each cell of the contingency table.

Simple Correspondence Analysis – Graphs

Stat > Multivariate > Simple Correspondence Analysis > Graphs

Allows you display various plots to complement your analysis. See Simple correspondence analysis graphs.

In all plots, row points are plotted with red circles--solid circles for regular points, and open circles for supplementary points. Column points are plotted with blue squares--solid squares for regular points, and open squares for supplementary points.

The aspect ratio of the plots is one-to-one so that a unit on the x-axis is equal to a unit on the y-axis.

Dialog box items

Axis pairs for all plots (Y then X): Enter between 1 and 15 axis pairs for each requested plot. The axes you list must be axes in the subspace you defined in the main dialog box. For example, if you entered 4 in number of components, you can only list axes 1, 2, 3, and 4.

The first axis in a pair will be the Y or vertical axis of the plot; the second axis will be the X or horizontal axis of the plot. For example, if you enter 2 1 3 1 plots component 2 versus component 1, and component 3 versus component 1.

Show supplementary points in all plots: Check to display supplementary points on all plots.

Plots:

- **Symmetric plot showing rows only:** Check to display a plot that shows the row principal coordinates.
- **Symmetric plot showing columns only:** Check to display a plot that shows the column principal coordinates.
- **Symmetric plot showing rows and columns:** Check to display a symmetric plot that shows both row principal coordinates and column principal coordinates overlaid in a joint display.
- **Asymmetric row plot showing rows and columns:** Check to display an asymmetric row plot.
- **Asymmetric column plot showing rows and columns:** Check to display an asymmetric column plot.

To display simple correspondence analysis plots

1. Perform steps 1–2 of To perform a simple correspondence analysis.
2. Click **Graphs** and check all of the plots that you would like to display.
3. If you like, you can specify the component pairs and their axes for plotting. Enter between 1 and 15 component pairs in **Axis pairs for all plots (Y then X)**. Minitab plots the first component in each pair on the vertical or y-axis of the

plot; the second component in the pair on the horizontal or x-axis of the plot.

4 If you have supplementary data and would like to include this data in the plot(s), check **Show supplementary points in all plots**. Click **OK** in each dialog box.

In all plots, row points are plotted with red circles–solid circles for regular points, and open circles for supplementary points. Column points are plotted with blue squares–blue squares for regular points, and open squares for supplementary points.

Choosing a simple correspondence analysis graph

You can display the following simple correspondence-analysis plots:

- A row plot or a column plot
- A symmetric plot
- An asymmetric row plot or an asymmetric column plot

A row plot is a plot of row principal coordinates. A column plot is a plot of column principal coordinates.

A symmetric plot is a plot of row and column principal coordinates in a joint display. An advantage of this plot is that the profiles are spread out for better viewing of distances between them. The row-to-row and column-to-column distances are approximate χ^2 distances between the respective profiles. However, this same interpretation cannot be made for row-to-column distances. Because these distances are two different mappings, you must interpret these plots carefully [4].

An asymmetric row plot is a plot of row principal coordinates and of column standardized coordinates in the same plot. Distances between row points are approximate χ^2 distances between the row profiles. Choose the asymmetric row plot over the asymmetric column plot if rows are of primary interest.

An asymmetric column plot is a plot of column principal coordinates and row standardized coordinates. Distances between column points are approximate χ^2 distances between the column profiles. Choose an asymmetric column plot over an asymmetric row plot if columns are of primary interest.

An advantage of asymmetric plots is that there can be an intuitive interpretation of the distances between row points and column points, especially if the two displayed components represent a large proportion of the total inertia [4]. Suppose you have an asymmetric row plot, as shown in Example of simple correspondence analysis. This graph plots both the row profiles and the column vertices for components 1 and 2. The closer a row profile is to a column vertex, the higher the row profile is with respect to the column category. In this example, of the row points, Biochemistry is closest to column category E, implying that biochemistry as a discipline has the highest percentage of unfunded researchers in this study. A disadvantage of asymmetric plots is that the profiles of interest are often bunched in the middle of the graph [4], as happens with the asymmetric plot of this example.

Simple Correspondence Analysis – Storage

Stat > Multivariate > Simple Correspondence Analysis > Storage

Allows you to store results. In the four cases that store coordinates, the coordinate for the first component is stored in the first column, the coordinate for the second component in the second column, and so on. If there are supplementary points, their coordinates are stored at the ends of the columns.

Dialog box items

Columns of the contingency table: Enter one worksheet column for each column of the contingency table. Minitab does not store supplementary rows and columns.

Row principal coordinates: Check to store the row principal coordinates. Minitab stores the coordinate for the first component in a column named RPC1, the coordinate for the second component in a column that named RPC2, etc. If there are supplementary points, their coordinates are stored at the ends of the columns.

Row standardized coordinates: Check to store the row standardized coordinates. Minitab stores the coordinate for the first component in a column named RSC1, the coordinate for the second component in a column that named RSC2, etc. If there are supplementary points, their coordinates are stored at the ends of the columns.

Column principal coordinates: Check to store the column principal coordinates. Minitab stores the coordinate for the first component in a column named CPC1, the coordinate for the second component in a column that named CPC2, etc. If there are supplementary points, their coordinates are stored at the ends of the columns.

Column standardized coordinates: Check to store the column standardized coordinates. Minitab stores the coordinate for the first component in a column named CSC1, the coordinate for the second component in a column that named CSC2, etc. If there are supplementary points, their coordinates are stored at the ends of the columns.

Example of Simple Correspondence Analysis

The following example is from Correspondence Analysis in Practice, by M. J. Greenacre, p.75. Seven hundred ninety-six researchers were cross-classified into ten academic disciplines and five funding categories, where A is the highest funding category, D is the lowest, and category E is unfunded. Here, disciplines are rows and funding categories are columns. You wish to see how the disciplines compare to each other relative to the funding categories so you perform correspondence analysis from a row orientation. Supplementary data include: a row for museum researchers not included in the study and a row for mathematical sciences, which is the sum of Mathematics and Statistics.

1 Open the worksheet EXH_TABL.MTW.

2 Choose **Stat > Multivariate > Simple Correspondence Analysis**.

3 Choose **Columns of a contingency table,** and enter *CT1-CT5*. In **Row names**, enter *RowNames*. In **Column names**, enter *ColNames*.

4 Click **Results** and check **Row profiles**. Click **OK**.

5 Click **Supp Data**. In **Supplementary Rows**, enter *RowSupp1 RowSupp2*. In **Row names**, enter *RSNames*. Click **OK**.

6 Click **Graphs**. Check **Show supplementary points in all plots**. Check **Symmetric plot showing rows only** and **Asymmetric row plot showing rows and columns**.

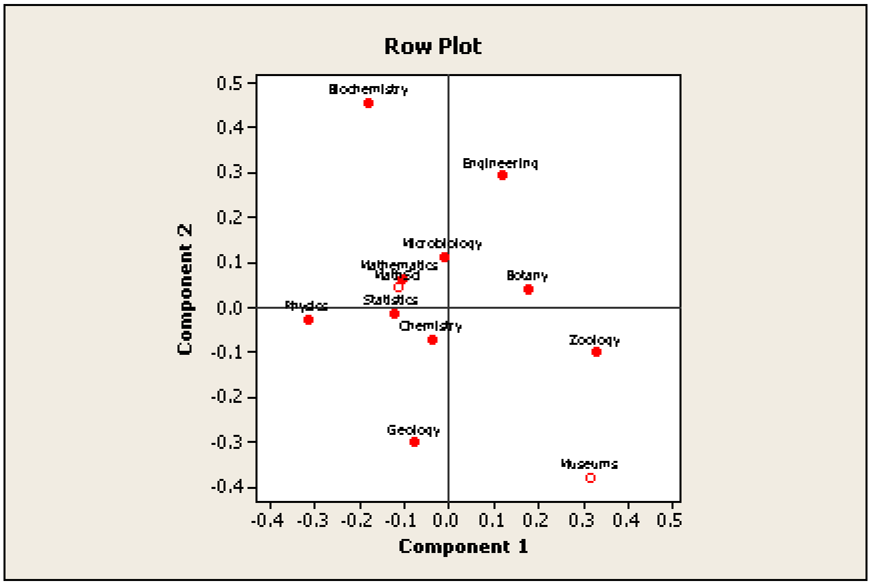
7 Click **OK** in each dialog box.

Session window output

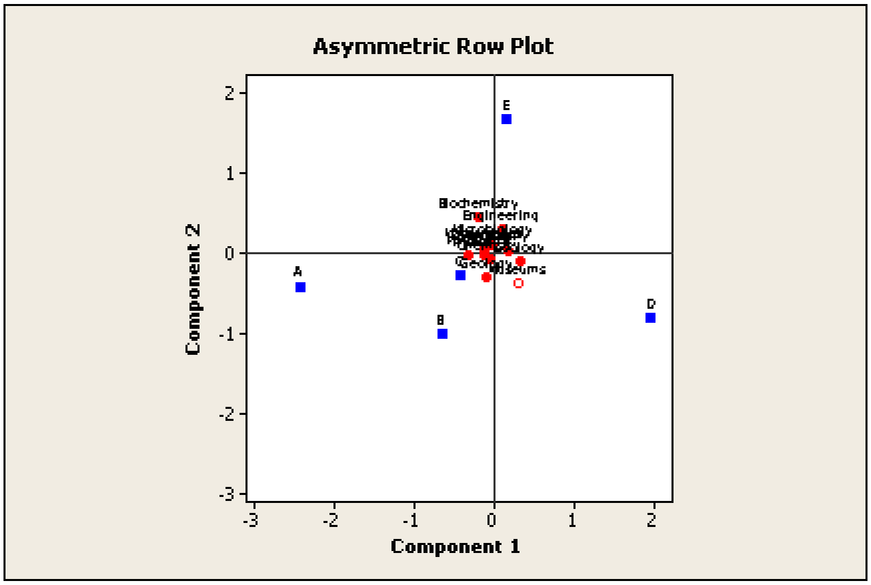
Simple Correspondence Analysis: CT1, CT2, CT3, CT4, CT5

```
Row Profiles

                  A      B      C      D      E   Mass
Geology       0.035  0.224  0.459  0.165  0.118  0.107
Biochemistry  0.034  0.069  0.448  0.034  0.414  0.036
Chemistry     0.046  0.192  0.377  0.162  0.223  0.163
Zoology       0.025  0.125  0.342  0.292  0.217  0.151
Physics       0.088  0.193  0.412  0.079  0.228  0.143
Engineering   0.034  0.125  0.284  0.170  0.386  0.111
Microbiology  0.027  0.162  0.378  0.135  0.297  0.046
Botany        0.000  0.140  0.395  0.198  0.267  0.108
Statistics    0.069  0.172  0.379  0.138  0.241  0.036
Mathematics   0.026  0.141  0.474  0.103  0.256  0.098
Mass          0.039  0.161  0.389  0.162  0.249


Analysis of Contingency Table

 Axis  Inertia  Proportion  Cumulative  Histogram
    1   0.0391      0.4720      0.4720  ******************************
    2   0.0304      0.3666      0.8385  ***********************
    3   0.0109      0.1311      0.9697  ********
    4   0.0025      0.0303      1.0000  *
Total   0.0829


Row Contributions

                                           Component  1
ID  Name           Qual   Mass  Inert   Coord   Corr  Contr
 1  Geology       0.916  0.107  0.137  -0.076  0.055  0.016
 2  Biochemistry  0.881  0.036  0.119  -0.180  0.119  0.030
 3  Chemistry     0.644  0.163  0.021  -0.038  0.134  0.006
 4  Zoology       0.929  0.151  0.230   0.327  0.846  0.413
 5  Physics       0.886  0.143  0.196  -0.316  0.880  0.365
 6  Engineering   0.870  0.111  0.152   0.117  0.121  0.039
 7  Microbiology  0.680  0.046  0.010  -0.013  0.009  0.000
 8  Botany        0.654  0.108  0.067   0.179  0.625  0.088
 9  Statistics    0.561  0.036  0.012  -0.125  0.554  0.014
10  Mathematics   0.319  0.098  0.056  -0.107  0.240  0.029

                      Component  2
ID  Name           Coord   Corr  Contr
 1  Geology       -0.303  0.861  0.322
 2  Biochemistry   0.455  0.762  0.248
 3  Chemistry     -0.073  0.510  0.029

 4  Zoology       -0.102  0.083  0.052
 5  Physics       -0.027  0.006  0.003
 6  Engineering    0.292  0.749  0.310
 7  Microbiology   0.110  0.671  0.018
 8  Botany         0.039  0.029  0.005
 9  Statistics    -0.014  0.007  0.000
10  Mathematics    0.061  0.079  0.012


Supplementary Rows

                                      Component  1          Component  2
ID  Name      Qual   Mass  Inert   Coord   Corr  Contr   Coord   Corr  Contr
 1  Museums  0.556  0.067  0.353   0.314  0.225  0.168  -0.381  0.331  0.318
 2  MathSci  0.559  0.134  0.041  -0.112  0.493  0.043   0.041  0.066  0.007


Column Contributions

                                 Component  1          Component  2
ID  Name   Qual   Mass  Inert   Coord   Corr  Contr   Coord   Corr  Contr
 1  A     0.587  0.039  0.187  -0.478  0.574  0.228  -0.072  0.013  0.007
 2  B     0.816  0.161  0.110  -0.127  0.286  0.067  -0.173  0.531  0.159
 3  C     0.465  0.389  0.094  -0.083  0.341  0.068  -0.050  0.124  0.032
 4  D     0.968  0.162  0.347   0.390  0.859  0.632  -0.139  0.109  0.103
 5  E     0.990  0.249  0.262   0.032  0.012  0.006   0.292  0.978  0.699
```

Graph window output



Graph window output



Interpreting the results

Row Profiles. The first table gives the proportions of each row category by column. Thus, of the class Geology, 3.5% are in column A, 22.4% are column B, etc. The mass of the Geology row, 0.107, is the proportion of all Geology subjects in the data set.

Analysis of Contingency Table. The second table shows the decomposition of the total inertia. For this example, the table gives a summary of the decomposition of the 10 x 5 contingency table into 4 components. The column labeled Inertia contains the χ squared / n value accounted for by each component. Of the total inertia, 65.972 / 796 or 0.0829, 47.2% is accounted for by the first component, 36.66% by the second component, and so on. Here, 65.972 is the χ squared statistic you would obtain if you performed a χ squared test of association with this contingency table.

Row Contributions. You can use the third table to interpret the different components. Since the number of components was not specified, Minitab calculates 2 components.

- The column labeled Qual, or quality, is the proportion of the row inertia represented by the two components. The rows Zoology and Geology, with quality = 0.928 and 0.916, respectively, are best represented among the rows by the two component breakdown, while Math has the poorest representation, with a quality value of 0.319.
- The column labeled Mass has the same meaning as in the Row Profiles table– the proportion of the class in the whole data set.
- The column labeled Inert is the proportion of the total inertia contributed by each row. Thus, Geology contributes 13.7% to the total χ squared statistic.

Next, Minitab displays information for each of the two components (axes).

- The column labeled Coord gives the principal coordinates of the rows.
- The column labeled Corr represents the contribution of the component to the inertia of the row. Thus, Component 1 accounts for most of the inertia of Zoology and Physics (Coor = 0.846 and 0.880, respectively), but explains little of the inertia of Microbiology (Coor = 0.009).
- Contr, the contribution of each row to the axis inertia, shows that Zoology and Physics contribute the most, with Botany contributing to a smaller degree, to Component 1. Geology, Biochemistry, and Engineering contribute the most to Component 2.

Supplementary rows. You can interpret this table in a similar fashion as the row contributions table.

Column Contributions. The fifth table shows that two components explain most of the variability in funding categories B, D, and E. The funded categories A, B, C, and D contribute most to component 1, while the unfunded category, E, contributes most to component 2.

Row Plot. This plot displays the row principal coordinates. Component 1, which best explains Zoology and Physics, shows these two classes well removed from the origin, but with opposite sign. Component 1 might be thought of as contrasting the biological sciences Zoology and Botany with Physics. Component 2 might be thought of as contrasting Biochemistry and Engineering with Geology.

Asymmetric Row Plot. Here, the rows are scaled in principal coordinates and the columns are scaled in standard coordinates. Among funding classes, Component 1 contrasts levels of funding, while Component 2 contrasts being funded

(A to D) with not being funded (E). Among the disciplines, Physics tends to have the highest funding level and Zoology has the lowest. Biochemistry tends to be in the middle of the funding level, but highest among unfunded researchers. Museums tend to be funded, but at a lower level than academic researchers

Multiple Correspondence Analysis

Multiple Correspondence Analysis

Stat > Multivariate > Multiple Correspondence Analysis

Multiple correspondence analysis extends simple correspondence analysis to the case of three or more categorical variables. Multiple correspondence analysis performs a simple correspondence analysis on a matrix of indicator variables where each column of the matrix corresponds to a level of categorical variable. Rather than having the two-way table of simple correspondence analysis, here the multi-way table is collapsed into one dimension. By moving from the simple to multiple procedure, you gain information on a potentially larger number of variables, but you may lose information on how rows and columns relate to each other.

Dialog box items

Input Data

Categorical variables: Choose If your data are in raw form and then enter the columns containing the categorical variables.

Indicator variables: Choose if your data are arranged as indicator variables and then enter the columns containing the indicator in the text box. The entries in all columns must be either the integers 0 and 1.

Category names: Enter the column that contains the category names if you want to assign category names. The name column must be a text column whose length matches the number of categories on all categorical variables.

For example, suppose there are 3 categorical variables: Sex (male, female), Hair color (blond, brown, black), and Age (under 20, from 20 to 50, over 50), and no supplementary variables. You would assign 2 + 3 + 3 = 8 category names, so the name column would contain 8 rows.

Minitab only uses the first 8 characters of the names in printed tables, but uses all characters on graphs.

Number of components: Enter the number of components to calculate. The default number of components is 2.

<Supp Data>

<Results>

<Graphs>

<Storage>

Data – Multiple Correspondence Analysis

Worksheet data may be arranged in two ways: raw or indicator variable form. See Arrangement of Input Data. Worksheet data arrangement determines acceptable data values.

- If your data are in **raw form**, you can have one or more classification columns with each row representing one observation. The data represent categories and may be numeric, text, or date/time. If you wish to change the order in which text categories are processed from their default alphabetized order, you can define your own order. See Ordering Text Categories. You must delete missing data before using this procedure.
- If your data are in **indicator variable form**, each row will also represent one observation. There will be one indicator column for each category level. You can use **Calc > Make Indicator Variables** to create indicator variables from raw data. You must delete missing data before using this procedure.

Supplementary data

When performing a multiple correspondence analysis, you have a main classification set of data on which you perform your analysis. However, you may also have additional or **supplementary data** in the same form as the main set, and you might want to see how this supplementary data are "scored" using the results from the main set. These supplementary data are typically a classification of your variables that can help you to interpret the results. Minitab does not include these data when calculating the components, but you can obtain a profile and display supplementary data in graphs.

Set up your supplementary data in your worksheet using the same form, either raw data or indicator variables, as you did for the input data. Because your supplementary data will provide additional information about your observations, your supplementary data column(s) must be the same length as your input data.

To perform a multiple correspondence analysis

1 Choose **Stat > Multivariate > Multiple Correspondence Analysis**.

2 To enter your data, do one of the following:

- For raw data, enter the columns containing the raw data in **Categorical variables**.
- For indicator variable data, enter the columns containing the indicator variable data in **Indicator variables**.

3 If you like, use any dialog box options, then click **OK**.

Multiple Correspondence Analysis – Supplementary Data

Stat > Multivariate > Multiple Correspondence Analysis > Supp Data

Dialog box items

Supplementary data (in same form as input data): Enter one or more columns containing supplementary column data. See What are Supplementary Data?

Category names: Enter a column containing a text name for each category of all the supplementary data, arranged by numerical order of the corresponding categories by variable.

What are Supplementary Data?

When performing a simple correspondence analysis, you have a main classification set of data on which you perform your analysis. However, you may also have additional or supplementary data in the same form as the main set, because you can see how these supplementary data are "scored" using the results from the main set. These supplementary data may be further information from the same study, information from other studies, or target profiles [4]. Minitab does not include these data when calculating the components, but you can obtain a profile and display supplementary data in graphs.

You can have row supplementary data or column supplementary data. Row supplementary data constitutes an additional row(s) of the contingency table, while column supplementary data constitutes an additional column(s) of the contingency table. Supplementary data must be entered in contingency table form. Therefore, each worksheet column of these data must contain c entries (where c is the number of contingency table columns) or r entries (where r is the number of contingency table rows).

Multiple Correspondence Analysis – Results

Stat > Multivariate > Multiple Correspondence Analysis > Results

Allows you to control displayed output.

Dialog box items

Indicator table: Check to display a table of indicator variables.

Burt table: Check to display the Burt table.

Multiple Correspondence Analysis – Graphs

Stat > Multivariate > Multiple Correspondence Analysis > Graphs

Allows you display column plots.

Points are plotted with blue squares--solid squares for regular points, and open squares for supplementary points.

The aspect ratio of the plots is one-to-one so that a unit on the x-axis is equal to a unit on the y-axis.

Dialog box items

Axis pairs for plots (Y then X): Enter one to 15 axis pairs to use for the column plots. The axes you list must be axes in the subspace you defined in the main dialog box. For example, if you entered 4 in number of components, you can only list axes 1, 2, 3, and 4.

The first axis in a pair will be the Y or vertical axis of the plot; the second axis will be the X or horizontal axis of the plot. For example, if you enter 2 1 3 1 plots component 2 versus component 1, and component 3 versus component 1.

Show supplementary points in all plots: Check to display supplementary points on all plots.

Display column plot: Check to display a plot that shows the column coordinates.

Multiple Correspondence Analysis – Storage

Stat > Multivariate > Multiple Correspondence Analysis > Storage

Allows you to store the column coordinates.

Dialog box item

Coordinates for the components: Check to store the column coordinates. Minitab stores the coordinate for the first component in the first listed column, the coordinate for the second component in the second listed column, etc. If there are supplementary points, their coordinates are stored at the ends of the columns.

Example of Multiple Correspondence Analysis

Automobile accidents are classified [8] (data from [3]) according to the type of accident (collision or rollover), severity of accident (not severe or severe), whether or not the driver was ejected, and the size of the car (small or standard). Multiple correspondence analysis was used to examine how the categories in this four-way table are related to each other.

1 Open the worksheet EXH_TABL.MTW.

2 Choose **Stat > Multivariate > Multiple Correspondence Analysis**.

3 Choose **Categorical variables**, and enter *CarWt DrEject AccType AccSever*.

4 In **Category names**, enter *AccNames*.

5 Click **Graphs**. Check **Display column plot**.

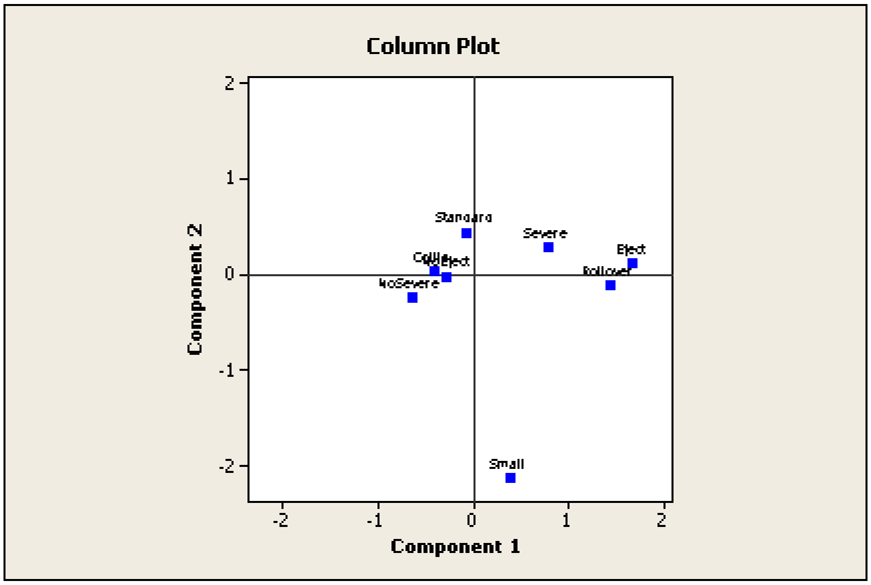
6 Click **OK** in each dialog box.

Session window output

Multiple Correspondence Analysis: CarWt, DrEject, AccType, AccSever

```
Analysis of Indicator Matrix

 Axis  Inertia  Proportion  Cumulative  Histogram
    1   0.4032      0.4032      0.4032  ******************************
    2   0.2520      0.2520      0.6552  ******************
    3   0.1899      0.1899      0.8451  **************
    4   0.1549      0.1549      1.0000  ***********
Total   1.0000


Column Contributions

                                       Component  1          Component  2
ID  Name       Qual   Mass  Inert  Coord   Corr  Contr  Coord   Corr  Contr
 1  Small     0.965  0.042  0.208   0.381  0.030  0.015  -2.139  0.936  0.771
 2  Standard  0.965  0.208  0.042  -0.078  0.030  0.003   0.437  0.936  0.158
 3  NoEject   0.474  0.213  0.037  -0.284  0.472  0.043  -0.020  0.002  0.000
 4  Eject     0.474  0.037  0.213   1.659  0.472  0.250   0.115  0.002  0.002
 5  Collis    0.613  0.193  0.057  -0.426  0.610  0.087   0.034  0.004  0.001
 6  Rollover  0.613  0.057  0.193   1.429  0.610  0.291  -0.113  0.004  0.003
 7  NoSevere  0.568  0.135  0.115  -0.652  0.502  0.143  -0.237  0.066  0.030
 8  Severe    0.568  0.115  0.135   0.769  0.502  0.168   0.280  0.066  0.036
```

Graph window output



Interpreting the results

Analysis of Indicator Matrix. This table gives a summary of the decomposition of variables. The column labeled Inertia is the χ squared / n value accounted for by each component. Of the total inertia of 1, 40.3%, 25.2%, 19.0%, and, 15.5% are accounted for by the first through fourth components, respectively.

Column Contributions. Use the column contributions to interpret the different components. Since we did not specify the number of components, Minitab calculates 2 components.

- The column labeled Qual, or quality, is the proportion of the column inertia represented by the all calculated components. The car-size categories (Small, Standard) are best represented by the two component breakdown with Qual = 0.965, while the ejection categories are the least represented with Qual = 0.474. When there are only two categories for each class, each is represented equally well by any component, but this rule would not necessarily be true for more than two categories.
- The column labeled Mass is the proportion of the class in the whole data set. In this example, the CarWt, DrEject, AccType, and AccSever classes combine for a proportion of 0.25.
- The column labeled Inert is the proportion of inertia contributed by each column. The categories small cars, ejections, and collisions have the highest inertia, summing 61.4%, which indicates that these categories are more dissociated from the others.

Next, Minitab displays information for each of the two components (axes).

- The column labeled Coord gives the column coordinates. Eject and Rollover have the largest absolute coordinates for component 1 and Small has the largest absolute coordinate for component 2. The sign and relative size of the coordinates are useful in interpreting components.
- The column labeled Corr represents the contribution of the respective component to the inertia of the row. Here, Component 1 accounts for 47 to 61% of the inertia of the ejection, collision type, and accident severity categories, but explains only 3.0% of the inertia of car size.
- Contr, the contribution of the row to the axis inertia, shows Eject and Rollover contributing the most to Component 1 (Contr = 0.250 and 0.291, respectively). Component 2, on the other hand accounts for 93.6% of the inertia of the car size categories, with Small contributing 77.1% of the axis inertia.

Column Plot. As the contribution values for Component 1 indicate, Eject and Rollover are most distant from the origin. This component contrasts Eject and Rollover and to some extent Severe with NoSevere. Component 2 separates Small with the other categories. Two components may not adequately explain the variability of these data, however.

Nonparametrics

Overview

Nonparametric Analysis Overview

Minitab provides the following types of nonparametric procedures:

- 1-sample median test (sign test and Wilcoxon test)
- 2-sample median test (Mann-Whitney test)
- Analysis of variance (Kruskal-Wallis, Mood's median, and Friedman test)
- Test for randomness (runs test)
- Pairwise statistics (pairwise averages, pairwise differences, and pairwise slopes)

Parametric implies that a distribution is assumed for the population. Often, an assumption is made when performing a hypothesis test that the data are a sample from a certain distribution, commonly the normal distribution. **Nonparametric** implies that there is no assumption of a specific distribution for the population.

An advantage of a parametric test is that if the assumptions hold, the power, or the probability of rejecting H_0 when it is false, is higher than is the power of a corresponding nonparametric test with equal sample sizes. The nonparametric test results are more robust against violation of the assumptions. Therefore, if assumptions are violated for a test based upon a parametric model, the conclusions based on parametric test p-values may be more misleading than conclusions based upon nonparametric test p-values. See [1] for comparing the power of some of these nonparametric tests to their parametric equivalent.

Tests of population location

These nonparametric tests are analogous to the parametric t-tests and analysis of variance procedures in that they are used to perform tests about population location or center value. The center value is the mean for parametric tests and the median for nonparametric tests.

- 1-Sample Sign:– performs a 1-sample sign test of the median and calculates the corresponding point estimate and confidence interval. Use this test as a nonparametric alternative to 1-sample Z and 1-sample t-tests.
- 1-Sample Wilcoxon:– performs a 1-sample Wilcoxon signed rank test of the median and calculates the corresponding point estimate and confidence interval. Use this test as a nonparametric alternative to 1-sample Z and 1-sample t-tests.
- Mann-Whitney:– performs a hypothesis test of the equality of two population medians and calculates the corresponding point estimate and confidence interval. Use this test as a nonparametric alternative to the 2-sample t-test.
- Kruskal-Wallis:– performs a hypothesis test of the equality of population medians for a one-way design (two or more populations). This test is a generalization of the procedure used by the Mann-Whitney test and, like Mood's median test, offers a nonparametric alternative to the one-way analysis of variance. The Kruskal-Wallis test looks for differences among the populations' medians.

  The Kruskal-Wallis test is more powerful (the confidence interval is narrower, on average) than Mood's median test for analyzing data from many distributions, including data from the normal distribution, but is less robust against outliers.
- Mood's Median Test:– performs a hypothesis test of the equality of population medians in a one-way design. Mood's median test, like the Kruskal-Wallis test, provides a nonparametric alternative to the usual one-way analysis of variance. Mood's median test is sometimes called a median test or sign scores test.

  Mood's median test is robust against outliers and errors in data, and is particularly appropriate in the preliminary stages of analysis. Mood's median test is more robust against outliers than the Kruskal-Wallis test, but is less powerful (the confidence interval is wider, on the average) for analyzing data from many distributions, including data from the normal distribution.
- Friedman:– performs a nonparametric analysis of a randomized block experiment and thus provides an alternative to the two-way analysis of variance.

  Randomized block experiments are a generalization of paired experiments. The Friedman test is a generalization of the paired sign test with a null hypothesis of treatments having no effect. This test requires exactly one observation per treatment-block combination.

Tests for randomness

Runs Tests test whether or not the data order is random. No assumptions are made about population distribution parameters. Use **Stat > Quality Tools > Run Chart** to generate a run chart and perform additional tests for randomness.

Procedures for calculating pairwise statistics

Pairwise Averages, Pairwise Differences, and Pairwise Slopes compute averages, differences, and slopes, respectively, for all possible pairs of values. These statistics are sometimes used in nonparametric statistical calculations.

Nonparametrics

Stat > Nonparametrics

Select one of the following commands:

1-Sample Sign

1-Sample Wilcoxon

Mann-Whitney

Kruskal-Wallis

Mood's Median Test

Friedman test

Runs Test

Pairwise Averages

Pairwise Differences

Pairwise Slopes

Examples of Nonparametric Tests

The following examples illustrate how to use the various nonparametric techniques available. Choose an example below:

1-Sample Sign Test

1-Sample Sign Confidence Interval

1-Sample Wilcoxon Test

1-Sample Wilcoxon Confidence Interval

Mann-Whitney Test

Kruskal-Wallis Test

Mood's Median Test

Friedman Test

Runs Test

Pairwise Averages

Pairwise Differences

Pairwise Slopes

References – Nonparametrics

[1] Gibbons, J.D. (1976). Nonparametric Methods for Quantitative Analysis. Holt, Rhinehart, and Winston.

[2] T.P. Hettmansperger and S.J. Sheather (1986). "Confidence Intervals Based on Interpolated Order Statistics," Statistics and Probability Letters, 4, pp.75–79.

[3] M. Hollander and D.A. Wolfe (1973). Nonparametric Statistical Methods, John Wiley & Sons.

[4] D.B. Johnson and T. Mizoguchi (1978). "Selecting the Kth Element in X + Y and X1 + X2 + ... + Xm," SIAM Journal of Computing 7, pp.147–153.

[5] E.L. Lehmann (1975). Nonparametrics: Statistical Methods Based on Ranks, Holden–Day.

[6] J.W. McKean and T.A. Ryan, Jr. (1977). "An Algorithm for Obtaining Confidence Intervals and Point Estimates Based on Ranks in the Two Sample Location Problem," Transactions on Mathematical Software, pp.183–185.

[7] G. Noether (1971). Statistics–A Non–Parametric Approach, Houghton-Mifflin.

1-Sample Sign

1-Sample Sign

Stat > Nonparametrics > 1-Sample Sign

You can perform a 1-sample sign test of the median or calculate the corresponding point estimate and confidence interval. For the one-sample sign test, the hypotheses are

H0: median = hypothesized median versus H1: median ≠ hypothesized median

Use the sign test as a nonparametric alternative to 1-sample Z-tests and to 1-sample t-tests , which use the mean rather than the median.

Dialog box items

Variables: Select the column(s) containing the variable(s) you want to test.

Confidence interval: Choose to calculate a sign confidence interval.

Level: Enter a confidence level between 0 and 100 (default is 95.0).

Note Minitab calculates confidence interval for the level closest to the requested level.

Test median: Choose to perform a sign-test, then specify the null hypothesis test value.

Alternative: Click the arrow to choose the kind of test performed by selecting less than (lower-tailed), not equal (two-tailed), or greater than (upper-tailed) from the drop-down box.

Data – 1-Sample Sign

You need at least one column of numeric data. If you enter more than one column of data, Minitab performs a one-sample sign test separately for each column. Minitab automatically omits missing data from the calculations.

To calculate a sign confidence interval and test for the median

1 Choose **Stat > Nonparametrics > 1-Sample Sign**.

2 In **Variables**, enter the column(s) containing the data.

3 Choose one of the following:

- to calculate a sign confidence interval for the median, choose **Confidence interval**
- to perform a sign test, choose **Test median**

4 If you like, use one or more of the available dialog box options, then click **OK**.

Example of 1-Sample Sign Confidence Interval

Using data for the 29 houses in the previous example, you also want to obtain a 95% confidence interval for the population median.

1 Open the worksheet EXH_STAT.MTW.

2 Choose **Stat > Nonparametrics > 1-Sample Sign**.

3 In **Variables**, enter **PriceIndex**. Choose **Confidence interval**. Click **OK**.

Session window output

Sign CI: PriceIndex

```
Sign confidence interval for median

                                       Confidence
                             Achieved   Interval
             N  Median  Confidence  Lower  Upper  Position
PriceIndex  29   144.0      0.9386  110.0  210.0        10
                            0.9500  108.5  211.7       NLI
                            0.9759  101.0  220.0         9
```

Interpreting the results

Minitab calculates three intervals. The first and third intervals have confidence levels below and above the requested level, respectively. The confidence levels are calculated according to binomial probabilities. For example, the interval that goes from the 9th smallest observation to the 9th largest observation has a confidence of $1 - 2P(X < 9) = 0.9759$, where X has a binomial distribution with n = 29 and p = 0.5. The middle confidence interval of (110.0, 211.7) is found by a nonlinear interpolation procedure [2], and has a confidence level equal to the requested level or the default of 95%.

Example of 1-Sample Sign Test of the Median

Price index values for 29 homes in a suburban area in the Northeast were determined. Real estate records indicate the population median for similar homes the previous year was 115. This test will determine if there is sufficient evidence for judging if the median price index for the homes was greater than 115 using $\alpha = 0.10$.

1 Open the worksheet EXH_STAT.MTW.

2 Choose **Stat > Nonparametrics > 1-Sample Sign**.

3 In **Variables**, enter **PriceIndex**.

4 Choose **Test median** and enter **115** in the text box.

5 In **Alternative**, choose **greater than**. Click **OK**.

Session window output

Sign Test for Median: PriceIndex

```
Sign test of median =  115.0 versus > 115.0

             N  Below  Equal  Above        P  Median
PriceIndex  29     12      0     17  0.2291   144.0
```

Interpreting the results

Of the 29 price index data, 12 are below and 17 are above the hypothesize value, 115. Because an upper one-sided test was chosen, the p-value is the binomial probability of observing 17 or more observations greater than 115 if p is 0.5. If your α level was less than a p-value of 0.2291, you would fail to conclude that the population median was greater than 115, which seems likely for most situations.

If you had performed a two-sided test using the same sample (H_0: median = 115 versus H_1 median ≠ 115), there would be 12 observations below 115 and 17 above. Since you would be performing a two-sided test, you would look at the number of observations below and above 115, and take the larger of these, 17. The binomial probability of observing this many observations or more is 0.2291, and the p-value of the two-sided test is twice this value, or 2 (0.2291) = 0.4582. If n had been > 50, Minitab would have used a normal approximation to the binomial in calculating the p-value.

1-Sample Wilcoxon

1-Sample Wilcoxon

Stat > Nonparametrics > 1-Sample Wilcoxon

You can perform a 1-sample Wilcoxon signed rank test of the median or calculate the corresponding point estimate and confidence interval. The Wilcoxon signed rank test hypotheses are

H_0: median = hypothesized median versus H_1: median ≠ hypothesized median

An assumption for the one-sample Wilcoxon test and confidence interval is that the data are a random sample from a continuous, symmetric population. When the population is normally distributed, this test is slightly less powerful (the confidence interval is wider, on the average) than the t-test. It may be considerably more powerful (the confidence interval is narrower, on the average) for other populations.

Dialog box items

Variables: Select the column(s) containing the variable(s) you want to test.

Confidence interval: Choose to calculate a one-sample Wilcoxon confidence interval for each column listed.

Level: Enter a confidence level between 0 and 100 (default is 95.0.)

Note Minitab calculates confidence interval for the level closest to the requested level.

Test median: Choose to perform a Wilcoxon signed-rank test, then specify the null hypothesis test value.

Alternative: Click the arrow to choose the kind of test performed by selecting less than (lower-tailed), not equal (two-tailed), or greater than (upper-tailed) from the drop-down box.

Data – 1-Sample Wilcoxon

You need at least one column of numeric data. If you enter more than one column of data, Minitab performs a one-sample Wilcoxon test separately for each column. Minitab automatically omits missing data from the calculations.

To calculate 1-sample Wilcoxon confidence interval and test for the median

1 Choose **Stat > Nonparametrics > 1-Sample Wilcoxon**.

2 In **Variables**, enter the column(s) containing the variable(s).

3 Choose one of the following:

- to calculate a Wilcoxon confidence interval for the median, choose **Confidence interva**l
- to perform a Wilcoxon signed rank test, choose **Test median**

4 If you like, use one or more of the available dialog box options, then click **OK**

Note If you do not specify a hypothesized median, a one-sample Wilcoxon test tests whether the sample median is different from zero.

Example of 1-Sample Wilcoxon Confidence Interval

A 95% confidence interval for the population median can be calculated by the one-sample Wilcoxon method.

1 Open the worksheet EXH_STAT.MTW.

2 Choose **Stat > Nonparametrics > 1-Sample Wilcoxon**.

3 In **Variables**, enter **Achievement**.

4 Choose **Confidence interval**. Click **OK**.

Session window output

Wilcoxon Signed Rank CI: Achievement

```
                                         Confidence
              Estimated    Achieved       Interval
            N    Median  Confidence    Lower   Upper
Achievement  9      77.5        95.6    70.0    84.0
```

Interpreting the results

The computed confidence interval (70, 84) has a confidence level of 95.6%. You can also perform the above two-sided hypothesis test at $\alpha = 1–0.956$ or 0.044 by noting that 77 is within the confidence interval, and thus fail to reject H0. The estimated median is the median of the Walsh averages.

Example of a 1-Sample Wilcoxon Test

Achievement test scores in science were recorded for 9 students. This test enables you to judge if there is sufficient evidence for the population median being different than 77 using $\alpha = 0.05$.

1 Open the worksheet EXH_STAT.MTW.

2 Choose **Stat > Nonparametrics > 1-Sample Wilcoxon**.

3 In **Variables**, enter **Achievement**.

4 Choose **Test median**, and enter **77** in the box. Click **OK**.

Session window output

Wilcoxon Signed Rank Test: Achievement

```
Test of median = 77.00 versus median not = 77.00

                  N
                for   Wilcoxon         Estimated
            N  Test  Statistic      P     Median
Achievement  9     8       19.5  0.889      77.50
```

Interpreting the results

The Wilcoxon test statistic of 19.5 is the number of Walsh averages exceeding 77. Because one test score was equal to the hypothesized value, the sample size used for the test was reduced by one to 8, as indicated under "N for Test".

There is insufficient evidence to reject the null hypothesis ($p > 0.05$). The population median is not statistically different from 77. The estimated median, here 77.5, is the median of the Walsh averages. This median may be different from the median of the data, which is 77 in this example.

Mann-Whitney

Mann-Whitney

Stat > Nonparametrics > Mann-Whitney

You can perform a 2-sample rank test (also called the Mann-Whitney test, or the two-sample Wilcoxon rank sum test) of the equality of two population medians, and calculate the corresponding point estimate and confidence interval. The hypotheses are

H0: $\eta_1 = \eta_2$ versus H1: $\eta_1 \neq \eta_2$, where η is the population median.

An assumption for the Mann-Whitney test is that the data are independent random samples from two populations that have the same shape and whose variances are equal and a scale that is continuous or ordinal (possesses natural ordering) if discrete. The 2-sample rank test is slightly less powerful (the confidence interval is wider on the average) than the 2-sample test with pooled sample variance when the populations are normal, and considerably more powerful (confidence interval is narrower, on the average) for many other populations. If the populations have different shapes or different standard deviations, a 2-Sample t without pooling variances may be more appropriate.

Dialog box items

First Sample: Select the column containing the sample data from one population.

Second Sample: Select the column containing the sample data from the other population.

Confidence level: Specify the level of confidence desired between 0 and 100; the attained level will be as close as possible.

Note Minitab calculates confidence interval for the level closest to the requested level.

Alternative: Click the arrow to choose the kind of test performed by selecting less than (lower-tailed), not equal (two-tailed), or greater than (upper-tailed) from the drop-down box.

Data – Mann-Whitney

You will need two columns containing numeric data drawn from two populations. The columns do not need to be the same length. Minitab automatically omits missing data from the calculations.

To calculate a Mann-Whitney test

1 Choose **Stat > Nonparametrics > Mann-Whitney**.

2 In **First Sample**, enter the column containing the sample data from one population.

3 In **Second Sample**, enter the column containing the other sample data.

4 If you like, use one or more of the available dialog box options, then click **OK**.

Example of 2-sample Mann-Whitney test

Samples were drawn from two populations and diastolic blood pressure was measured. You will want to determine if there is evidence of a difference in the population locations without assuming a parametric model for the distributions. Therefore, you choose to test the equality of population medians using the Mann-Whitney test with α = 0.05 rather than using a two-sample t-test, which tests the equality of population means.

1 Open the worksheet EXH_STAT.MTW.

2 Choose **Stat > Nonparametrics > Mann-Whitney**.

3 In **First Sample**, enter **DBP1**. In **Second Sample**, enter **DBP2**. Click **OK**.

Session window output

Mann-Whitney Test and CI: DBP1, DBP2

```
      N  Median
DBP1  8   69.50
DBP2  9   78.00


Point estimate for ETA1-ETA2 is -7.50
95.1 Percent CI for ETA1-ETA2 is (-18.00,4.00)
W = 60.0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.2685
The test is significant at 0.2679 (adjusted for ties)
```

Interpreting the results

Minitab calculates the sample medians of the ordered data as 69.5 and 78. The 95.1% confidence interval for the difference in population medians (ETA1–ETA2) is [–18 to 4]. The test statistic W = 60 has a p-value of 0.2685 or 0.2679 when adjusted for ties. Since the p-value is not less than the chosen α level of 0.05, you conclude that there is insufficient evidence to reject H0. Therefore, the data does not support the hypothesis that there is a difference between the population medians.

Kruskal-Wallis

Kruskal-Wallis

Stat > Nonparametrics > Kruskal-Wallis

You can perform a Kruskal-Wallis test of the equality of medians for two or more populations.

This test is a generalization of the procedure used by the Mann-Whitney test and, like Mood's Median test, offers a nonparametric alternative to the one-way analysis of variance. The Kruskal-Wallis hypotheses are:

H0: the population medians are all equal versus H1: the medians are not all equal

An assumption for this test is that the samples from the different populations are independent random samples from continuous distributions, with the distributions having the same shape. The Kruskal-Wallis test is more powerful than Mood's median test for data from many distributions, including data from the normal distribution, but is less robust against outliers.

Dialog box items

Response: Enter the column that contains the response variable from all the samples.

Factor: Enter the column that contains the factor levels.

Data – Kruskal-Wallis

The response (measurement) data must be stacked in one numeric column. You must also have a column that contains the factor levels or population identifiers. Factor levels can be numeric, text, or date/time data. If you wish to change the order in which text levels are processed, you can define your own order. See Ordering Text Categories. Calc > Make Patterned Data can be helpful in entering the level values of a factor.

Minitab automatically omits rows with missing responses or factor levels from the calculations.

To perform a Kruskal-Wallis test

1 Choose **Stat > Nonparametrics > Kruskal-Wallis**.

2 In **Response**, enter the column containing the response data.

3 In **Factor**, enter the column containing the factor levels. Click **OK**.

Example of Kruskal-Wallis Test

Measurements in growth were made on samples that were each given one of three treatments. Rather than assuming a data distribution and testing the equality of population means with one-way ANOVA, you decide to select the Kruskal-Wallis procedure to test H_0: $\eta1 = \eta2 = \eta3$, versus H_1: not all η's are equal, where the η's are the population medians.

1 Open the worksheet EXH_STAT.MTW.

2 Choose **Stat > Nonparametrics > Kruskal-Wallis**.

3 In **Response**, enter **Growth**.

4 In **Factor**, enter **Treatment**. Click **OK**.

Session window output

Kruskal-Wallis Test: Growth versus Treatment

```
Kruskal-Wallis Test on Growth

Treatment   N  Median  Ave Rank      Z
1           5   13.20       7.7  -0.45
2           5   12.90       4.3  -2.38
3           6   15.60      12.7   2.71
Overall    16               8.5

H = 8.63  DF = 2  P = 0.013
H = 8.64  DF = 2  P = 0.013  (adjusted for ties)
```

Interpreting the results

The sample medians for the three treatments were calculated 13.2, 12.9, and 15.6. The z-value for level 1 is –0.45, the smallest absolute z-value. This size indicates that the mean rank for treatment 1 differed least from the mean rank for all observations. The mean rank for treatment 2 was lower than the mean rank for all observations, as the z-value is negative (z = –2.38). The mean rank for treatment 3 is higher than the mean rank for all observations, as the z-value is positive (z = 2.71).

The test statistic (H) had a p-value of 0.013, both unadjusted and adjusted for ties, indicating that the null hypothesis can be rejected at a levels higher than 0.013 in favor of the alternative hypothesis of at least one difference among the treatment groups.

Mood's Median Test

Mood's Median Test

Stat > Nonparametrics > Mood's Median Test

Mood's median test can be used to test the equality of medians from two or more populations and, like the Kruskal-Wallis Test, provides an nonparametric alternative to the one-way analysis of variance. Mood's median test is sometimes called a median test or sign scores test. Mood's median test tests:

H_0: the population medians are all equal versus H_1: the medians are not all equal

An assumption of Mood's median test is that the data from each population are independent random samples and the population distributions have the same shape. Mood's median test is robust against outliers and errors in data and is particularly appropriate in the preliminary stages of analysis. Mood's median test is more robust than is the Kruskal-Wallis test against outliers, but is less powerful for data from many distributions, including the normal.

Dialog box items

Response: Enter the column that contains the response data from all the samples.

Factor: Enter the column that contains the factor levels.

Store residuals: Check to store the residuals.

Store fits: Check to store the fitted values. These are the group medians.

Data – Mood's Median Test

The response (measurement) data must be stacked in one numeric column. You must also have a column that contains the factor levels or population identifiers. Factor levels can be numeric, text, or date/time data. If you wish to change the order in which text levels are processed, you can define your own order. See Ordering Text Categories. Calc > Make Patterned Data can be helpful in entering the level values of a factor.

Minitab automatically omits rows with missing responses or factor levels from the calculations.

To perform a Mood's median test

1 Choose **Stat > Nonparametrics > Mood's Median Test**.

2 In **Response**, enter the column containing the measurement data.

3 In **Factor**, enter the column containing the factor levels.

4 If you like, use one or more of the available dialog box options, then click **OK**.

Example of Mood's median test

One hundred seventy-nine participants were given a lecture with cartoons to illustrate the subject matter. Subsequently, they were given the OTIS test, which measures general intellectual ability. Participants were rated by educational level: 0 = preprofessional, 1 = professional, 2 = college student. The Mood's median test was selected to test H_0: $\eta 1 = \eta 2 = \eta 3$, versus H_1: not all η's are equal, where the η's are the median population OTIS scores for the three education levels.

1 Open the worksheet CARTOON.MTW.

2 Choose **Stat > Nonparametrics > Mood's Median Test**.

3 In **Response**, enter **Otis**. In **Factor**, enter **ED**. Click **OK**.

Session window output

Mood Median Test: Otis versus ED

```
Mood median test for Otis
Chi-Square = 49.08    DF = 2    P = 0.000

                            Individual 95.0% CIs
ED  N<=  N>  Median  Q3-Q1  ----+---------+---------+---------+--
0    47   9    97.5   17.3  (-----*-----)
1    29  24   106.0   21.5           (------*------)
2    15  55   116.5   16.3                           (----*----)
                            ----+---------+---------+---------+--
                             96.0     104.0     112.0     120.0

Overall median = 107.0
```

Interpreting the results

The participant scores are classified as being below or above the overall median, and a chi-square test for association is performed. The χ^2 value of 49.08 with a p-value of < 0.0005 indicates that there is sufficient evidence to reject H_0 in favor of H_1 at commonly used α levels.

For each factor level, Minitab prints the median, interquartile range, and a sign confidence interval for the population median. The confidence interval is the nonlinear interpolation interval done by the one-sample sign procedure (see methods and formulas – 1-sample sign). Test scores are highest for college students. (You might conjecture that it is the college student whose intellect is most stimulated by cartoons.)

If a level has less than six observations, the confidence level would be less than 95%. When there are only two factor levels, Minitab displays a 95% two-sample confidence interval for the difference between the two population medians.

Friedman

Friedman

Stat > Nonparametrics > Friedman

Friedman test is a nonparametric analysis of a randomized block experiment, and thus provides an alternative to the Two-way analysis of variance. The hypotheses are:

H0: all treatment effects are zero versus H1: not all treatment effects are zero

Randomized block experiments are a generalization of paired experiments, and the Friedman test is a generalization of the paired sign test. Additivity (fit is sum of treatment and block effect) is not required for the test, but is required for the estimate of the treatment effects.

Dialog box items

Response: Enter the column containing the response variable.

Treatment: Enter the column that contains the treatments.

Blocks: Enter the column that contains the blocks.

Store residuals: Check to store the residuals. The residuals are calculated as the (observation adjusted for treatment effect) – (adjusted block median).

Store fits: Check to store the fitted values. The fits are calculated as the (treatment effect) + (adjusted block median).

Output

Minitab prints the test statistic, which has an approximately chi-square distribution, and the associated degrees of freedom (number of treatments minus one). If there are ties within one or more blocks, the average rank is used, and a test statistic corrected for ties is also printed. If there are many ties, the uncorrected test statistic is conservative; the corrected version is usually closer, but may be either conservative or liberal. Minitab displays an estimated median for each treatment level. The estimated median is the grand median plus the treatment effect. For details of the method used see [2].

Data – Friedman

The response (measurement) data must be stacked in one numeric column. You must also have a column that contains the treatment levels and a column that contains the block levels. Treatment and block levels can be numeric, text, or date/time data. If you wish to change the order in which text levels are processed, you can define your own order. See Ordering Text Categories. Calc > Make Patterned Data can be helpful in entering the level values of a factor.

You must have exactly one nonmissing observation per treatment–block combination. Minitab automatically omits rows with missing responses, treatment levels, or block levels from the calculations.

To perform a Friedman test

1 Choose **Stat > Nonparametrics > Friedman**.

2 In **Response**, enter the column containing the measurement data.

3 In **Treatment**, enter the column containing the treatment levels.

4 In **Blocks**, enter the column that contains the block levels.

5 If you like, use any of the dialog box options, then click **OK**.

Example of Friedman test

A randomized block experiment was conducted to evaluate the effect of a drug treatment on enzyme activity. Three different drug therapies were given to four animals, with each animal belonging to a different litter. The Friedman test provides the desired test of H0: all treatment effects are zero vs. H1: not all treatment effects are zero.

1 Open the worksheet EXH_STAT.MTW.

2 Choose **Stat > Nonparametrics > Friedman**.

3 In **Response**, enter **EnzymeActivity**.

4 In **Treatment**, enter **Therapy**. In **Blocks**, enter **Litter**. Click **OK**.

Session window output

Friedman Test: EnzymeActivity versus Therapy blocked by Litter

```
S = 2.38  DF = 2  P = 0.305
S = 3.80  DF = 2  P = 0.150 (adjusted for ties)

                          Sum
                           of
Therapy  N  Est Median  Ranks
1        4      0.2450    6.5
2        4      0.3117    7.0
3        4      0.5783   10.5

Grand median = 0.3783
```

Interpreting the results

The test statistic, S, has a p-value of 0.305, unadjusted for ties, and 0.150, adjusted for ties. For a levels 0.05 or 0.10, there is insufficient evidence to reject H0 because the p-value is greater than the α level. You therefore conclude that the data do not support the hypothesis that any of the treatment effects are different from zero.

The estimated medians associated with treatments are the grand median plus estimated treatment effects. The sum of ranks value is the sum of the treatment ranks, when ranked within each block. These values can serve as a measure of the relative size of treatment medians and are used in calculating the test statistic. The grand median is the median of the adjusted block medians. See Calculating treatment effects, adjusted block medians, and the grand median for more information.

Runs Test

Runs Test

Stat > Nonparametrics > Runs Test

Use Runs Test to see if a data order is random. Runs Test is a nonparametric test because no assumption is made about population distribution parameters. Use this test when you want to determine if the order of responses above or below a specified value is random. A run is a set of consecutive observations that are all either less than or greater than a specified value.

Dialog box items

Variables: Select the columns containing the variables you want to test for randomness.

Above and below the mean: Choose to use the mean as the baseline to determine the number of runs.

Above and below: Choose to use a value other than the mean as the baseline to determine the number of runs, then enter a value.

Data – Runs Test

You need at least one column of numeric data. If you have more than one column of data, Minitab performs a Runs Test separately for each column.

You may have missing data at the beginning or end of a data column, but not in the middle. You must omit missing data from the middle of a column before using this procedure.

To perform a runs test

1 Choose **Stat > Nonparametrics > Runs Test**.

2 In **Variables**, enter the columns containing the data you want to test for randomness.

3 If you like, use any dialog box options, then click **OK**.

Example of a runs test

Suppose an interviewer selects 30 people at random and asks them each a question for which there are four possible answers. Their responses are coded 0, 1, 2, 3. You wish to perform a runs test in order to check the randomness of answers. Answers that are not in random order may indicate that a gradual bias exists in the phrasing of the questions or that subjects are not being selected at random.

1 Open the worksheet EXH_STAT.MTW.

2 Choose **Stat > Nonparametrics > Runs Test**.

3 In **Variables**, enter *Response*. Click **OK**.

Session window output

Runs Test: Response

```
Runs test for Response

Runs above and below K = 1.23333

The observed number of runs = 8
The expected number of runs = 14.9333
11 observations above K, 19 below
P-value = 0.005
```

Interpreting the results

Because the option of a value other than the mean was not specified as the comparison criterion (K), the mean, 1.233, was used. There are eight runs.

(0, 0, 1, 0, 1, 1, 1, 1) (2, 3, 3, 2) (0, 0, 0) (2) (1, 1, 1, 1) (2, 3, 3) (0, 0, 1, 0) (2, 2, 3)

To determine if this is an unusual number of runs, Minitab calculates the number of observations above and below K. From these values, Minitab calculates the expected number of runs. Because the resulting p-value (0.0055) is smaller than the alpha level of 0.05, there is sufficient evidence to conclude that the data are not in random order.

Pairwise Averages

Pairwise Averages

Stat > Nonparametrics > Pairwise Averages

Pairwise Averages calculates and stores the average for each possible pair of values in a single column, including each value with itself. Pairwise averages are also called Walsh averages. Pairwise averages are used, for example, for the Wilcoxon method.

Dialog box items

Variable: Enter the column for which you want to obtain averages.

Store averages in: Specify the storage column for the Walsh averages. For n data values, Minitab stores n (n + 1) / 2 pairwise (or Walsh) averages.

Store indices in: Check to store the indices for each average, then specify two columns which will contain the indices. The Walsh average, (xi + xj) / 2, has indices i and j. The value of i is put in the first column and the value of j is put in the second storage column.

Data – Pairwise Averages

You must have one numeric data column. If you have missing data, the pairwise averages involving the missing values are set to missing.

To calculate Pairwise Averages

1 Choose **Stat > Nonparametrics > Pairwise Averages**.

2 In **Variable**, enter the column for which you want to obtain averages.

3 In **Store averages** in, enter a column name or number to store the pairwise (Walsh) averages. For n data values, Minitab stores n (n + 1) / 2 pairwise averages.

4 If you like, use one or more of the dialog box options, then click **OK**.

Example of pairwise averages

The example illustrates how to calculates the average of all possible pairs of values, including each value with itself.

1 In C1 of the Data window, type *1 2 3*.

2 Choose **Stat > Nonparametrics > Pairwise Averages**.

2 In **Variable**, enter *C1*.

3 In **Store averages in**, enter *C2*.

4 Check **Store indices in**, then type *C3* and *C4* in the text boxes. Click **OK**.

Data window output

This command does not produce output in the session window. To see the results, look in the data window or use **Data > Display Data**.

Row	C1	C2	C3	C4
1	1	1.0	1	1
2	2	1.5	1	2
3	3	2.0	2	2
4		2.0	1	3
5		2.5	2	3
6		3.0	3	3

Pairwise Differences

Pairwise Differences

Stat > Nonparametrics > Pairwise Differences

Pairwise Differences calculates and stores the differences between all possible pairs of values formed from two columns. These differences are useful for nonparametric tests and confidence intervals. For example, the point estimate given by Mann-Whitney can be computed as the median of the differences.

Dialog box items

First variable: Select the first column. The column you select in **Second variable** will be subtracted from this column.

Second variable: Select the second column. This column will be subtracted from the column you selected in **First Variable**.

Store differences in: Specify the storage column for the differences. For n data values, Minitab stores n (n + 1) / 2 pairwise differences.

Store indices in: Check to store the indices for each difference, then specify two columns which will contain the indices. The difference, ($x_i – y_j$), has indices i and j. The value of i is put in the first column and the value of j is put in the second storage column.

Data – Pairwise Differences

You must have two numeric data columns. If you have missing data, the pairwise differences involving the missing values are set to missing.

To calculate Pairwise Differences

1 Choose **Stat > Nonparametrics > Pairwise Differences**.

2 In **First variable**, enter a column. The column you enter in **Second variable** will be subtracted from this column.

3 In **Second variable**, enter a column. This column will be subtracted from the column you entered in **First variable**.

4 In **Store differences** in, enter a column name or number to store the differences. For n data values, Minitab stores n (n + 1) / 2 pairwise differences.

5 If you like, use one or more of the dialog box options, then click **OK**.

Example of pairwise differences

The example illustrates how to computes all possible differences between pairs of elements from two columns.

1 In the Data window, in rows 1 to 3 of C1, type *3 5 2*. In rows 1 to 3 of C2, type *1.1 2.0 1.1*.

2 Choose **Stat > Nonparametrics > Pairwise Differences.**

3 In **First Variable**, enter *C1*.

4 In **Second variable**, enter *C2*.

3 In **Store differences in**, type *C3*.

4 Check **Store indices in**, then type *C4* and *C5* in the text boxes. Click **OK**.

Data window output

This command does not produce output in the session window. To see the results, look in the data window or use **Data > Display Data**.

Row	C1	C2	C3	C4	C5
1	3	1.1	1.9	1	1
2	5	2.0	1.0	1	2
3	2	1.1	1.9	1	3
4			3.9	2	1
5			3.0	2	2
6			3.9	2	3
7			0.9	3	1
8			0.0	3	2
9			0.9	3	3

Pairwise Slopes

Pairwise Slopes

Stat > Nonparametrics > Pairwise Slopes

Pairwise Slopes calculates and stores the slope between all possible pairs of points, where a row in y–x columns defines a point in the plane. This procedure is useful for finding robust estimates of the slope of a line through the data.

Dialog box items

Y variable: Select the column containing the y-variable.

X variable: Select the column containing the x-variable.

Store slopes in: Specify the storage column for the slopes.

Store indices in: Check to store the indices for each slope, then specify two columns which will contain the indices.

Data – Pairwise Slopes

You must have two numeric data columns, one that contains the response variable (y) and one that contains the predictor variable (x). If you have missing data or the slope is not defined (e.g. slope of a line parallel to the y axis), the slope will be stored as missing.

To calculate Pairwise Slopes

1 Choose **Stat > Nonparametrics > Pairwise Slopes**.

2 In **Y variable**, enter the column containing the response data.

3 In **X variable**, enter the column containing the predictor data.

4 In **Store slopes** in, enter a column name or number to store the pairwise slopes. For n pairs, Minitab stores (n–1) / 2 slopes.

5 If you like, use one or more of the dialog box options, then click **OK**.

Example of calculating pairwise slopes

The example illustrates how compute the slope between every pair of data from two columns of equal length.

1. In the Data window, in rows 1 to 3 of C1, type *3 5 2 6*. In rows 1 to 3 of C2, type *1.1 2.0 1.1 3.0*.
2. Choose **Stat > Nonparametrics > Pairwise Slopes**.
3. In **Y Variable**, enter *C1*.
4. In **X variable**, enter *C2*.
3. In **Store Slopes in**, type *C3*.
4. Check **Store indices in**, then type *C4* and *C5* in the text boxes. Click **OK**.

Data window output

This command does not produce output in the session window. To see the results, look in the data window or use **Data > Display Data**.

Row	C1	C2	C3	C4	C5
1	3	1.1	2.22222	2	1
2	5	2.0	*	3	1
3	2	1.1	3.33333	3	2
4	6	3.0	1.57895	4	1
5			1.00000	4	2
6			2.10526	4	3

Time Series

Overview

Time Series Overview

Mınitab's time series procedures can be used to analyze data collected over time, commonly called a time series. These procedures include simple forecasting and smoothing methods, correlation analysis methods, and ARIMA modeling. Although correlation analysis may be performed separately from ARIMA modeling, we present the correlation methods as part of ARIMA modeling.

Simple forecasting and smoothing methods are based on the idea that reliable forecasts can be achieved by modeling patterns in the data that are usually visible in a time series plot, and then extrapolating those patterns to the future. Your choice of method should be based upon whether the patterns are *static* (constant in time) or *dynamic* (changes in time), the nature of the trend and seasonal components, and how far ahead that you wish to forecast. These methods are generally easy and quick to apply.

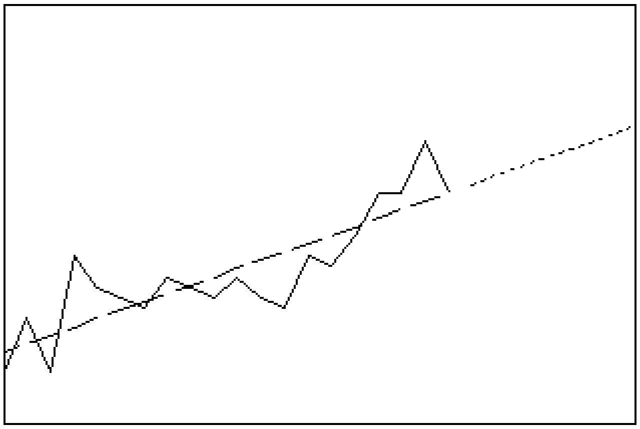
ARIMA modeling also makes use of patterns in the data, but these patterns may not be easily visible in a plot of the data. Instead, ARIMA modeling uses differencing and the autocorrelation and partial autocorrelation functions to help identify an acceptable model. ARIMA stands for Autoregressive Integrated Moving Average, which represent the filtering steps taken in constructing the ARIMA model until only random noise remains. While ARIMA models are valuable for modeling temporal processes and are also used for forecasting, fitting a model is an iterative approach that may not lend itself to application speed and volume.

Simple forecasting and smoothing methods

The simple forecasting and smoothing methods model components in a series that are usually easy to see in a time series plot of the data. This approach decomposes the data into its component parts, and then extends the estimates of the components into the future to provide forecasts. You can choose from the static methods of trend analysis and decomposition, or the dynamic methods of moving average, single and double exponential smoothing, and Winters' method. *Static* methods have patterns that do not change over time; *dynamic* methods have patterns that do change over time and estimates are updated using neighboring values.

You may use two methods in combination. That is, you may choose a static method to model one component and a dynamic method to model another component. For example, you may fit a static trend using trend analysis and dynamically model the seasonal component in the residuals using Winters' method. Or, you may fit a static seasonal model using decomposition and dynamically model the trend component in the residuals using double exponential smoothing. You might also apply a trend analysis and decomposition together so that you can use the wider selection of trend models offered by trend analysis (see Example of trend analysis and Example of decomposition). A disadvantage of combining methods is that the confidence intervals for forecasts are not valid.

For each of the methods, the following table provides a summary and a graph of fits and forecasts of typical data.

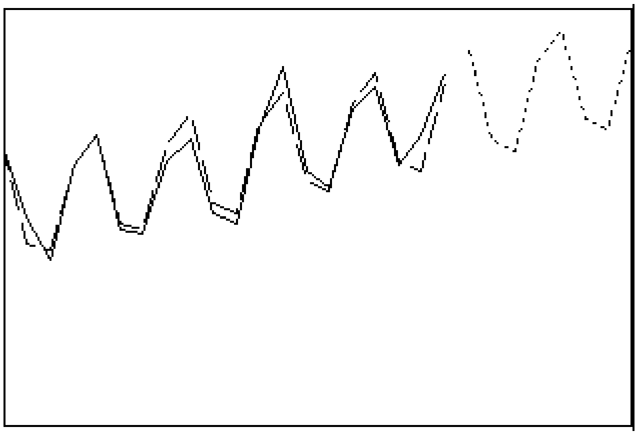
Command	Forecast	Example
Trend Analysis		
Fits a general trend model to time series data. Choose among the linear, quadratic, exponential growth or decay, and S-curve models. Use this procedure to fit trend when there is no seasonal component in your series.	Length: long Profile: extension of trend line	

Decomposition

Separates the times series into linear trend and seasonal components, as well as error. Choose whether the seasonal component is additive or multiplicative with the trend. Use this procedure to forecast when there is a seasonal component in your series or if you simply want to examine the nature of the component parts

Length: long

Profile: trend with seasonal pattern

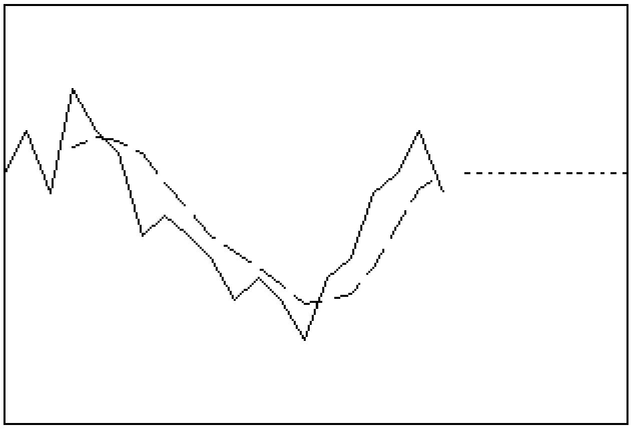


Moving Average

Smoothes your data by averaging consecutive observations in a series. This procedure can be a likely choice when your data do not have a trend or seasonal component. There are ways, however, to use moving averages when your data possess trend and/or seasonality.

Length: short

Profile: flat line

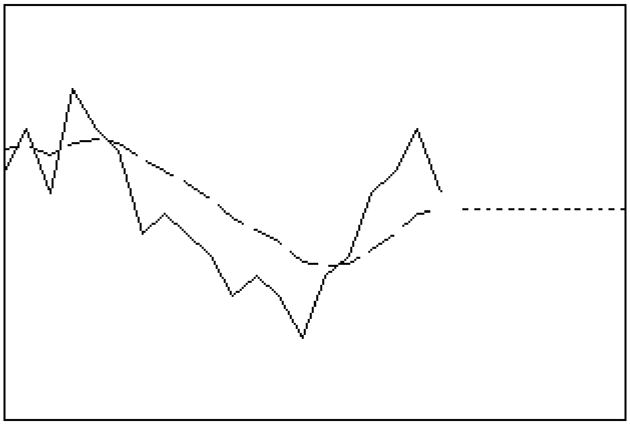


Single Exp Smoothing

Length: short

Smoothes your data using the optimal one-step ahead ARIMA (0,1,1) forecasting formula. This procedure works best without a trend or seasonal component. The single dynamic component in a moving average model is the level.

Profile: flat line

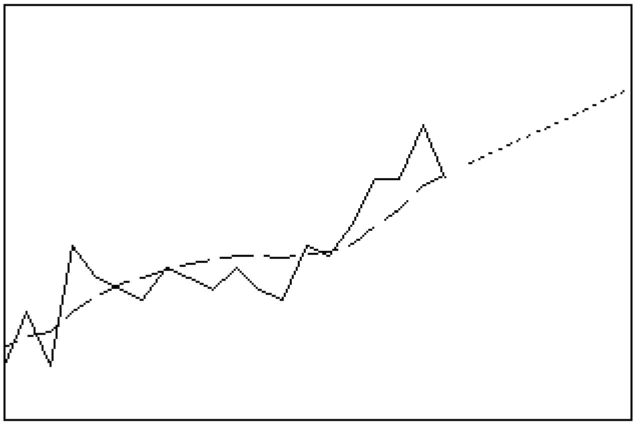


Double Exp Smoothing

Smoothes your data using the optimal one-step-ahead ARIMA (0,2,2) forecasting formula. This procedure can work well when trend is present but it can also serve as a general smoothing method. Double Exponential Smoothing calculates dynamic estimates for two components: level and trend.

Length: short

Profile: straight line with slope equal to last trend estimate

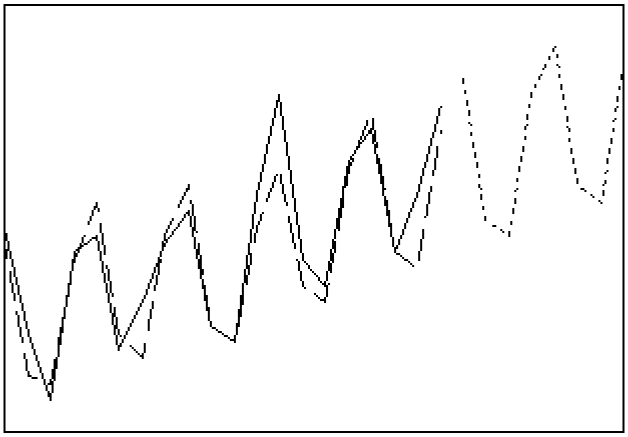


Winters' Method

Smoothes your data by Holt-Winters exponential smoothing. Use this procedure when trend and seasonality are present, with these two components being either additive or multiplicative. Winters' Method calculates dynamic estimates for three components: level, trend, and seasonal.

Length: short to medium

Profile: trend with seasonal pattern



Correlation analysis and ARIMA modeling

Examining correlation patterns within a time series or between two time series is an important step in many statistical analyses. The correlation analysis tools of differencing, autocorrelation, and partial autocorrelation are often used in ARIMA modeling to help identify an appropriate model.

ARIMA modeling can be used to model many different time series, with or without trend or seasonal components, and to provide forecasts. The forecast profile depends upon the model that is fit. The advantage of ARIMA modeling compared to the simple forecasting and smoothing methods is that it is more flexible in fitting the data. However, identifying and fitting a model may be time-consuming, and ARIMA modeling is not easily automated.

- Differences computes and stores the differences between data values of a time series. If you wish to fit an ARIMA model but there is trend or seasonality present in your data, differencing data is a common step in assessing likely ARIMA models. Differencing is used to simplify the correlation structure and to reveal any underlying pattern.
- Lag computes and stores the lags of a time series. When you lag a time series, Minitab moves the original values down the column, and inserts missing values at the top of the column. The number of missing values inserted depends on the length of the lag.
- Autocorrelation computes and plots the autocorrelations of a time series. Autocorrelation is the correlation between observations of a time series separated by k time units. The plot of autocorrelations is called the autocorrelation function or acf. View the acf to guide your choice of terms to include in an ARIMA model.
- Partial Autocorrelation computes and plots the partial autocorrelations of a time series. Partial autocorrelations, like autocorrelations, are correlations between sets of ordered data pairs of a time series. As with partial correlations in the regression case, partial autocorrelations measure the strength of relationship with other terms being accounted for. The partial autocorrelation at a lag of k is the correlation between residuals at time t from an autoregressive model and observations at lag k with terms for all intervening lags present in the autoregressive model. The plot of partial autocorrelations is called the partial autocorrelation function or pacf. View the pacf to guide your choice of terms to include in an ARIMA model.
- Cross Correlation computes and graphs correlations between two time series.
- ARIMA fits a Box-Jenkins ARIMA model to a time series. ARIMA stands for Autoregressive Integrated Moving Average. The terms in the name–Autoregressive, Integrated, and Moving Average–represent filtering steps taken in constructing the ARIMA model until only random noise remains. Use ARIMA to model time series behavior and to generate forecasts.

Time Series

Stat > Time Series

Time Series Plot – plots a time series

Trend Analysis – fits a trend line using a linear, quadratic, growth, or S-curve model

Decomposition – performs classical decomposition on a time series, using either a multiplicative or an additive model

Moving Average – calculates moving averages, which can be used to either smooth a time series or to generate forecasts

Single Exp Smoothing – smoothes out the noise in a time series and forecasts future values of the series

Double Exp Smoothing – smoothes out the noise in a time series and forecasts data that exhibit a trend; uses either Holt or Brown double exponential smoothing method

Winters' Method – does Holt-Winters seasonal exponential smoothing, using either a multiplicative or an additive model, to smooth and forecast data which exhibit both a trend and a seasonal pattern

Differences – differences a series

Lag – lags a series

Autocorrelation – computes an autocorrelation function

Partial Autocorrelation – computes a partial autocorrelation function

Cross Correlation – computes a cross correlation function

ARIMA – fits a Box-Jenkins ARIMA model

Time Series Examples

The following examples illustrate how evaluate time series with Minitab. Choose an example below:

Time Series Plot

Trend Analysis

Decomposition

Fits and Forecasts of Combined Trend Analysis and Decomposition

Moving Average

Single Exponential Smoothing

Double Exponential Smoothing

Winters' Method

Autocorrelation

Partial Autocorrelation

ARIMA

References – Time Series

[1] B.L. Bowerman and R. T. O' Connell (1993). *Forecasting and Time Series: An Applied Approach*, 3rd edition. Duxbury Press.

[2] G.E.P. Box and G.M. Jenkins (1994). *Time Series Analysis: Forecasting and Control*, 3rd Edition. Prentice Hall.

[3] J.D. Cryer (1986). *Time Series Analysis*. Duxbury Press.

[4] N.R. Farnum and L.W. Stanton (1989). *Quantitative Forecasting Methods*. PWS-Kent.

[5] G.M. Ljung and G.E.P. Box (1978). "On a Measure of Lack of Fit in Time Series Models," *Biometrika*, 65, 67–72.

[6] S. Makridakis, S.C. Wheelwright, and R. J. Hyndman (1998). *Forecasting: Methods and Applications*. Wiley.

[7] D.W. Marquardt (1963). "An Algorithm for Least Squares Estimation of Nonlinear Parameters," *Journal Soc. Indust. Applied Mathematics*, 11, 431–441.

[8] W.Q. Meeker, Jr. (1977). "TSERIES–A User-oriented Computer Program for Identifying, Fitting and Forecasting ARIMA Time Series Models," ASA 1977 Proceedings of the Statistical Computing Section.

[9] W.Q. Meeker, Jr. (1977). *TSERIES User's Manual*, Statistical Laboratory, Iowa State University.

[10] W. Vandaele (1983). *Applied Time Series and Box-Jenkins Models*. Academic Press, Inc.

Acknowledgment

The ARIMA algorithm is based on the fitting routine in the TSERIES package written by Professor William Q. Meeker, Jr., of Iowa State University [8], [9]. We are grateful to Professor Meeker for his help in the adaptation of his routine to Minitab.

Time Series Plot

Time Series Plot

Gallery

Data

Simple

To display a simple time series plot

Example, simple

With Groups

To display a time series plot with groups

Example, with groups

Multiple

To display multiple overlaid time series plots

Example, multiple

Multiple With Groups

To display multiple overlaid time series plots with groups

Example, multiple with groups

Time

To customize the time scale

Example, using a stamp for the x-axis

Example, using custom start times

Trend Analysis

Trend Analysis

Stat > Time Series > Trend Analysis

Trend analysis fits a general trend model to time series data and provides forecasts. Choose among the linear, quadratic, exponential growth or decay, and S-curve models. Use this procedure to fit trend when there is no seasonal component to your series.

Dialog box items

Variable: Enter the column containing the time series.

Model Type: Select the model that you want. Use care when interpreting the coefficients from the different models, as they have different meanings. See [4] for details.

Linear: Click to fit the linear trend model.

Quadratic: Click to fit the quadratic trend model.

Exponential growth: Click to fit the exponential growth trend model.

S-Curve (Pearl-Reed logistic): Click to fit the Pearl-Reed logistic S-curve trend model. You cannot have missing data when fitting the S-curve model.

Generate forecasts: Check to generate forecasts.

Number of forecasts: Enter an integer to indicate how many forecasts that you want.

Starting from origin: Enter a positive integer to specify a starting point for the forecasts. For example, if you specify 4 forecasts and 48 for the origin, Minitab computes forecasts for periods 49, 50, 51, and 52. If you leave this space blank, Minitab generates forecasts from the end of the data. Minitab uses data up to the origin for fitting the trend model used to generate forecasts.

<Time>

<Options>

<Storage>

<Graphs>

<Results>

Data – Trend Analysis

The time series must be in one numeric column. If you choose the S-curve trend model, you must delete missing data from the worksheet before performing the trend analysis. Minitab automatically omits missing values from the calculations when you use one of the other three trend models.

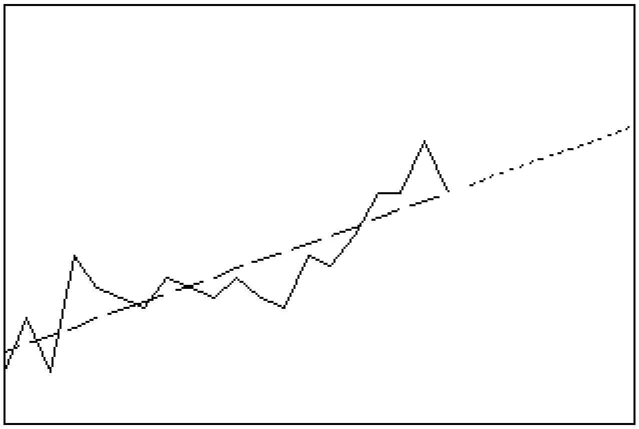
To perform a Trend Analysis

1 Choose **Stat > Time Series > Trend Analysis**.

2 In **Variable**, enter the column containing the series.

3 If you like, use any dialog box options, then click **OK**.

Trend Analysis : When to Use



Use for:

- Data with constant trend, and
- Data with no seasonal pattern
- Long range forecasting

Forecast profile:

- Continuation of trend line fit to data

ARIMA equivalent: none

Measures of Accuracy

Minitab computes three measures of accuracy of the fitted model: MAPE, MAD, and MSD for each of the simple forecasting and smoothing methods. For all three measures, the smaller the value, the better the fit of the model. Use these statistics to compare the fits of the different methods.

MAPE, or Mean Absolute Percentage Error, measures the accuracy of fitted time series values. It expresses accuracy as a percentage.

$$\text{MAPE} = \frac{\Sigma \left| (y_t - \hat{y}_t)/y_t \right|}{n} \times 100 \qquad (y_t \neq 0)$$

where y_t equals the actual value, $\hat{y}_t$ equals the fitted value, and n equals the number of observations.

MAD, which stands for Mean Absolute Deviation, measures the accuracy of fitted time series values. It expresses accuracy in the same units as the data, which helps conceptualize the amount of error.

$$\mathrm{MAD} = \frac{\sum_{t=1}^{n} |y_t - \hat{y}_t|}{n}$$

where y_t equals the actual value, $\hat{y}_t$ equals the fitted value, and n equals the number of observations.

MSD stands for Mean Squared Deviation. MSD is always computed using the same denominator, n, regardless of the model, so you can compare MSD values across models. MSD is a more sensitive measure of an unusually large forecast error than MAD.

$$\mathrm{MSD} = \frac{\sum_{t=1}^{n} |y_t - \hat{y}_t|^2}{n}$$

where y_t equals the actual value, $\hat{y}_t$ equals the forecast value, and n equals the number of forecasts.

Forecasts – Trend Analysis

Forecasts are extrapolations of the trend model fits. Data prior to the forecast origin are used to fit the trend.

Trend models

There are four different trend models you can choose from: linear (default), quadratic, exponential growth curve, or S-curve (Pearl-Reed logistic). Use care when interpreting the coefficients from the different models, as they have different meanings. See [4] for details.

Trend analysis by default uses the *linear trend* model:

$Y_t = \beta_0 + (\beta_1 * t) + e_t$

In this model, β_1 represents the average change from one period to the next.

The *quadratic trend model* which can account for simple curvature in the data, is:

$Y_t = \beta 0 + \beta_1 * t + (\beta_2 * t^2) + e_t$

The *exponential growth trend model* accounts for exponential growth or decay. For example, a savings account might exhibit exponential growth. The model is:

$Y_t = \beta_0 * \beta_1^t * e_t$

The *S-curve model* fits the Pearl-Reed logistic trend model. This accounts for the case where the series follows an S-shaped curve. The model is:

$Y_t = (10^a) / (\beta_0 + \beta_1 \beta_2^t)$

Time

Stat > Time Series > *menu command* **> Time**

Specify the time scale.

Dialog box items

Time Scale: Change the x-axis of the scale you specify.

Index: Choose to number the x-axis with a single scale from 1 to n by ones (where n is the number of observations in the column containing the time series).

Calendar: Choose to label the x-axis with different time unit scales. For example, if you choose Month Quarter Year, Minitab assumes that your data are in intervals of 1 month and generates 3 scales on the x-axis: 1 for months, 1 for quarters, and 1 for years. At each 4th tick on the month scale, Minitab generates a tick on the quarter scale. At each 4th tick on the quarter scale, Minitab generates a tick on the Year scale.

Clock: Choose to label the x-axis with different time unit scales. For example, if you choose Day Hour Minute, Minitab assumes that your data are in minute intervals and generates 3 scales on the x-axis: 1 for days, 1 for hours, and 1 for minutes. At each 60th tick on the minute scale, Minitab generates a tick on the hour scale. At each 24th tick on the hour scale, Minitab generates a tick on the Day scale.

Start value(s): Enter the start values.

Increment: Enter a value to increment the time scale.

Stamp: Choose to add a date/time stamp on the x-axis. Enter a date/time column in the text box. The values are stamped on the plot in the same format as they appear in the column.

Trend Analysis – Options

Stat > Time Series > Trend Analysis > Options

Specify a customized title and the trend analysis fit to be a weighted average of a fit of prior data and the current data.

Dialog box items

Title: To replace the default title with your own custom title, type the desired text in this box.

Prior parameter values: Specify the parameter values of the trend analysis fit to the prior data for the weighted average fit. See Weighted average trend analysis.

Weights for blending priors with new estimates: Enter weights of coefficients of current data for the weighted average fit. Enter numbers between 0 and 1 for weights of each trend model parameter estimate. See Weighted average trend analysis.

To perform a weighted average trend analysis

1. Choose **Stat > Time Series > Trend Analysis > Options**.
2. Enter the coefficient estimates from the prior trend analysis in the order in which they are given in the Session window or the graph.
3. Optionally enter weights between 0 and 1 for each new coefficient, in the same order as for coefficients. Default weights of 0.2 will be used for each coefficient if you don't enter any. If you do enter weights, the number that you enter must be equal to the number of coefficients.
4. If you like, use any dialog box options, then click **OK**.

Minitab generates a time series plot of the data, plus a second time series plot that shows trend lines for three models. The Session window displays the coefficients and accuracy measures for all three models.

Weighted average trend analysis

You can perform a weighted average trend analysis to incorporate knowledge learned from fitting the same trend model to prior data in order to obtain an "improved" fit to the present data. The smoothed trend line combines prior and new information in much the same way that exponential smoothing works. In a sense, this smoothing of the coefficients filters out some of the noise from the model parameters estimated in successive cycles.

If you supply coefficients from a prior trend analysis fit, Minitab performs a weighted trend analysis. If the weight for a particular coefficient is a, Minitab estimates the new coefficient by:

$\alpha p_1 + (1 - \alpha) p_2$, where p_1 is the coefficient estimated from the current data and p_2 is the prior coefficient.

Trend Analysis – Storage

Stat > Time Series > Trend Analysis > Storage

Stores fits, residuals, and forecasts in the worksheet.

Dialog box items

Fits (trend line): Check to store the fitted values. These are the values used to plot the trend line. You should store the fitted values if you want to generate diagnostic residual plots.

Residuals (detrended data): Check to store the residuals. If you store the residuals you can generate diagnostic plots, using Autocorrelation. For the linear, quadratic, and S-curve models, the detrended data equal the original data minus the fits.

Forecasts: Check to store the forecasts. This option is available only if you generated forecasts in the main Trend Analysis dialog box.

Trend Analysis – Graphs

Stat > Time Series > Trend Analysis > Graphs

Displays time series plot and residual plots for diagnosis of model fit.

Dialog box items

Time series plot (including optional forecasts)

Display plot: Choose to display the time series plot..

Do not display plot: Choose to suppress the time series plot.

Residual Plots

Individual plots: Choose to display one or more plots.

Histogram of residuals: Check to display a histogram of the residuals.

Normal plot of residuals: Check to display a normal probability plot of the residuals.

Residuals versus fits: Check to plot the residuals versus the fitted values.

Residuals versus order: Check to plot the residuals versus the order of the data. The row number for each data point is shown on the x-axis–for example, 1 2 3 4... n.

Four in one: Choose to display a layout of a histogram of residuals, a normal plot of residuals, a plot of residuals versus fits, and a plot of residuals versus order.

Residuals versus the variables: Enter one or more columns containing the variables against which you want to plot the residuals. Minitab displays a separate graph for each column.

Trend Analysis – Results

Stat > Time Series > Trend Analysis > Results

Controls Session window output.

Dialog box items

Control the Display of Results

Display nothing: Choose to suppress output.

Summary table: Choose to display the default output – fitted trend equation and accuracy measures.

Summary table and results table: Choose to display the default output plus a table of the original series, the trend and the detrended data.

Example of Trend Analysis

You collect employment data in a trade business over 60 months and wish to predict employment for the next 12 months. Because there is an overall curvilinear pattern to the data, you use trend analysis and fit a quadratic trend model. Because there is also a seasonal component, you save the fits and residuals to perform decomposition of the residuals (see Example of Decomposition).

1 Open the worksheet EMPLOY.MTW.

2 Choose **Stat > Time Series > Trend Analysis**.

3 In **Variable**, enter *Trade*.

4 Under **Model Type**, choose **Quadratic**.

5 Check **Generate forecasts** and enter *12* in **Number of forecasts**.

6 Click **Storage**.

7 Check **Fits (Trend Line)**, **Residuals (detrended data)**, and **Forecasts**. Click **OK** in each dialog box.

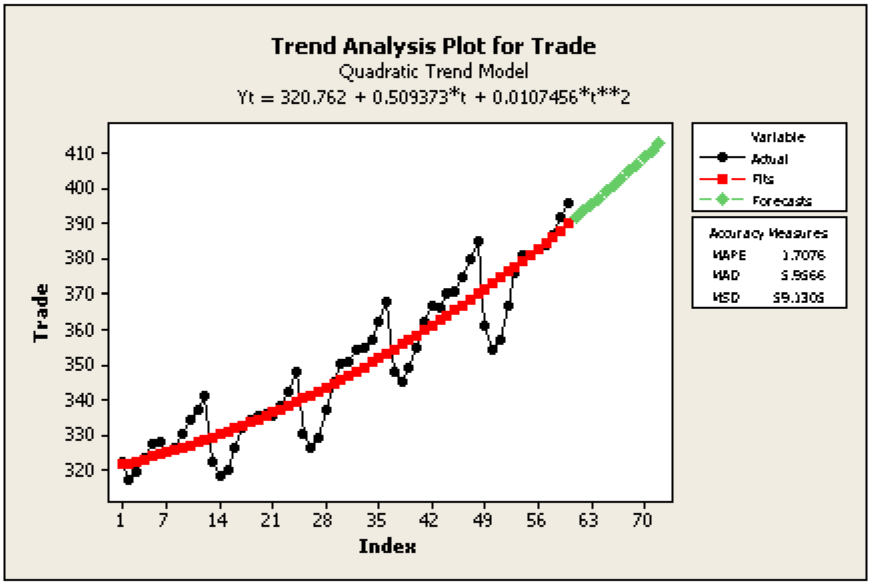
Session Window Output

Trend Analysis for Trade

```
Data      Trade
Length    60
NMissing  0


Fitted Trend Equation

Yt = 320.762 + 0.509373*t + 0.0107456*t**2


Accuracy Measures

MAPE   1.7076
MAD    5.9566
```

```
MSD   59.1305


Forecasts

Period  Forecast
61        391.818
62        393.649
63        395.502
64        397.376
65        399.271
66        401.188
67        403.127
68        405.087
69        407.068
70        409.071
71        411.096
72        413.142
```

Graph window output



Interpreting the results

The trend plot that shows the original data, the fitted trend line, and forecasts. The Session window output also displays the fitted trend equation and three measures to help you determine the accuracy of the fitted values: MAPE, MAD, and MSD. The trade employment data show a general upward trend, though with an evident seasonal component. The trend model appears to fit well to the overall trend, but the seasonal pattern is not well fit. To better fit these data, you also use decomposition on the stored residuals and add the trend analysis and decomposition fits and forecasts (see Example of decomposition).

Decomposition

Decomposition

Stat > Time Series > Decomposition

You can use decomposition to separate the time series into linear trend and seasonal components, as well as error, and provide forecasts. You can choose whether the seasonal component is additive or multiplicative with the trend. Use this procedure when you wish to forecast and there is a seasonal component to your series, or if you simply want to examine the nature of the component parts. See [6] for a discussion of decomposition methods.

Dialog box items

Variable: Enter the column containing the time series.

Seasonal Length: Enter a positive integer greater than or equal to 2. This is the length of the seasonal component. For example, if you have monthly data, you might use a seasonal length of 12.

Model Type:

Multiplicative: Choose to use the multiplicative model.

Additive: Choose to use the additive model.

Model Components:

Trend plus seasonal: Choose to include the trend component in the decomposition.

Seasonal only: Choose to omit the trend component from the decomposition. You might want to do this if you have already detrended your data with Trend Analysis.

Caution If the data contain a trend component but you omit it from the decomposition, the estimates of the seasonal indices may be affected.

Generate forecasts: Check if you want to generate forecasts.

Number of forecasts: Enter an integer to indicate how the number of forecasts.

Starting from origin: Enter a positive integer to specify a starting point for the forecasts. For example, if you specify 4 forecasts and 48 for the origin, Minitab computes forecasts for periods 49, 50, 51, and 52. If you leave this space blank, Minitab generates forecasts from the end of the data.

<Time>

<Options>

<Storage>

<Graphs>

<Results>

Data – Decomposition

The time series must be in one numeric column. Minitab automatically omits missing data from the calculations.

The data that you enter depends upon how you use this procedure. Usually, decomposition is performed in one step by simply entering the time series. Alternatively, you can perform a decomposition of the trend model residuals. This process may improve the fit of the model by combining the information from the trend analysis and the decomposition. See Decomposition of trend model residuals .

To perform a Decomposition

1 Choose **Stat > Time Series > Decomposition**.

2 In **Variable**, enter the column containing the series.

3 In **Seasonal length**, enter the seasonal length or period.

4 If you like, use any dialog box options, then click **OK**.

To combine trend analysis and decomposition

1 Perform a Trend Analysis and store the fits, residuals, and forecasts (see Example of a trend analysis).

2 Choose **Stat > Time Series > Decomposition**.

3 In **Variable**, enter the column containing trend analysis residuals.

4 Under **Model Type**, choose **Additive**.

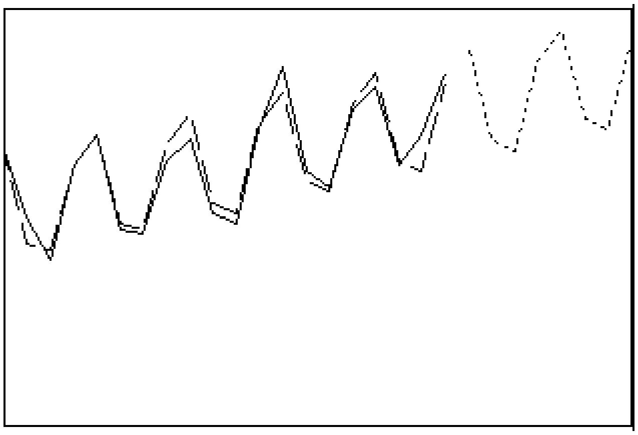
5 Under **Model Components**, choose **Seasonal only**.

6 Click **Storage** and check **Fits**. Click **OK** in each dialog box.

7 Next, you need to calculate the fits from the combined procedure. If you want these components to be additive, add the respective fits together.

Note The MAPE, MAD, MSD accuracy measures from decomposition used in this manner are not comparable to these statistics calculated from other procedures, but you can calculate the comparable values fairly easily. We demonstrate this with MSD in the decomposition example.

Decomposition, Additive Model : When to Use



Use for:

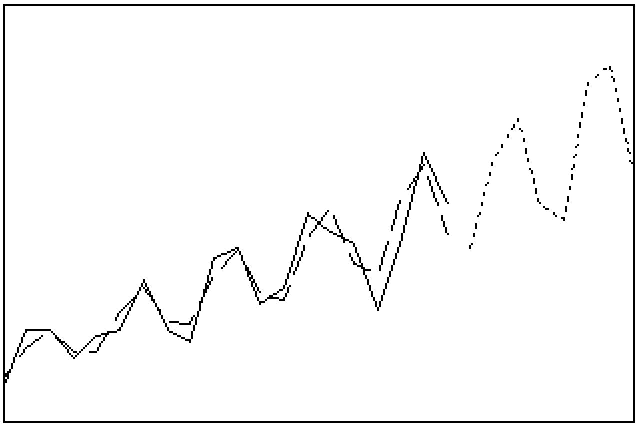
- Data with either no trend or constant trend, and
- Data with constant seasonal pattern
- Size of seasonal pattern not proportional to data
- Long range forecasting

Forecast profile:

- Straight line with seasonal pattern added

ARIMA equivalent: none

Decomposition, Multiplicative Model : When to Use



Use for:

- Data with either no trend or constant trend, and
- Data with constant seasonal pattern
- Size of seasonal pattern proportional to data
- Long range forecasting

Forecast profile:

- Straight line multiplied by seasonal pattern

ARIMA equivalent: none

How Minitab Does Decomposition

The decomposition model

By default, Minitab uses a *multiplicative model*. Use the multiplicative model when the size of the seasonal pattern in the data depends on the level of the data. This model assumes that as the data increase, so does the seasonal pattern. Most time series exhibit such a pattern. The multiplicative model is

Y_t = Trend ∗ Seasonal ∗ Error

where Y_t is the observation at time t.

The *additive model* is:

Y_t = Trend + Seasonal + Error

where Y_t is the observation at time t.

You can also omit the trend component from the decomposition. You will probably choose this if you have already detrended your data with the trend analysis procedure. If the data contain a trend component but you omit it from the decomposition, this can influence the estimates of the seasonal indices.

Method

Decomposition involves the following steps:

1 Minitab fits a trend line to the data, using least squares regression.

2 Next, the data are detrended by either dividing the data by the trend component (multiplicative model) or subtracting the trend component from the data (additive model).

3 Then, the detrended data are smoothed using a centered moving average with a length equal to the length of the seasonal cycle. When the seasonal cycle length is an even number, a two-step moving average is required to synchronize the moving average correctly.

4 Once the moving average is obtained, it is either divided into (multiplicative model) or subtracted from (additive model) the detrended data to obtain what are often referred to as raw seasonals.

5 Within each seasonal period, the median value of the raw seasonals is found. The medians are also adjusted so that their mean is one (multiplicative model) or their sum is zero (additive model). These adjusted medians constitute the seasonal indices.

6 The seasonal indices are used in turn to seasonally adjust the data.

Forecasts – Decomposition

Decomposition calculates the forecast as the linear regression line multiplied by (multiplicative model) or added to (additive model) the seasonal indices. Data prior to the forecast origin are used for the decomposition.

Time

Stat > Time Series > *menu command* **> Time**

Specify the time scale.

Dialog box items

Time Scale: Change the x-axis of the scale you specify.

Index: Choose to number the x-axis with a single scale from 1 to n by ones (where n is the number of observations in the column containing the time series).

Calendar: Choose to label the x-axis with different time unit scales. For example, if you choose Month Quarter Year, Minitab assumes that your data are in intervals of 1 month and generates 3 scales on the x-axis: 1 for months, 1 for quarters, and 1 for years. At each 4th tick on the month scale, Minitab generates a tick on the quarter scale. At each 4th tick on the quarter scale, Minitab generates a tick on the Year scale.

Clock: Choose to label the x-axis with different time unit scales. For example, if you choose Day Hour Minute, Minitab assumes that your data are in minute intervals and generates 3 scales on the x-axis: 1 for days, 1 for hours, and 1 for minutes. At each 60th tick on the minute scale, Minitab generates a tick on the hour scale. At each 24th tick on the hour scale, Minitab generates a tick on the Day scale.

Start value(s): Enter the start values.

Increment: Enter a value to increment the time scale.

Stamp: Choose to add a date/time stamp on the x-axis. Enter a date/time column in the text box. The values are stamped on the plot in the same format as they appear in the column.

Decomposition – Options

Stat > Time Series > Decomposition > Options

Specify a customized title and the first observation in the seasonal period.

Dialog box items

Title: To replace the default title with your own custom title, type the desired text in this box. This title will be used for all three sets of plots.

First obs. is in seasonal period: Enter a number to specify a starting value. For example, if you have monthly data and the first observation is in June, then enter 6 to set the seasonal period correctly. By default, the starting value is 1 because Minitab assumes that the first data value in the series corresponds to the first seasonal period.

Decomposition – Storage

Stat > Time Series > Decomposition > Storage

Stores various values in the worksheet.

Dialog box items

Trend line: Check to store the trend component data. This is the trend component; it does not contain the error or seasonal component.

Detrended data: Check to store the detrended data (has seasonal and error components, no trend)

Seasonals: Check to store the seasonal component data (has seasonal component, no error or trend).

Seasonally adjusted data: Check to store the data which has had the seasonal component removed (has trend and error components, no seasonal).

Fits: Check to store the fitted values. These are the combination of trend and seasonal components, with no error component.

Residuals: Check to store the residuals. If you store the residuals you can generate diagnostic plots using Autocorrelation.

Forecasts: Check to store the forecasts. This option is available only if you generated forecasts in the main Decomposition dialog box.

Decomposition – Graphs

Stat > Time Series > Decomposition > Graphs

Displays time series plot and residual plots for diagnosis of model fit.

Dialog box items

Time series plot (including optional forecasts)

- **Display plot:** Choose to display the time series plot.
- **Do not display plot:** Choose to suppress the time series plot.

Residual Plots

- **Individual plots:** Choose to display one or more plots.
  - **Histogram of residuals:** Check to display a histogram of the residuals.
  - **Normal plot of residuals:** Check to display a normal probability plot of the residuals.
  - **Residuals versus fits:** Check to plot the residuals versus the fitted values.
  - **Residuals versus order:** Check to plot the residuals versus the order of the data. The row number for each data point is shown on the x-axis–for example, 1 2 3 4... n.
- **Four in one:** Choose to display a layout of a histogram of residuals, a normal plot of residuals, a plot of residuals versus fits, and a plot of residuals versus order.
- **Residuals versus the variables:** Enter one or more columns containing the variables against which you want to plot the residuals. Minitab displays a separate graph for each column.

Decomposition – Results

Stat > Time Series > Decomposition > Results

Controls Session window output.

Dialog box items

Control the Display of Results

Display nothing: Choose to suppress output.

Summary table: Choose to display the default output – fitted trend equation, the seasonal indices, and accuracy measures.

Summary table and results table: Choose to display the default output plus a table of the original series, the trend and seasonal components, the detrended data, the seasonally adjusted data, the residuals, and the fits.

Decomposition of trend model residuals

You can use trend analysis and decomposition in combination when your data have a trend that is fit well by the quadratic, exponential growth curve, or S-curve models of trend analysis and possess seasonality that can be fit well by decomposition.

Example of Fits and Forecasts of Combined Trend Analysis and Decomposition

You collect employment data in a trade business over 60 months and wish to predict employment for the next 12 months. Because there is an overall curvilinear pattern to the data, you use trend analysis and fit a quadratic trend model (see Example of Trend). Because a seasonal component also exists, you save the fits and forecasts to perform decomposition of the residuals (see Example of Decomposition). You now want to combine the trend analysis and decomposition results.

Step 1: Calculate the fits and forecasts of the combined trend analysis and decomposition

1. Open the worksheet EMPLOY.MTW.
2. Choose **Calc > Calculator**.
3. In **Store result in variable**, type *NewFits*.
4. In **Expression**, add the fits from trend analysis to the fits from decomposition. Click **OK**.
5. Choose **Calc > Calculator**.
6. Clear the **Expression** box by selecting the contents and pressing [Delete].
7. In **Store result in variable**, type *NewFore*.
8. In **Expression**, add the forecasts from trend analysis to the forecasts from decomposition. Click **OK**.

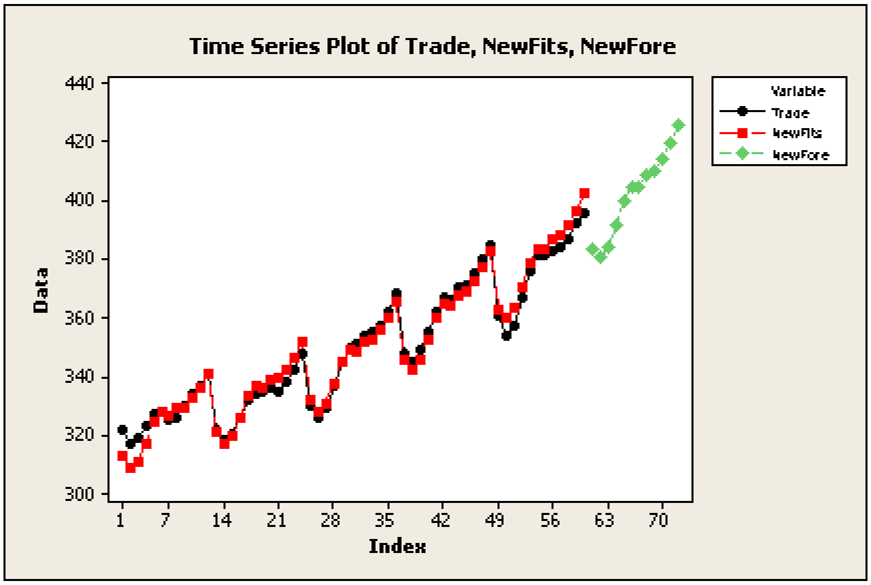
Step 2: Plot the fits and forecasts of the combined trend analysis and decomposition

1. Choose **Stat > Time Series > Time Series Plot >** choose ***Multiple*** **> OK**.
2. Under **Series**, enter *Trade NewFits NewFore*.
3. Click **Time/Scale**.
4. Under **Start Values**, choose **One set for each variable**, then enter *1 1 61* in rows 1–3 respectively.
5. Click **OK** in each dialog box.

Step 3: Calculate MSD

1. Choose **Calc > Calculator**.
2. In **Store result in variable**, type *MSD*.
3. Clear the **Expression** box by selecting the contents and pressing [Delete].
4. In **Functions**, double-click **Sum**.
5. Within the parentheses in **Expression**, enter *((Trade – NewFits)**2) / 60*. Click **OK**.
6. Choose **Data > Display Data**.
7. In **Columns, constants, and matrices to display**, enter *MSD*. Click **OK**.

Graph window output



Session window output

Trend Analysis for Trade

```
Time Series Decomposition for RESI1


Data Display


MSD
   11.8989
```

Interpreting the results

In the time series plot, the combined fits from the trend analysis and decomposition (*NewFits*) are close to the actual Trade graphed values (*Trade*).

You can compare fits of different models using MSD. MSD for the quadratic trend model was 59.13. Additive and multiplicative decomposition models with a linear trend (not shown) give MSD values of 20.39 and 18.54, respectively. The MSD value of 11.8989 for the combined quadratic trend and decomposition of residuals indicates a better fit using the additive trend analysis and decomposition models. You might also check the fit to these data of the multiplicative trend analysis and decomposition models.

Example of Decomposition

You wish to predict trade employment for the next 12 months using data collected over 60 months. Because the data have a trend that is fit well by trend analysis' quadratic trend model and possess a seasonal component, you use the residuals from trend analysis example (see Example of a trend analysis) to combine both trend analysis and decomposition for forecasting.

1 Do the trend analysis example.

2 Choose **Stat > Time Series > Decomposition**.

3 In **Variable**, enter the name of the residual column you stored in the trend analysis.

4 In **Seasonal length**, enter *12*.

5 Under **Model Type**, choose **Additive**. Under **Model Components**, choose **Seasonal only**.

6 Check **Generate forecasts** and enter *12* in **Number of forecasts**.

7 Click **Storage**. Check **Forecasts** and **Fits**.

8 Click **OK** in each dialog box.

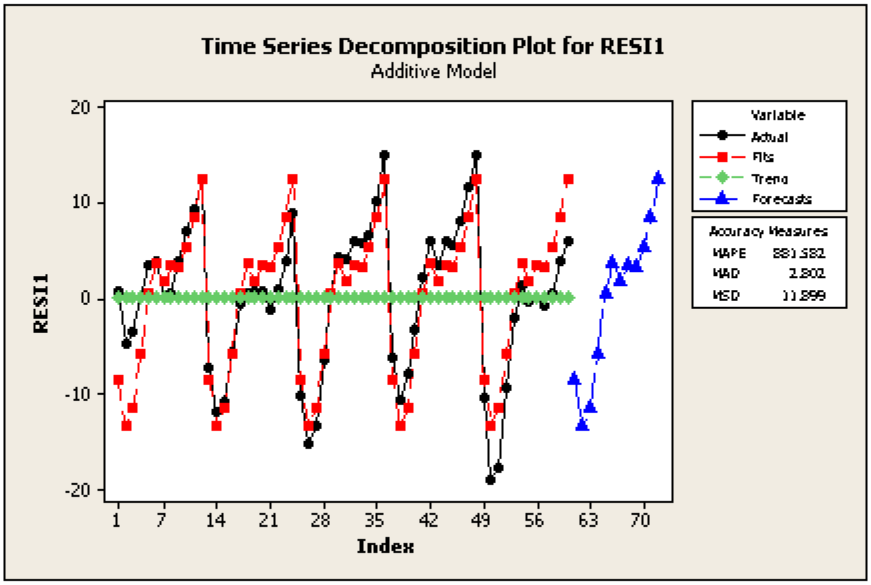
Session window output

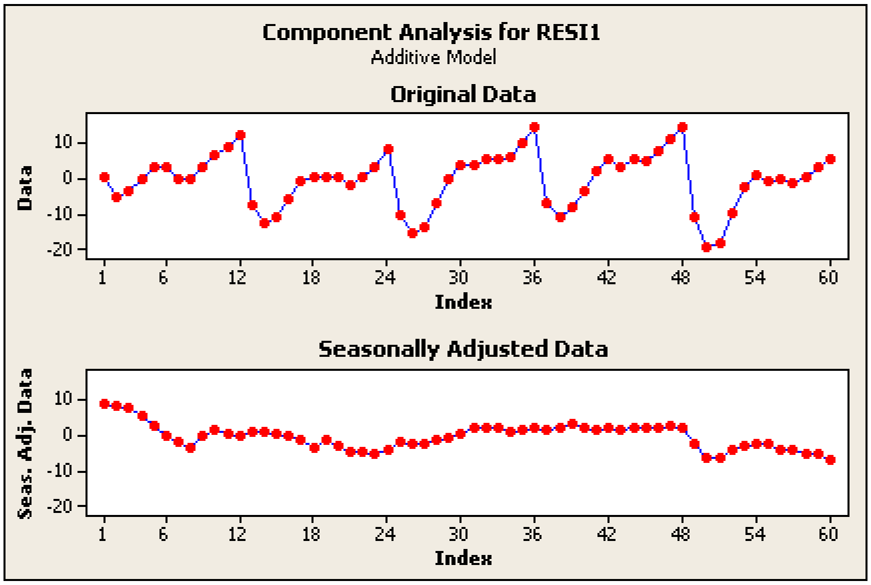
Trend Analysis for Trade

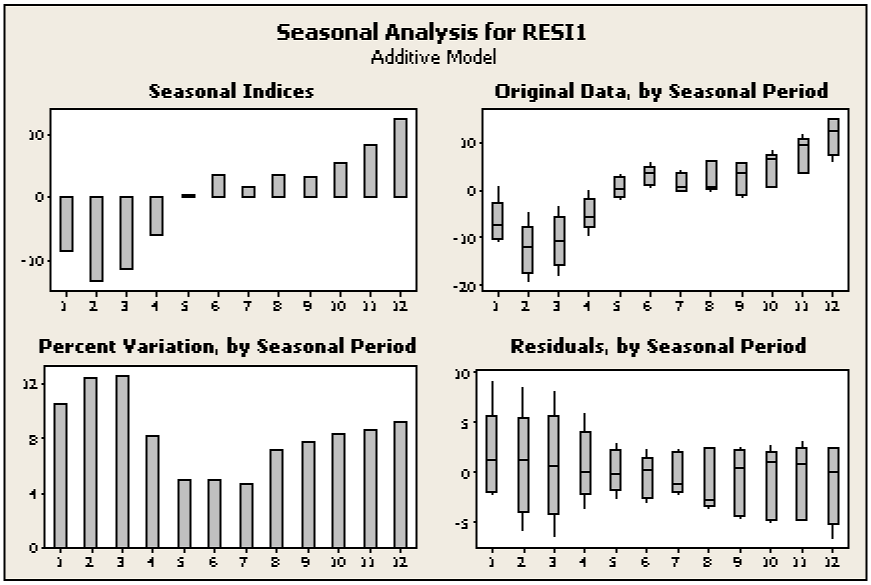
```
Data      Trade
Length    60
NMissing  0


Fitted Trend Equation

Yt = 320.762 + 0.509373*t + 0.0107456*t**2


Accuracy Measures

MAPE   1.7076
MAD    5.9566
MSD   59.1305


Forecasts

Period  Forecast
61       391.818
62       393.649
63       395.502
64       397.376
65       399.271
66       401.188
67       403.127
68       405.087
69       407.068
70       409.071
71       411.096
72       413.142


Time Series Decomposition for RESI1


Additive Model


Data      RESI1
Length    60
NMissing  0


Seasonal Indices

Period     Index
     1   -8.4826

     2  -13.3368
     3  -11.4410
     4   -5.8160
     5    0.5590
     6    3.5590
     7    1.7674
     8    3.4757
     9    3.2674
    10    5.3924
    11    8.4965
    12   12.5590
```

```
Accuracy Measures

MAPE   881.582
MAD      2.802
MSD     11.899


Forecasts

Period  Forecast
61       -8.4826
62      -13.3368
63      -11.4410
64       -5.8160
65        0.5590
66        3.5590
67        1.7674
68        3.4757
69        3.2674
70        5.3924
71        8.4965
72       12.5590
```

Graph window output







Interpreting the results

Decomposition generates three sets of plots:

- A time series plot that shows the original series with the fitted trend line, predicted values, and forecasts.
- A component analysis – in separate plots are the series, the detrended data, the seasonally adjusted data, the seasonally adjusted and detrended data (the residuals).
- A seasonal analysis – charts of seasonal indices and percent variation within each season relative to the sum of variation by season and boxplots of the data and of the residuals by seasonal period.

In addition, Minitab displays the fitted trend line, the seasonal indices, the three accuracy measures– MAPE, MAD, and MSD (see Measures of accuracy) – and forecasts in the Session window.

In the example, the first graph shows that the detrended residuals from trend analysis are fit fairly well by decomposition, except that part of the first annual cycle is underpredicted and the last annual cycle is overpredicted. This is also evident in the lower right plot of the second graph; the residuals are highest in the beginning of the series and lowest at the end.

Moving Average

Moving Average

Stat > Time Series > Moving Average

Moving Average smoothes your data by averaging consecutive observations in a series and provides short-term forecasts. This procedure can be a likely choice when your data do not have a trend or seasonal component. There are ways, however, to use moving averages when your data possess trend and/or seasonality.

Dialog box items

Variable: Enter the column containing the time series.

MA Length: Enter a positive integer to indicate desired length for the moving average. With non-seasonal time series, it is common to use short moving averages to smooth the series, although the length you select may depend on the amount of noise in the series. A longer moving average filters out more noise, but is also less sensitive to changes in the series. With seasonal series, it is common to use a moving average of length equal to the length of an annual cycle.

Center the moving averages: If you check this option, Minitab places the moving average values at the period which is in the center of the range rather than at the end of the range. This is called centering the moving average, and is done to position the moving average values at their central positions in time. Click here for more information on centering.

Generate forecasts: Check to generate forecasts. Forecasts appear in green on the time series plot with 95% prediction interval bands.

> **Number of forecasts:** Enter an integer to indicate how many forecasts that you want.
>
> **Starting from origin:** Enter a positive integer to specify a starting point for the forecasts. For example, if you specify 4 forecasts and 48 for the origin, Minitab computes forecasts for periods 49, 50, 51, and 52, based on the moving average at period 48. If you leave this space blank, Minitab generates forecasts from the end of the data.

<Time>

<Options>

<Storage>

<Graphs>

<Results>

Data – Moving Average

The time series must be in one numeric column. Minitab automatically omits missing data from the calculations.

To perform a Moving Average

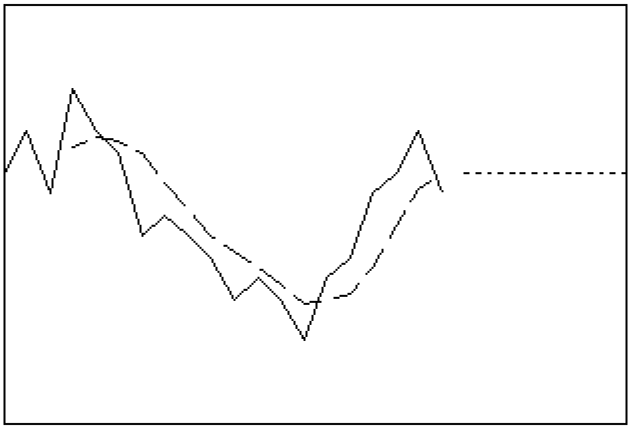
1 Choose **Stat > Time Series > Moving Average**.

2 In **Variable**, enter the column containing the time series.

3 In **MA length**, enter a number to indicate the moving average length. See Determining the moving average length.

4 If you like, use any dialog box options, then click **OK**.

Moving Average : When to Use



Use for:

- Data with no trend, and
- Data with no seasonal pattern
- Short term forecasting

Forecast profile:

- Flat line

ARIMA equivalent: none

Linear Moving Averages Method

You can use this method with a time series that exhibits a trend as well as with moving average schemes involving more than two moving averages. First, compute and store the moving average of the original series. Then compute and store the moving average of the previously stored column to obtain a second moving average.

To compute and store the moving average, choose Stat > Time Series > Moving Average, complete that dialog box, choose Storage, and check the "Moving averages" box.

Determining the Moving Average Length

With non-seasonal time series, it is common to use short moving averages to smooth the series, although the length you select may depend on the amount of noise in the series. A longer moving average filters out more noise, but is also less sensitive to changes in the series. With seasonal series, it is common to use a moving average of length equal to the length of the period. For example, you might choose a moving average length of 12 for monthly data with an annual cycle.

Centering Moving Average Values

By default, moving average values are placed at the period in which they are calculated. For example, for a moving average length of 3, the first numeric moving average value is placed at period 3, the next at period 4, and so on.

When you center the moving averages, they are placed at the center of the range rather than the end of it. This is done to position the moving average values at their central positions in time.

- **If the moving average length is odd:** Suppose the moving average length is 3. In that case, Minitab places the first numeric moving average value at period 2, the next at period 3, and so on. In this case, the moving average value for the first and last periods is missing (∗).
- **If the moving average length is even:** Suppose the moving average length is 4. The center of that range is 2.5, but you cannot place a moving average value at period 2.5. This is how Minitab works around the problem. Calculate the average of the first four values, call it MA1. Calculate the average of the next four values, call it MA2. Average those two numbers (MA1 and MA2), and place that value at period 3. Repeat throughout the series. In this case, the moving average values for the first two and last two periods are missing (∗).

Forecasts – Moving Average

The fitted value at time t is the uncentered moving average at time t -1. The forecasts are the fitted values at the forecast origin. If you forecast 10 time units ahead, the forecasted value for each time will be the fitted value at the origin. Data up to the origin are used for calculating the moving averages.

You can use the linear moving average method by performing consecutive moving averages. This is often done when there is a trend in the data. First, compute and store the moving average of the original series. Then compute and store the moving average of the previously stored column to obtain a second moving average.

In naive forecasting, the forecast for time t is the data value at time t -1. Using moving average procedure with a moving average of length one gives naive forecasting.

See [1], [4], and [6] for a discussion.

Time

Stat > Time Series > *menu command* **> Time**

Specify the time scale.

Dialog box items

Time Scale: Change the x-axis of the scale you specify.

Index: Choose to number the x-axis with a single scale from 1 to n by ones (where n is the number of observations in the column containing the time series).

Calendar: Choose to label the x-axis with different time unit scales. For example, if you choose Month Quarter Year, Minitab assumes that your data are in intervals of 1 month and generates 3 scales on the x-axis: 1 for months, 1 for quarters, and 1 for years. At each 4th tick on the month scale, Minitab generates a tick on the quarter scale. At each 4th tick on the quarter scale, Minitab generates a tick on the Year scale.

Clock: Choose to label the x-axis with different time unit scales. For example, if you choose Day Hour Minute, Minitab assumes that your data are in minute intervals and generates 3 scales on the x-axis: 1 for days, 1 for hours, and 1 for minutes. At each 60th tick on the minute scale, Minitab generates a tick on the hour scale. At each 24th tick on the hour scale, Minitab generates a tick on the Day scale.

Start value(s): Enter the start values.

Increment: Enter a value to increment the time scale.

Stamp: Choose to add a date/time stamp on the x-axis. Enter a date/time column in the text box. The values are stamped on the plot in the same format as they appear in the column.

Moving Average – Options

Stat > Time Series > Moving Average > Options

Specify a customized title.

Dialog box items

Title: To replace the default title with your own custom title, type the desired text in this box.

Moving Average – Storage

Stat > Time Series > Moving Average > Storage

Stores various statistics from the moving average fit in the worksheet.

Dialog box items

Moving averages: Check to store the moving averages, which are averages of consecutive groups of data in a time series.

Fits (one-period-ahead forecasts): Check to store the fitted values. The uncentered moving average at time T is the fitted value for time T+1. You should store the fitted values if you want to generate diagnostic residual plots.

Residuals: Check to store the residuals. The residual at time T is the difference between the actual data at time T and the fitted value at time T. These residuals are used to calculate MAPE, MAD and MSD. If you store the residuals you can generate diagnostic plots using Autocorrelation.

Forecasts: Check to store the forecasts. This option is available only if you generated forecasts in the initial Moving Average dialog box.

Upper 95% Prediction Interval: Check to store the upper 95% prediction limits for the forecasts.

Lower 95% Prediction Interval: Check to store the lower 95% prediction limits for the forecasts.

Moving Average – Graphs

Stat > Time Series > Moving Average > Graphs

Displays time series plot and residual plots for diagnosis of model fit.

Dialog box items

Time series plot (including optional forecasts)

Plot predicted vs. actual: Choose to generate a time series plot which displays the data and one-period-ahead forecasts, or fitted values.

Plot smoothed vs. actual: Choose to generate a time series plot which displays the data and the smoothed, or moving average, values.

Do not display plot: Choose to suppress the time series plot.

Residual Plots

Individual plots: Choose to display one or more plots.

Histogram of residuals: Check to display a histogram of the residuals.

Normal plot of residuals: Check to display a normal probability plot of the residuals.

Residuals versus fits: Check to plot the residuals versus the fitted values.

Residuals versus order: Check to plot the residuals versus the order of the data. The row number for each data point is shown on the x-axis–for example, 1 2 3 4... n.

Four in one: Choose to display a layout of a histogram of residuals, a normal plot of residuals, a plot of residuals versus fits, and a plot of residuals versus order.

Residuals versus the variables: Enter one or more columns containing the variables against which you want to plot the residuals. Minitab displays a separate graph for each column.

Moving Average – Results

Stat > Time Series > Moving Average > Results

Controls Session window output.

Dialog box items

Control the Display of Results

Display nothing: Choose to suppress output.

Summary table: Choose to display the default non-graphical output – summary information about the data, the moving average length, and the accuracy measures.

Summary table and results table: Choose to display the default output plus a table of the original series, moving averages, the predicted values and the errors. If you generate forecasts, they are also listed, with corresponding lower and upper 95% prediction limits.

Example of Moving Average

You wish to predict employment over the next 6 months in a segment of the metals industry using data collected over 60 months. You use the moving average method as there is no well-defined trend or seasonal pattern in the data.

1 Open the worksheet EMPLOY.MTW.

2 Choose **Stat > Time Series > Moving Average**.

3 In **Variable**, enter *Metals*. In **MA length**, enter *3*.

4 Check **Center the moving averages**.

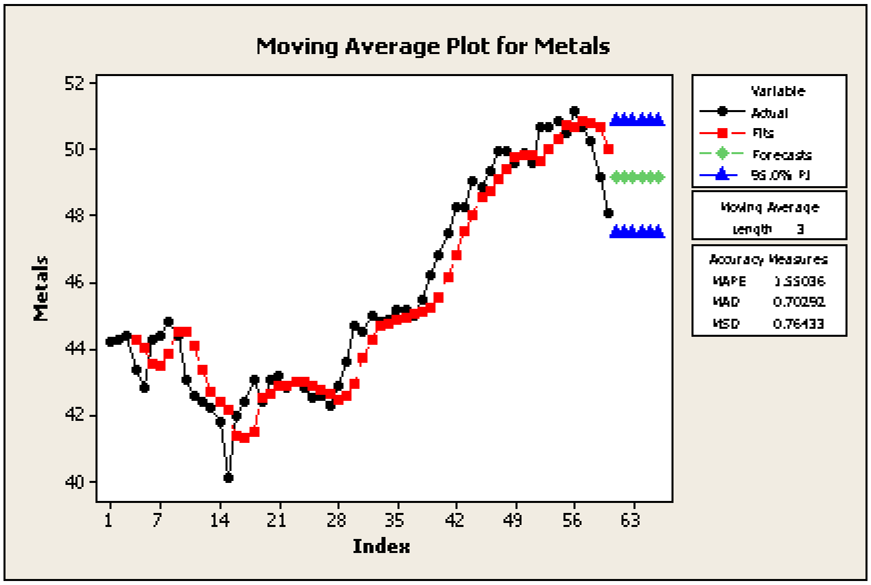
5 Check **Generate forecasts**, and enter *6* in **Number of forecasts**. Click **OK**.

Session window output

Moving Average for Metals

```
Data      Metals
Length    60
NMissing  0


Moving Average

Length  3


Accuracy Measures

MAPE  1.55036
MAD   0.70292
MSD   0.76433


Forecasts

Period  Forecast    Lower    Upper
61          49.2  47.4865  50.9135
62          49.2  47.4865  50.9135
63          49.2  47.4865  50.9135
64          49.2  47.4865  50.9135
65          49.2  47.4865  50.9135
66          49.2  47.4865  50.9135
```

Graph window output



Interpreting the results

Minitab generated the default time series plot which displays the series and fitted values (one-period-ahead forecasts), along with the six forecasts. Notice that the fitted value pattern lags behind the data pattern. This is because the fitted values are the moving averages from the previous time unit. If you wish to visually inspect how moving averages fit your data, plot the smoothed values rather than the predicted values.

To see exponential smoothing methods applied to the same data, see Example of single exponential smoothing and Example of double exponential smoothing.

In the Session window, Minitab displays three measures to help you determine the accuracy of the fitted values: MAPE, MAD, and MSD. See Measures of accuracy. Minitab also displays the forecasts along with the corresponding lower and upper 95% prediction limits.

Single Exponential Smoothing

Single Exponential Smoothing

Stat > Time Series > Single Exp Smoothing

Single exponential smoothing smoothes your data by computing exponentially weighted averages and provides short-term forecasts. This procedure works best for data without a trend or seasonal component. See [1], [4], and [6] for a discussion of exponential smoothing methods.

Dialog box items

Variable: Enter the column containing the time series.

Weight to Use in Smoothing: Use the options below to specify which weight to use. See Computing Weights, or Smoothed Values for details.

> **Optimal ARIMA:** Choose to use the default weight, which Minitab computes by fitting an ARIMA (0, 1, 1) model to the data. With this option, Minitab calculates the initial smoothed value by backcasting.
>
> **Use:** Choose to enter a specific weight, then type a number greater than or equal to 0 and less than 2. With this option, Minitab uses the average of the first six observations (or all the observations if there are less than six observations) for the initial smoothed value by default. You can change this default by specifying a different value in the Single Exponential Smoothing - Options dialog box.

Generate forecasts: Check to generate forecasts. Forecasts appear in green on the time series plot with 95% prediction interval bands.

> **Number of forecasts:** Enter an integer to indicate how many forecasts you want.
>
> **Starting from origin:** Enter a positive integer to specify a starting point for the forecasts. For example, if you specify 4 forecasts and 48 for the origin, Minitab computes forecasts for periods 49, 50, 51, and 52, based on the smoothed value at period 48. If you leave this space blank, Minitab generates forecasts from the end of the data.

<Time>

<Options>

<Storage>

<Graphs>

<Results>

Data – Single Exponential Smoothing

Your time series must be in a numeric column.

The time series *cannot* include any missing values. If you have missing values, you may want to provide estimates of the missing values. If you

- Have seasonal data, estimate the missing values as the fitted values from the decomposition procedure. Replace the missing values in the series with the corresponding fitted values computed by Decomposition.
- Do not have seasonal data, estimate the missing values as the fitted values from the moving average procedure. Replace the missing value with the fitted value computed by Moving Average.

To perform single exponential smoothing

1. Choose **Stat > Time Series > Single Exp Smoothing**.
2. In **Variable**, enter the column containing the time series.
3. If you like, use any dialog box options, then click **OK**.

Computing Weights, or Smoothed Values

The weight is the smoothing parameter. You can have Minitab supply some optimal weight (the default) or you can specify a weight.

Large weights result in more rapid changes in the fitted line; small weights result in less rapid changes in the fitted line. Therefore, the larger the weights, the more the smoothed values follow the data; the smaller the weights, the less jagged the pattern is in the smoothed values. Thus, small weights are usually recommended for a series with a high noise level around the signal or pattern.

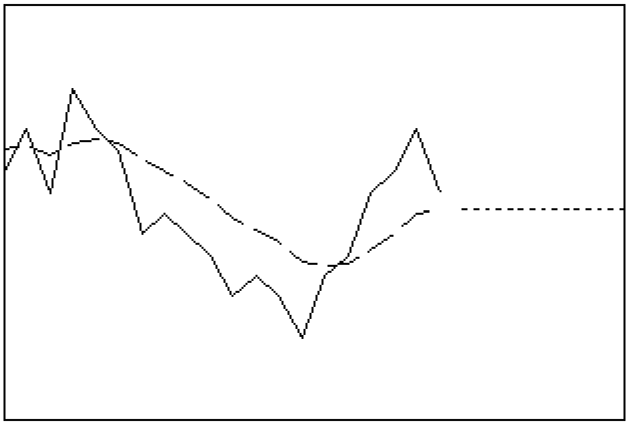
To specify your own weight – Single Exponential Smoothing

In the main Single Exponential Smoothing dialog box, choose **Use** under **Weight to use in smoothing**, and enter a value between 0 and 2, although the usual choices are between 0 and 1.

You can use a rule of thumb for choosing a weight.

- A weight α will give smoothing that is approximately equivalent to an unweighted moving average of length $(2 - \alpha) / \alpha$.
- Conversely, if you want a weight to give a moving average of approximate length l, specify the weight to be 2 / (l +1).

Single Exponential Smoothing : When to Use



Use for:

- Data with no trend, and
- Data with no seasonal pattern
- Short term forecasting

Forecast profile:

- Flat line

ARIMA equivalent: (0,1,1) model

Forecasts – Single Exponential Smoothing

The fitted value at time t is the smoothed value at time t-1. The forecasts are the fitted value at the forecast origin. If you forecast 10 time units ahead, the forecasted value for each time will be the fitted value at the origin. Data up to the origin are used for the smoothing.

In naive forecasting, the forecast for time t is the data value at time t-1. Perform single exponential smoothing with a weight of one to give naive forecasting.

Time

Stat > Time Series > *menu command* **> Time**

Specify the time scale.

Dialog box items

Time Scale: Change the x-axis of the scale you specify.

Index: Choose to number the x-axis with a single scale from 1 to n by ones (where n is the number of observations in the column containing the time series).

Calendar: Choose to label the x-axis with different time unit scales. For example, if you choose Month Quarter Year, Minitab assumes that your data are in intervals of 1 month and generates 3 scales on the x-axis: 1 for months, 1 for quarters, and 1 for years. At each 4th tick on the month scale, Minitab generates a tick on the quarter scale. At each 4th tick on the quarter scale, Minitab generates a tick on the Year scale.

Clock: Choose to label the x-axis with different time unit scales. For example, if you choose Day Hour Minute, Minitab assumes that your data are in minute intervals and generates 3 scales on the x-axis: 1 for days, 1 for hours, and 1 for minutes. At each 60th tick on the minute scale, Minitab generates a tick on the hour scale. At each 24th tick on the hour scale, Minitab generates a tick on the Day scale.

Start value(s): Enter the start values.

Increment: Enter a value to increment the time scale.

Stamp: Choose to add a date/time stamp on the x-axis. Enter a date/time column in the text box. The values are stamped on the plot in the same format as they appear in the column.

Single Exponential Smoothing – Options

Stat > Time Series > Single Exp Smoothing > Options

Specify a customized title and the initial smoothed value.

Dialog box items

Title: To replace the default title with your own custom title, type the desired text in this box.

Set initial smoothed value

Use average of first [ ] observations: By default, if you specify a weight, Minitab uses the average of the first six observations for the initial smoothed value. You can change this default by entering a different value here. If you instructed Minitab to use an optimal weight, this option will not be available since in that case Minitab computes the initial smoothed value by backcasting.

Single Exponential Smoothing – Storage

Stat > Time Series > Single Exp Smoothing > Storage

Stores the smoothed values, the fits or predicted values (smoothed value at time t -1), the residuals (data - fits), forecasts, and upper and lower 95% prediction limits.

Dialog box items

Smoothed data: Check to store the smoothed data. The smoothed value at time T is the fitted value for time T+1.

Fits (one-period-ahead forecasts): Check to store the fitted values. You should store the fitted values if you want to generate diagnostic residual plots.

Residuals: Check to store the residuals. The residual at time T is the difference between the actual data at time T and the fitted value at time T. These residuals are used to calculate MAPE, MAD and MSD. If you store the residuals you can generate diagnostic plots using Autocorrelation.

Forecasts: Check to store the forecasts. This option is available only if you generated forecasts in the main Single Exponential Smoothing dialog box.

Upper 95% Prediction Interval: Check to store the upper 95% prediction limits for the forecasts.

Lower 95% Prediction Interval: Check to store the lower 95% prediction limits for the forecasts.

Single Exponential Smoothing – Graphs

Stat > Time Series > Single Exp Smoothing > Graphs

Displays time series plot and residual plots for diagnosis of model fit.

Dialog box items

Time series plot (including optional forecasts)

Plot predicted vs. actual: Choose to generate a time series plot which displays the data and one-period-ahead forecasts, or fitted values.

Plot smoothed vs. actual: Choose to generate a time series plot which displays the data and the smoothed, or moving average, values.

Do not display plot: Choose to suppress the time series plot.

Residual Plots

Individual plots: Choose to display one or more plots.

Histogram of residuals: Check to display a histogram of the residuals.

Normal plot of residuals: Check to display a normal probability plot of the residuals.

Residuals versus fits: Check to plot the residuals versus the fitted values.

Residuals versus order: Check to plot the residuals versus the order of the data. The row number for each data point is shown on the x-axis–for example, 1 2 3 4... n.

Four in one: Choose to display a layout of a histogram of residuals, a normal plot of residuals, a plot of residuals versus fits, and a plot of residuals versus order.

Residuals versus the variables: Enter one or more columns containing the variables against which you want to plot the residuals. Minitab displays a separate graph for each column.

Single Exponential Smoothing – Results

Stat > Time Series > Single Exp Smoothing > Results

Controls Session window output.

Dialog box items

Control the Display of Results

Display nothing: Choose to suppress output.

Summary table: Choose to display the default non-graphical output – the smoothing constant and accuracy measures.

Summary table and results table: Choose to display the default output plus a table of the original series, smoothed data, the one-period-ahead forecasts (or fitted values), and the one-period-ahead errors. If you generated forecasts, the table also includes the forecasts and their upper and lower 95% prediction limits.

Example of Single Exponential Smoothing

You wish to predict employment over 6 months in a segment of the metals industry using data collected over 60 months. You use single exponential smoothing because there is no clear trend or seasonal pattern in the data.

1 Open the worksheet EMPLOY.MTW.

2 Choose **Stat > Time Series > Single Exp Smoothing**.

3 In **Variable**, enter *Metals*.

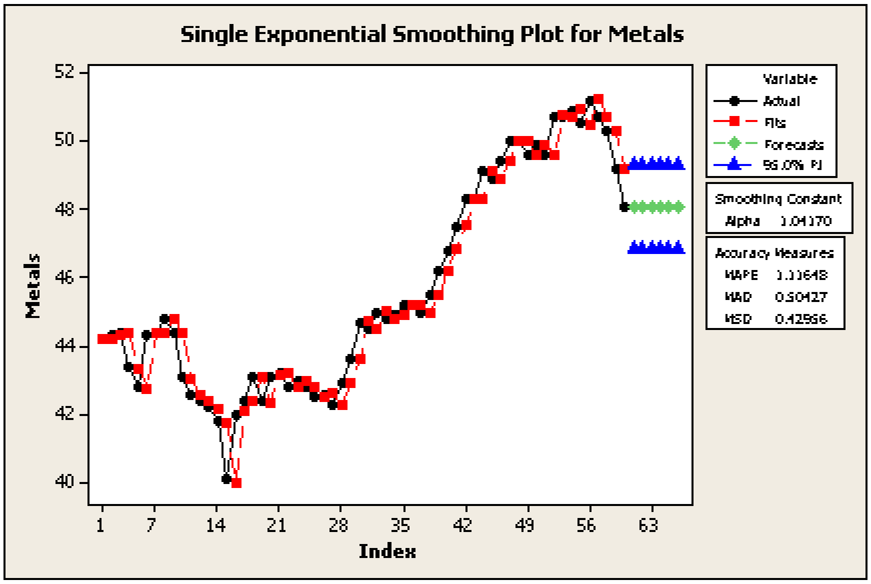
4 Check **Generate forecasts**, and enter *6* in **Number of forecasts**. Click **OK**.

Session window output

Single Exponential Smoothing for Metals

```
Data     Metals
Length   60


Smoothing Constant

Alpha  1.04170


Accuracy Measures

MAPE   1.11648
MAD    0.50427
MSD    0.42956


Forecasts

Period  Forecast    Lower    Upper
61        48.0560  46.8206  49.2914
62        48.0560  46.8206  49.2914
63        48.0560  46.8206  49.2914
64        48.0560  46.8206  49.2914
65        48.0560  46.8206  49.2914
66        48.0560  46.8206  49.2914
```

Graph window output



Interpreting the results

Minitab generated the default time series plot which displays the series and fitted values (one-period-ahead forecasts), along with the six forecasts.

In both the Session and Graph windows, Minitab displays the smoothing constant (weight) used and three measures to help you to determine the accuracy of the fitted values: MAPE, MAD, and MSD (see Measures of accuracy). The three accuracy measures, MAPE, MAD, and MSD, were 1.12, 0.50, and 0.43, respectively for the single exponential smoothing model, compared to 1.55, 0.70, and 0.76, respectively, for the moving average fit (see Example of moving average). Because these values are smaller for single exponential smoothing, you can judge that this method provides a better fit to these data.

Double Exponential Smoothing

Double Exponential Smoothing

Stat > Time Series > Double Exp Smoothing

Double exponential smoothing smoothes your data by Holt (and Brown as a special case) double exponential smoothing and provides short-term forecasts. This procedure can work well when a trend is present but it can also serve as a general smoothing method. Dynamic estimates are calculated for two components: level and trend.

Dialog box items

Variable: Enter the column containing the time series.

Weight to Use in Smoothing: The method Minitab uses to calculate level and trend components is determined by the option chosen below. See Calculating Level and Trend Components for details.

Optimal ARIMA: Choose to use the default weights, or smoothing parameters, which Minitab computes by fitting an ARIMA (0,2,2) model to the data.

Use: Choose to enter specific values for the smoothing parameters. You must specify the appropriate weights greater than 0 and less than 2 for the level component and greater than 0 and less than [4 / (weight for level component) – 2] for the trend component.

Generate forecasts: Check to generate forecasts. Forecasts appear in green on the time series plot with 95% prediction interval bands.

Number of forecasts: Enter an integer to indicate how many forecasts that you want.

Starting from origin: Enter a positive integer to specify a starting point for the forecasts. For example, if you specify 4 forecasts and 48 for the origin, Minitab computes forecasts for periods 49, 50, 51, and 52, based on the level and trend components at period 48. If you leave this space blank, Minitab generates forecasts from the end of the data.

<Time>

<Options>

<Storage>

<Graphs>

<Results>

Data – Double Exponential Smoothing

Your time series must be in a numeric column.

The time series cannot include any missing values. If you have missing values, you may want to provide estimates of the missing values. If you:

- Have seasonal data, estimate the missing values as the fitted values from the decomposition procedure. Replace the missing values in the series with the corresponding fitted values computed by Decomposition.
- Do not have seasonal data, estimate the missing values as the fitted values from the moving average procedure. Replace the missing value with the fitted value computed by Moving Average.

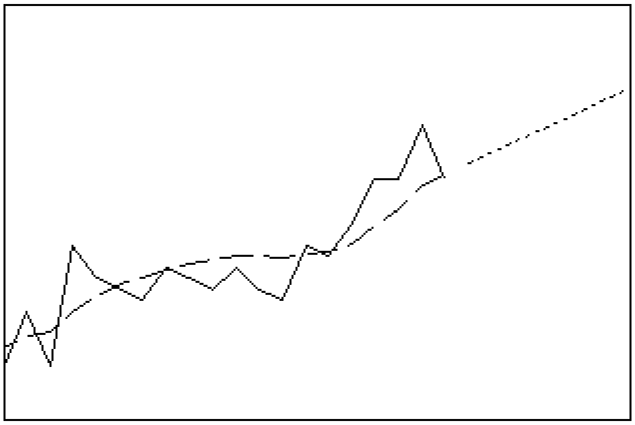
To perform double exponential smoothing

1 Choose **Stat > Time Series > Double Exp Smoothing**.

2 In **Variable**, enter the column containing the time series.

3 If you like, use any dialog box options, then click **OK**.

Double Exponential Smoothing : When to Use



Use for:

- Data with constant or non-constant trend, and
- Data with no seasonal pattern
- Short term forecasting

Forecast profile:

- Straight line with slope equal to last trend estimate

ARIMA equivalent: (0,2,2) model

Level and Trend Components – Double Exponential Smoothing

Double exponential smoothing employs a level component and a trend component at each period. It uses two weights, or smoothing parameters, to update the components at each period. The double exponential smoothing equations are:

$$L_t = \alpha Y_t + (1 - \alpha) [L_{t-1} + T_{t-1}]$$

$$T_t = \gamma [L_t - L_{t-1}] + (1 - \gamma) T_{t-1}$$

$$\hat{Y}_t = L_{t-1} + T_{t-1}$$

where L_t is the level at time t, α is the weight for the level, T_t is the trend at time t, γ is the weight for the trend, Y_t is the data value at time t, and $\hat{Y}_t$ is the fitted value, or one-step-ahead forecast, at time t.

....You must provide the initial level and trend to proceed in one of two ways..... the first observation is numbered one, then level and trend estimates at time zero must be initialized in order to proceed. The initialization method used to determine how the smoothed values are obtained in one of two ways: with Minitab generated weights or with specified weights.

Optimal ARIMA weights

1 Minitab fits an ARIMA (0,2,2) model to the data, in order to minimize the sum of squared errors.

2 The trend and level components are then initialized by backcasting.

Specified weights

1 Minitab fits a linear regression model to time series data (y variable) versus time (x variable).

2 The constant from this regression is the initial estimate of the level component, the slope coefficient is the initial estimate of the trend component.

When you specify weights that correspond to an equal-root ARIMA (0, 2, 2) model, Holt's method specializes to Brown's method. See [4] for more information on Holt's and Brown's methods.

Choosing weights – Double Exponential Smoothing

The weights are the smoothing parameters. You can have Minitab supply some optimal weights (the default) or you can specify weights between 0 and 2 for the level component and between 0 and [4 / weight for the level component) – 2] for the trend component. See Method for more information.

Regardless of the component, large weights result in more rapid changes in that component; small weights result in less rapid changes. Therefore, the larger the weights the more the smoothed values follow the data; the smaller the weights the smoother the pattern in the smoothed values. The components in turn affect the smoothed values and the predicted values. Thus, small weights are usually recommended for a series with a high noise level around the signal or pattern.

To specify your own weights

In the main Double Exponential Smoothing dialog box, choose **Use under Weight** to use in smoothing, and enter a value between 0 and 2 for the level component and between 0 and [4 / weight for the level component) – 2] for the trend component. Values between 0 and 1 are commonly used for both smoothing parameters.

Forecasts – Double Exponential Smoothing

Double exponential smoothing uses the level and trend components to generate forecasts. The forecast for m periods ahead from a point at time t is

$L_t + mT_t$, where L_t is the level and T_t is the trend at time t.

Data up to the forecast origin time will be used for the smoothing.

Time

Stat > Time Series > ***menu command*** **> Time**

Specify the time scale.

Dialog box items

Time Scale: Change the x-axis of the scale you specify.

Index: Choose to number the x-axis with a single scale from 1 to n by ones (where n is the number of observations in the column containing the time series).

Calendar: Choose to label the x-axis with different time unit scales. For example, if you choose Month Quarter Year, Minitab assumes that your data are in intervals of 1 month and generates 3 scales on the x-axis: 1 for months, 1 for quarters, and 1 for years. At each 4th tick on the month scale, Minitab generates a tick on the quarter scale. At each 4th tick on the quarter scale, Minitab generates a tick on the Year scale.

Clock: Choose to label the x-axis with different time unit scales. For example, if you choose Day Hour Minute, Minitab assumes that your data are in minute intervals and generates 3 scales on the x-axis: 1 for days, 1 for hours, and 1 for minutes. At each 60th tick on the minute scale, Minitab generates a tick on the hour scale. At each 24th tick on the hour scale, Minitab generates a tick on the Day scale.

Start value(s): Enter the start values.

Increment: Enter a value to increment the time scale.

Stamp: Choose to add a date/time stamp on the x-axis. Enter a date/time column in the text box. The values are stamped on the plot in the same format as they appear in the column.

Double Exponential Smoothing – Options

Stat > Time Series > Double Exp Smoothing > Options

Specify a customized title.

Dialog box items

Title: To replace the default title with your own custom title, type the desired text in this box.

Double Exponential Smoothing – Storage

Stat > Time Series > Double Exp Smoothing > Storage

Stores various items in the worksheet.

Dialog box items

Smoothed data: Check to store the smoothed data.

Level estimates: Check to store the level components. The level component at time T equals the smoothed value at time T, while adding the level and trend components at time T equals the one-period-ahead forecast for time T+1.

Trend estimates: Check to store the trend components.

Fits (one-period-ahead-forecasts): Check to store the fitted values. You should store the fitted values if you want to generate diagnostic residual plots.

Residuals: Check to store the residuals. The residual at time T is the difference between the actual data at time T and the fitted value at time T. These residuals are used to calculate MAPE, MAD and MSD. If you store the residuals you can generate diagnostic plots using Autocorrelation.

Forecasts: Check to store the forecasts. This option is available only if you generated forecasts in the main Double Exponential Smoothing dialog box.

Upper 95% Prediction Interval: Check to store the upper 95% prediction limits for the forecasts.

Lower 95% Prediction Interval: Check to store the lower 95% prediction limits for the forecasts.

Double Exponential Smoothing – Graphs

Stat > Time Series > Double Exp Smoothing > Graphs

Displays time series plot and residual plots for diagnosis of model fit.

Dialog box items

Time series plot (including optional forecasts)

Plot predicted vs. actual: Choose to generate a time series plot which displays the data and one-period-ahead forecasts, or fitted values.

Plot smoothed vs. actual: Choose to generate a time series plot which displays the data and the smoothed, or moving average, values.

Do not display plot: Choose to suppress the time series plot.

Residual Plots

Individual plots: Choose to display one or more plots.

Histogram of residuals: Check to display a histogram of the residuals.

Normal plot of residuals: Check to display a normal probability plot of the residuals.

Residuals versus fits: Check to plot the residuals versus the fitted values.

Residuals versus order: Check to plot the residuals versus the order of the data. The row number for each data point is shown on the x-axis–for example, 1 2 3 4... n.

Four in one: Choose to display a layout of a histogram of residuals, a normal plot of residuals, a plot of residuals versus fits, and a plot of residuals versus order.

Residuals versus the variables: Enter one or more columns containing the variables against which you want to plot the residuals. Minitab displays a separate graph for each column.

Double Exponential Smoothing – Results

Stat > Time Series > Double Exp Smoothing > Results

Controls Session window output.

Dialog box items

Control the Display of Results

Display nothing: Choose to suppress output.

Summary table: Choose to display the default non-graphical output – the smoothing constants for level and trend, and accuracy measures.

Summary table and results table: Choose to display the default output plus a table containing the original series, the smoothed data, the fitted data and the errors. If you generated forecasts, the table also includes the forecasts and their upper and lower 95% prediction limits.

Example of Double Exponential Smoothing

You wish to predict employment over six months in a segment of the metals industry. You use double exponential smoothing as there is no clear trend or seasonal pattern in the data, and you want to compare the fit by this method with that from single exponential smoothing (see Example of single exponential smoothing).

1 Open the worksheet EMPLOY.MTW.

2 Choose **Stat > Time Series > Double Exp Smoothing**.

3 In **Variable**, enter *Metals*.

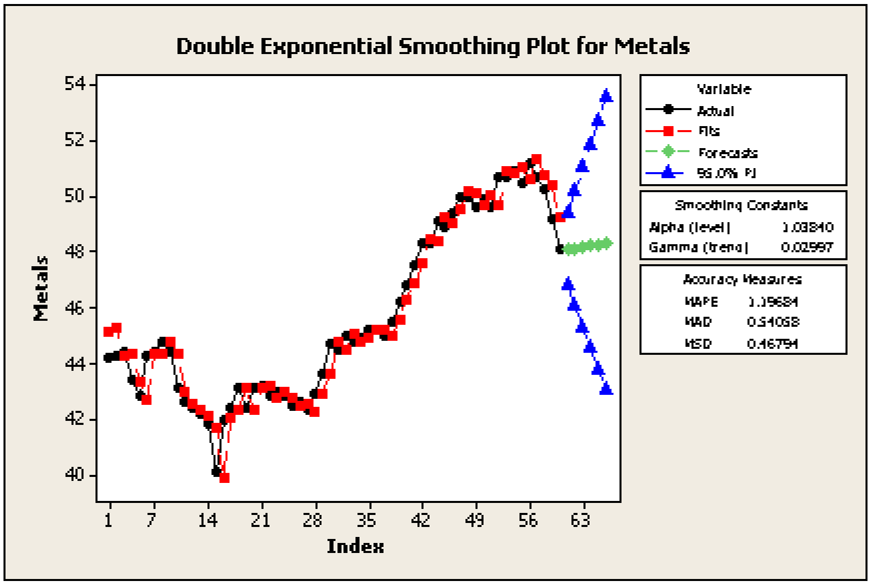
4 Check **Generate forecasts** and enter *6* in **Number of forecasts**. Click **OK**.

Session window output

Double Exponential Smoothing for Metals

```
Data     Metals
Length   60


Smoothing Constants

Alpha (level)   1.03840
Gamma (trend)   0.02997


Accuracy Measures

MAPE   1.19684
MAD    0.54058
MSD    0.46794


Forecasts

Period  Forecast    Lower    Upper
61       48.0961  46.7718  49.4205
62       48.1357  46.0600  50.2113
63       48.1752  45.3135  51.0368
64       48.2147  44.5546  51.8747
65       48.2542  43.7899  52.7184
66       48.2937  43.0221  53.5652
```

Graph window output



Interpreting the results

Minitab generated the default time series plot which displays the series and fitted values (one-step-ahead forecasts), along with the six forecasts.

In both the Session and Graph windows, Minitab displays the smoothing constants (weights) for the level and trend components and three measures to help you determine the accuracy of the fitted values: MAPE, MAD, and MSD (see Measures of accuracy).

The three accuracy measures, MAPE, MAD, and MSD, were respectively 1.19684, 0.54058, and 0.46794 for double exponential smoothing fit, compared to 1.11648, 0.50427, and 0.42956 for the single exponential smoothing fit (see

Example of single exponential smoothing). Because these values are smaller for single exponential smoothing, you can judge that this method provides a slightly better fit to these data.

Because the difference in accuracy measures for the two exponential smoothing methods are small, you might consider the type of forecast (horizontal line versus line with slope) in selecting between methods. Double exponential smoothing forecasts an employment pattern that is slightly increasing though the last four observations are decreasing. A higher weight on the trend component can result in a prediction in the same direction as the data, which may be more realistic, but the measured fit may not be as good as when Minitab generated weights are used.

Winters' Method

Winters' Method

Stat > Time Series > Winters' Method

Winters' Method smoothes your data by Holt-Winters exponential smoothing and provides short to medium-range forecasting. You can use this procedure when both trend and seasonality are present, with these two components being either additive or multiplicative. Winters' Method calculates dynamic estimates for three components: level, trend, and seasonal.

Dialog box items

Variable: Select the column containing the time series.

Seasonal Length: Enter the length of the seasonal pattern. This must be a positive integer greater than or equal to 2.

Model Type:

Multiplicative: Choose the multiplicative model when the seasonal pattern in the data depends on the size of the data. In other words, the magnitude of the seasonal pattern increases as the series goes up, and decreases as the series goes down.

Additive: Choose the additive model when the seasonal pattern in the data does not depend on the size of the data. In other words, the magnitude of the seasonal pattern does not change as the series goes up or down.

Weights to Use in Smoothing: By default, all three weights, or smoothing parameters, are set to 0.2. Since an equivalent ARIMA model exists only for a very restricted form of the Holt-Winters model, optimal parameters are not found for Winters' Method as they are for Single Exponential Smoothing and Double Exponential Smoothing.

Level: Specify the level component weight; must be a number from 0 to 1.

Trend: Specify the trend component weight; must be a number from 0 to 1.

Seasonal: Specify the seasonal component weight; must be a number from 0 to 1.

Generate forecasts: Check to generate forecasts. Forecasts appear in green on the time series plot with 95% prediction interval bands.

Number of forecasts: Enter an integer to indicate how many forecasts you want.

Starting from origin: Enter a positive integer to specify a starting point for the forecasts. For example, if you specify 4 forecasts and 48 for the origin, Minitab computes forecasts for periods 49, 50, 51, and 52, based on the level and trend components at period 48, and the corresponding seasonal components. If you leave this space blank, Minitab generates forecasts from the end of the data.

<Time>

<Options>

<Storage>

<Graphs>

<Results>

Data – Winters' Method

Your time series must be in one numeric column.

The time series *cannot* include any missing values. If you have missing values, you may want to provide estimates of the missing values. If you:

- Have seasonal data, estimate the missing values as the fitted values from the decomposition procedure. Replace the missing values in the series with the corresponding fitted values computed by Decomposition.
- Do not have seasonal data, estimate the missing values as the fitted values from the moving average procedure. Replace the missing value with the fitted value computed by Moving Average.

To perform an exponential smoothing by Winters' Method

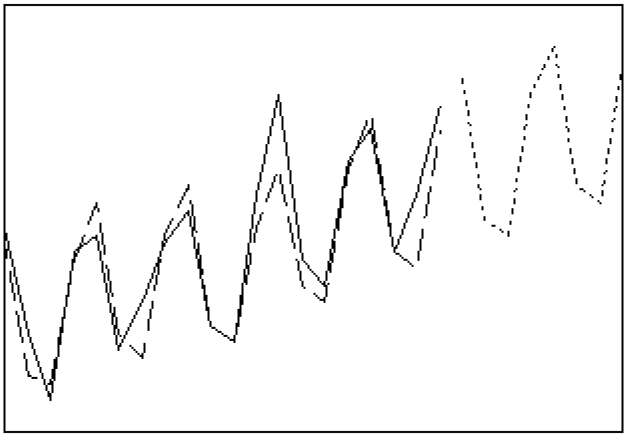
1 Choose **Stat > Time Series > Winters' Method**.

2 In **Variable**, enter the column containing the time series.

3 In **Seasonal length**, enter a number ≥ 2 for the period or seasonal length.

4 If you like, use any dialog box options, then click **OK**.

Winters' Method, Additive Model : When to Use



Use for:

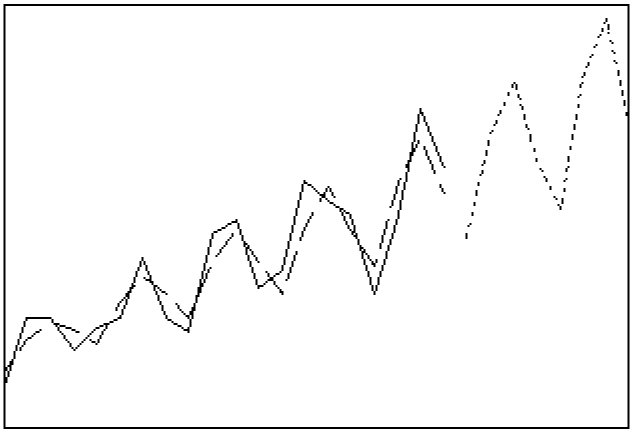
- Data with or without trend
- Data with seasonal pattern
- Size of seasonal pattern not proportional to data
- Short to medium range forecasting

Forecast profile:

- Straight line with seasonal pattern added on

ARIMA equivalent: none

Winters' Method, Multiplicative Model : When to Use



Use for:

- Data with or without trend
- Data with seasonal pattern
- Size of seasonal pattern proportional to data
- Short to medium range forecasting

Forecast profile:

- Straight line multiplied by seasonal pattern

ARIMA equivalent: none

Level, Trend, and Seasonal Components – Winters' Method

Winters' method employs a level component, a trend component, and a seasonal component at each period. It uses three weights, or smoothing parameters, to update the components at each period. Initial values for the level and trend components are obtained from a linear regression on time. Initial values for the seasonal component are obtained from a dummy-variable regression using detrended data. The Winters' method smoothing equations are:

- Additive model:

  $$L_t = \alpha\,(Y_t - S_{t-p}) + (1 - \alpha)\,[L_{t-1} + T_{t-1}]$$

  $$T_t = \gamma\,[L_t - L_{t-1}] + (1 - \gamma)T_{t-1}$$

  $$S_t = \delta\,(Y_t - L_t) + (1 - \delta)\,S_{t-p}$$

  $$\hat{Y}_t = L_{t-1} + T_{t-1} + S_{t-p}$$

- Multiplicative model:

  $$L_t = \alpha\,(Y_t / S_{t-p}) + (1-\alpha)\,[L_{t-1} + T_{t-1}]$$

  $$T_t = \gamma\,[L_t - L_{t-1}] + (1 - \gamma)T_{t-1}$$

  $$S_t = \delta\,(Y_t / L_t) + (1 - \delta)\,S_{t-p}$$

  $$\hat{Y}_t = (L_{t-1} + T_{t-1})\,S_{t-p}$$

  where

  - L_t is the level at time t
  - α is the weight for the level
  - T_t is the trend at time t
  - g is the weight for the trend
  - S_t is the seasonal component at time t
  - d is the weight for the seasonal component
  - p is the seasonal period
  - Y_t is the data value at time t
  - $\hat{Y}_t$ is the fitted value, or one-period-ahead forecast, at time t

Choosing weights – Winters' Method

You can enter weights, or smoothing parameters, for the level, trend, and seasonal components. The default weights are 0.2 and you can enter values between 0 and 1. Since an equivalent ARIMA model exists only for a very restricted form of the Holt-Winters model, Minitab does not compute optimal parameters for Winters' method as it does for single and double exponential smoothing.

Regardless of the component, large weights result in more rapid changes in that component; small weights result in less rapid changes. The components in turn affect the smoothed values and the predicted values. Thus, small weights are usually recommended for a series with a high noise level around the signal or pattern. Large weights are usually recommended for a series with a small noise level around the signal.

Winters' Method – An additive or a multiplicative model?

The Holt-Winters' model is multiplicative when the level and seasonal components are multiplied together and it is additive when they are added together. Choose the multiplicative model when the magnitude of the seasonal pattern in the data depends on the magnitude of the data. In other words, the magnitude of the seasonal pattern increases as the data values increase, and decreases as the data values decrease.

Choose the additive model when the magnitude of the seasonal pattern in the data does not depend on the magnitude of the data. In other words, the magnitude of the seasonal pattern does not change as the series goes up or down.

Forecasts – Winters' Method

Winters' method uses the level, trend, and seasonal components to generate forecasts. The forecast for m periods ahead from a point at time t is:

$L_t + mT_t$

where L_t is the level and T_t is the trend at time t, multiplied by (or added to for an additive model) the seasonal component for the same period from the previous year.

Winters' Method uses data up to the forecast origin time to generate the forecasts.

Time

Stat > Time Series > ***menu command*** **> Time**

Specify the time scale.

Dialog box items

Time Scale: Change the x-axis of the scale you specify.

Index: Choose to number the x-axis with a single scale from 1 to n by ones (where n is the number of observations in the column containing the time series).

Calendar: Choose to label the x-axis with different time unit scales. For example, if you choose Month Quarter Year, Minitab assumes that your data are in intervals of 1 month and generates 3 scales on the x-axis: 1 for months, 1 for quarters, and 1 for years. At each 4th tick on the month scale, Minitab generates a tick on the quarter scale. At each 4th tick on the quarter scale, Minitab generates a tick on the Year scale.

Clock: Choose to label the x-axis with different time unit scales. For example, if you choose Day Hour Minute, Minitab assumes that your data are in minute intervals and generates 3 scales on the x-axis: 1 for days, 1 for hours, and 1 for minutes. At each 60th tick on the minute scale, Minitab generates a tick on the hour scale. At each 24th tick on the hour scale, Minitab generates a tick on the Day scale.

Start value(s): Enter the start values.

Increment: Enter a value to increment the time scale.

Stamp: Choose to add a date/time stamp on the x-axis. Enter a date/time column in the text box. The values are stamped on the plot in the same format as they appear in the column.

Winters' Method – Options

Stat > Time Series > Winters' Method > Options

Specify a customized title and the first observation in the seasonal period.

Dialog box items

Title: To replace the default title with your own custom title, type the desired text in this box.

First obs. is in seasonal period: By default this value is 1 because Minitab assumes that the first data value in the series corresponds to the first seasonal period. Enter a different number to specify a different starting value. For example, if you have monthly data and the first observation is in June, then enter 6 to set the seasonal period correctly.

Winters' Method – Storage

Stat > Time Series > Winters' Method > Storage

Stores various values in the worksheet.

Dialog box items

Smoothed data: Check to store the smoothed data. The smoothed value for time T is equal to the seasonal estimate for time T – p (where p is the seasonal period) multiplied by (or added to for an additive model) the level estimate for time T – 1.

Level estimates: Check to store the level components.

Trend estimates: Check to store the trend components.

Seasonal estimates: Check to store the seasonal components.

Fits (one-period-ahead forecasts): Check to store the fitted values. You should store the fitted values if you want to generate diagnostic residual plots. The fitted value for time T is equal to the sum of the level and trend estimates for time T – 1, multiplied by (or added to for an additive model) the seasonal estimate for time T – p (where p is the seasonal period).

Residuals: Check to store the one-period-ahead forecast errors. These residuals are used to calculate MAPE, MAD and MSD. If you store the residuals you can generate diagnostic plots using Autocorrelation.

Forecasts: Check to store the forecasts. This option is available only if you generated forecasts in the main Winters' Method dialog box.

Upper 95% Prediction Interval: Check to store the upper 95% prediction limits for the forecasts.

Lower 95% Prediction Interval: Check to store the lower 95% prediction limits for the forecasts.

Winters' Method – Graphs

Stat > Time Series > Winters' Method > Graphs

Displays time series plot and residual plots for diagnosis of model fit.

Dialog box items

Time series plot (including optional forecasts)

Plot predicted vs. actual: Choose to generate a time series plot which displays the data and one-period-ahead forecasts, or fitted values.

Plot smoothed vs. actual: Choose to generate a time series plot which displays the data and the smoothed, or moving average, values.

Do not display plot: Choose to suppress the time series plot.

Residual Plots

Individual plots: Choose to display one or more plots.

Histogram of residuals: Check to display a histogram of the residuals.

Normal plot of residuals: Check to display a normal probability plot of the residuals.

Residuals versus fits: Check to plot the residuals versus the fitted values.

Residuals versus order: Check to plot the residuals versus the order of the data. The row number for each data point is shown on the x-axis–for example, 1 2 3 4... n.

Four in one: Choose to display a layout of a histogram of residuals, a normal plot of residuals, a plot of residuals versus fits, and a plot of residuals versus order.

Residuals versus the variables: Enter one or more columns containing the variables against which you want to plot the residuals. Minitab displays a separate graph for each column.

Winters' Method – Results

Stat > Time Series > Winters' Method > Results

Controls Session window output.

Dialog box items

Control the Display of Results

Display nothing: Choose to suppress output.

Summary table: Choose to display the default non-graphical output – the smoothing constants for level and trend, and accuracy measures.

Summary table and results table: Choose to display the default output plus a table containing the original series, the smoothed data, the fitted data and the errors. If you generated forecasts, the table also includes the forecasts and their upper and lower 95% prediction limits.

Example Winters' Method

You wish to predict employment for the next six months in a food preparation industry using data collected over the last 60 months. You use Winters' method with the default multiplicative model, because there is a seasonal component, and possibly trend, apparent in the data.

1 Open the worksheet EMPLOY.MTW.

2 Choose **Stat > Time Series > Winters' Method**.

3 In **Variable**, enter *Food*. In **Seasonal length**, enter *12*.

4 Under **Model Type**, choose **Multiplicative**.

5 Check **Generate forecasts** and enter *6* in **Number of forecasts**. Click **OK**.

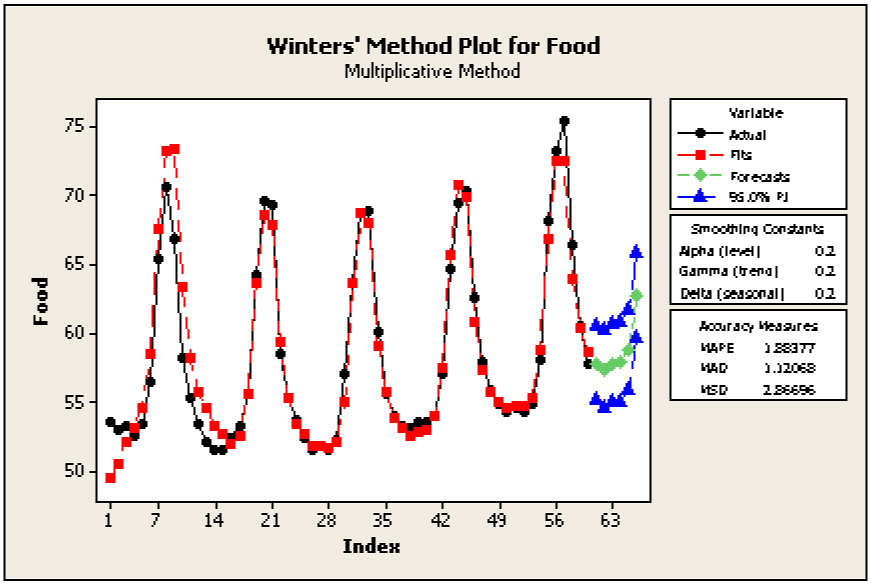
Session window output

Winters' Method for Food

```
Multiplicative Method
```

```
Data     Food
Length   60


Smoothing Constants

Alpha (level)     0.2
Gamma (trend)     0.2
Delta (seasonal)  0.2


Accuracy Measures

MAPE  1.88377
MAD   1.12068
MSD   2.86696


Forecasts

Period  Forecast    Lower    Upper
61       57.8102  55.0646  60.5558
62       57.3892  54.6006  60.1778
63       57.8332  54.9966  60.6698
64       57.9307  55.0414  60.8199
65       58.8311  55.8847  61.7775
66       62.7415  59.7339  65.7492
```

Graph window output



Interpreting the results

Minitab generated the default time series plot which displays the series and fitted values (one-period-ahead forecasts), along with the six forecasts.

In both the Session and Graph windows, Minitab displays the smoothing constants (weights) used for level, trend, and seasonal components used and three measures to help you determine the accuracy of the fitted values: MAPE, MAD, and MSD (see Measures of accuracy).

For these data, MAPE, MAD, and MSD were 1.88, 1.12, and 2.87, respectively, with the multiplicative model. MAPE, MAD, and MSD were 1.95, 1.15, and 2.67, respectively (output not shown) with the additive model, indicating that the multiplicative model provided a slightly better fit according to two of the three accuracy measures.

Differences

Differences

Stat > Time Series > Differences

Differences computes the differences between data values of a time series. If you wish to fit an ARIMA model but there is trend or seasonality present in your data, differencing data is a common step in assessing likely ARIMA models. Differencing is used to simplify the correlation structure and to help reveal any underlying pattern.

Dialog box items

Series: Enter the column containing the variable for which you want to compute differences.

Store differences in: Enter a storage column for the differences.

Lag: Enter the value for the lag. The default lag value is 1.

Data – Differences

Your time series must be in one numeric column. Minitab stores the difference for missing data as missing (*).

To perform Differences

1 Choose **Stat > Time Series > Differences**.

2 In **Series**, enter a column containing the series that you wish to difference.

3 In **Store differences in**, enter a name for the storage column.

4 If you like, change the lag, then click **OK**.

Lag

Lag

Stat > Time Series > Lag

Lag computes lags of a column and stores them in a new column. To lag a time series, Minitab moves the data down the column and inserts missing value symbols, *, at the top of the column. The number of missing values inserted depends upon the length of the lag.

Dialog box items

Series: Enter the column containing the variable that you want to lag.

Store lags in: Enter the storage column for the lags.

Lag: Enter the value for the lag. The default lag value is 1.

Data – Lag

Your time series must be in one numeric column. Minitab stores the lag for missing data with the missing value symbol (*).

To Lag a time series

1 Choose **Stat > Time Series > Lag**.

2 In **Series**, enter a column containing the series that you wish to lag.

3 In **Store lags in**, enter a name for the storage column.

4 If you like, change the value in **Lag** , then click **OK**.

Autocorrelation

Autocorrelation Function

Stat > Time Series > Autocorrelation

Autocorrelation computes and plots the autocorrelations of a time series. Autocorrelation is the correlation between observations of a time series separated by k time units. The plot of autocorrelations is called the autocorrelation function or ACF. View the ACF to guide your choice of terms to include in an ARIMA model. See Fitting an ARIMA model.

Dialog box items

Series: Enter the column containing the time series.

Default number of lags: Choose to use the default number of lags, which is n / 4 for a series with less than or equal to 240 observations or $\sqrt{n}$ + 45 for a series with more than 240 observations, where n is the number of observations in the series.

Number of lags: Choose to enter the number of lags to use instead of the default. The maximum number of lags is n – 1.

Store ACF: Check to store the autocorrelation values in the next available column.

Store t statistics: Check to store the t-statistics.

Store Ljung-Box Q statistics: Check to store the Ljung-Box Q statistics.

Title: Enter a new title to replace the default title on the graphical output.

Data – Autocorrelation

Your time series must be entered in one numeric column. You must either estimate or delete missing data before using this procedure.

To perform an autocorrelation function

1. Choose **Stat > Time Series > Autocorrelation**.
2. In **Series**, enter the column containing the time series.
3. If you like, use any dialog box options, then click **OK**.

Using the Ljung-Box Q statistic

You can use the Ljung-Box Q (LBQ) statistic to test the null hypothesis that the autocorrelations for all lags up to lag k equal zero. This is automatically done for you if you use ARIMA. However, you may find it useful to do this for autocorrelation.

See Example of Testing Autocorrelations for more information.

Example of Autocorrelation

You wish to predict employment in a food preparation industry using past employment data. You want to use ARIMA to do this but first you use autocorrelation in order to help identify a likely model. Because the data exhibit a strong 12 month seasonal component, you take a difference at lag 12 in order to induce stationarity and look at the autocorrelation of the differenced series. There may be some long-term trend in these data, but the magnitude of it appears to be small compared to the seasonal component. If the trend was larger, you might consider taking another difference at lag 1 to induce stationarity.

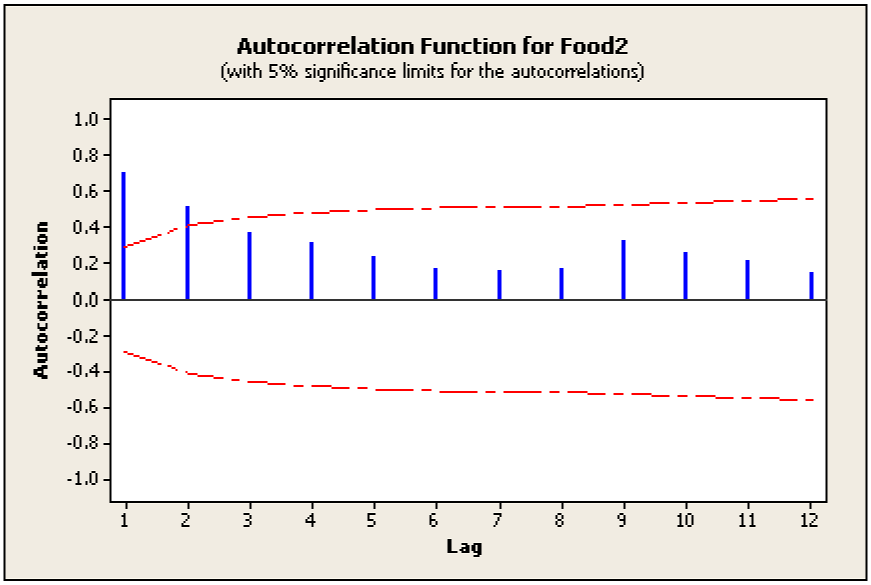
1. Open the worksheet EMPLOY.MTW.
2. Choose **Stat > Time Series > Differences**.
3. In **Series**, enter *Food*.
4. In **Store differences in**, enter *Food2*.
5. In **Lag**, enter *12*. Click **OK**.
6. Choose **Stat > Time Series > Autocorrelation**.
7. In **Series**, enter *Food2*. Click **OK**.

Session window output

Autocorrelation Function: Food2

```
Lag       ACF      T    LBQ
  1  0.701388   4.86  25.12
  2  0.512266   2.52  38.81
  3  0.366882   1.60  45.99
  4  0.310364   1.29  51.24
  5  0.234743   0.94  54.32
  6  0.173069   0.68  56.03
  7  0.162046   0.63  57.57
  8  0.170051  0.66  59.30
  9  0.322438  1.24  65.70
 10  0.252774  0.94  69.74
 11  0.208020  0.76  72.54
 12  0.150936  0.55  74.06


Autocorrelation for Food2
```

Graph window output



Interpreting the results

In the Session window, Minitab displays the autocorrelations, associated t-statistics, and Ljung-Box Q statistics. Because you did not specify the lag length, autocorrelation uses the default length of n / 4 for a series with less than or equal to 240 observations. Minitab generates an autocorrelation function (ACF) with approximate a = 0.05 critical bands for the hypothesis that the correlations are equal to zero.

The ACF for these data shows large positive, significant spikes at lags 1 and 2 with subsequent positive autocorrelations that do not die off quickly. This pattern is typical of an autoregressive process.

See Example of Testing Autocorrelations to test the null hypothesis that the autocorrelations for all lags up to a lag of 6 are zero.

Example of Testing Autocorrelations

Using the results from the Example of Autocorrelation, you can use the Ljung-Box Q (LBQ) statistic to test the null hypothesis that the autocorrelations for all lags up to lag k equal zero. Let's test that all autocorrelations up to a lag of 6 are zero. The LBQ statistic is 56.03.

Step 1: Compute the cumulative probability function

1 Choose **Calc > Probability Distributions > Chi-Square**.

2 Choose **Cumulative Probability**.

3 In **Degrees of freedom**, enter *6* (the lag of your test).

4 Choose **Input constant** and enter *56.03* (the LBQ value).

5 In **Optional storage**, enter *Cumprob*. This stores the cumulative probability function in a constant named *Cumprob*. Click **OK**.

Step 2: Compute the p-value

1 Choose **Calc > Calculator**.

2 In **Store result in variable** enter *pvalue*.

3 In **Expression**, enter *1 - 'Cumprob'*. Click **OK**.

Interpreting the results

Examine the value in the Data window. In this example, the p-value is 0.000000, which means the p-value is less than 0.0000005. The very small p-value implies that one or more of the autocorrelations up to lag 6 can be judged as significantly different from zero at any reasonable a level.

Partial Autocorrelation

Partial Autocorrelation Function

Stat > Time Series > Partial Autocorrelation

Partial autocorrelation computes and plots the partial autocorrelations of a time series. Partial autocorrelations, like autocorrelations, are correlations between sets of ordered data pairs of a time series. As with partial correlations in the regression case, partial autocorrelations measure the strength of relationship with other terms being accounted for. The partial autocorrelation at a lag of k is the correlation between residuals at time t from an autoregressive model and observations at lag k with terms for all intervening lags present in the autoregressive model. The plot of partial autocorrelations is called the partial autocorrelation function or PACF. View the PACF to guide your choice of terms to include in an ARIMA model. See Fitting an ARIMA model.

Dialog box items

Series: Choose the column containing the time series.

Default number of lags: Choose to use the default number of lags. This is n / 4 for a series with less than or equal to 240 observations or $\sqrt{n}$ + 45 for a series with more than 240 observations, where n is the number of observations in the series.

Number of lags: Choose to enter the number of lags to use instead of the default. The maximum number of lags is equal to n – 1.

Store PACF: Check to store the partial autocorrelations in the next available column.

Store t statistics: Check to store the t-statistics.

Title: Enter a new title to replace the default title on the graphical output.

To perform a partial autocorrelation function

1 Choose **Stat > Time Series > Partial Autocorrelation**.

2 In **Series**, enter the column containing the time series.

3 If you like, use any available dialog box options, then click **OK**.

Data – Partial Autocorrelation

Your time series must be entered in one numeric column. You must either estimate or delete missing data before using this procedure.

Example of partial autocorrelation

You obtain a partial autocorrelation function (PACF) of the food industry employment data, after taking a difference of lag 12, in order to help determine a likely ARIMA model.

1 Open the worksheet EMPLOY.MTW.

2 Choose **Stat > Time Series > Differences**.

3 In **Series**, enter *Food*.

4 In **Store differences in**, enter *Food2*.

5 In **Lag**, enter *12*. Click **OK**.

6 Choose **Stat > Time Series > Partial Autocorrelation**.

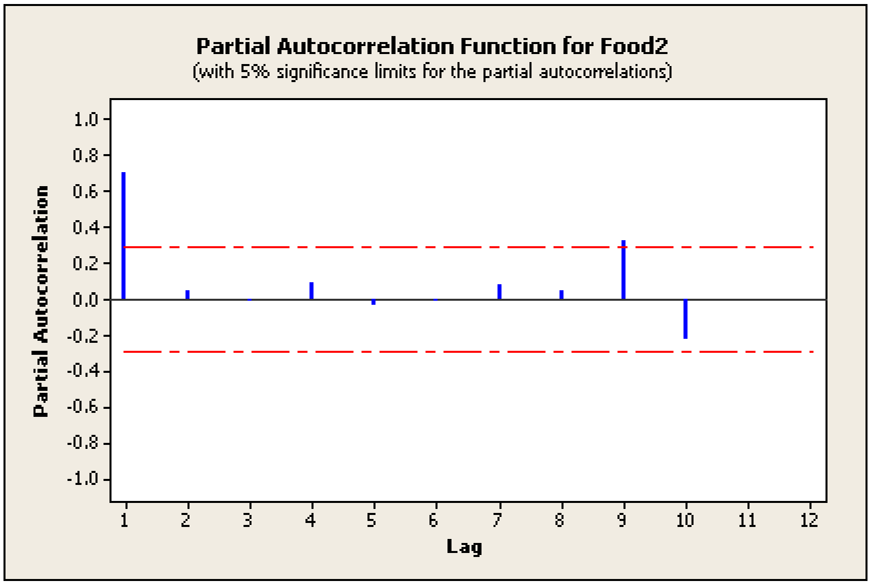
7 In **Series**, enter *Food2*. Click **OK**.

Session window output

Partial Autocorrelation Function: Food2

```
Lag        PACF      T
  1    0.701388   4.86
  2    0.039998   0.28
  3   -0.012022  -0.08
  4    0.092572   0.64
  5   -0.034921  -0.24
  6   -0.014194  -0.10
  7    0.075222   0.52
  8    0.049848   0.35
  9    0.326936   2.27
 10   -0.227678  -1.58
 11    0.005302   0.04
 12   -0.000979  -0.01


Partial Autocorrelation for Food2
```

Graph window output



Interpreting the results

Minitab generates a partial autocorrelation function with critical bands at approximately α = 0.05 for the hypothesis that the correlations are equal to zero. In the Data window, Minitab stores the partial autocorrelations and associated t-statistics.

In the food data example, there is a single large spike of 0.7 at lag 1, which is typical of an autoregressive process of order one. There is also a significant spike at lag 9, but you have no evidence of a nonrandom process occurring there.

Cross Correlation

Cross Correlation Function

Stat > Time Series > Cross Correlation

Computes and plots the cross correlation between two time series.

Dialog box items

First series: Select the column containing the response variable of first time series.

Second series: Select the column containing the response variable second time series.

Default number of lags: Choose to have Minitab set K = $-(\sqrt{n} + 10)$ to K = $+(\sqrt{n} + 10)$, where K is the number of lags and n is the number of observations in the series.

Number of lags: Choose to specify the number of lags desired. Then enter the number of lags.

Data – Cross Correlation

You must have two time series in separate numeric columns of equal length. You must either estimate or delete missing data before using this procedure.

To perform a cross correlation function

1. Choose **Stat > Time Series > Cross Correlation**.
2. In **First Series**, enter the column containing one time series.
3. In **Second Series**, enter the column containing other time series.
4. If you like, specify the number of lags for which to display cross correlations, then click **OK**.

ARIMA

ARIMA

Stat > Time Series > ARIMA

Use ARIMA to model time series behavior and to generate forecasts. ARIMA fits a Box-Jenkins ARIMA model to a time series. ARIMA stands for Autoregressive Integrated Moving Average with each term representing steps taken in the model construction until only random noise remains. ARIMA modeling differs from the other time series methods discussed in this chapter in the fact that ARIMA modeling uses correlational techniques. ARIMA can be used to model patterns that may not be visible in plotted data. The concepts used in this procedure follow Box and Jenkins [2]. For an elementary introduction to time series, see [3], [10]

For information on creating an ARIMA model, see Entering the ARIMA model and ARIMA specifications.

Dialog box items

Series: Enter the column containing the response variable of the time series you want to fit.

Fit seasonal model: Check to fit a seasonal model.

  Period: Specify the number of units in a complete cycle.

Autoregressive

  Nonseasonal: Enter the order of the autoregressive (AR) component (p).

  Seasonal: If you have a seasonal model, enter the order of the seasonal autoregressive component (P) .

Difference

  Nonseasonal: Enter the number of differences (d) used to discount trends over time. At least three data points must remain after differencing.

  Seasonal: If you have a seasonal model, enter the number of differences for the seasonal component (D).

Moving average

  Nonseasonal: Enter the order of the moving average (MA) component (q).

  Seasonal: If you have a seasonal model, enter the order of the seasonal moving average component (Q).

Include constant term in model: Check to include a constant term in the ARIMA model.

Starting values for coefficients: Check to specify the initial parameter values and then enter the column containing the values. The values must be entered in the order that the parameters appear in the output: p (AR values), P (seasonal AR values), q (MA values), Q (seasonal MA values), and then (optionally) the constant. If you do not specify the initial parameter values, Minitab uses 0.1 for the parameters with the exception of the constant.

<Forecasts>

<Graphs>

<Storage>

<Results>

Data – ARIMA

Your time series must be in a numeric column. Missing data in the middle of your series are not allowed. If you have missing values, you may want to provide estimates of the missing values.

To fit an ARIMA model

1 Choose **Stat > Time Series > ARIMA**.

2 In **Series**, enter the column containing the time series.

3 For at least one of **Autoregressive** or **Moving Average** under either **Nonseasonal** or **Seasonal**, enter the number of parameters. See Entering the ARIMA model.

4 If you like, use any dialog box options, then click **OK**.

Entering the ARIMA model

After you have identified one or more likely models, you need to specify the model in the main ARIMA dialog box.

- If you want to fit a seasonal model, check **Fit seasonal model** and enter a number to specify the period. The period is the span of the seasonality or the interval at which the pattern is repeated. The default period is 12.

  You must check **Fit seasonal model** before you can enter the seasonal autoregressive and moving average parameters or the number of seasonal differences to take.

- To specify autoregressive and moving average parameters to include in nonseasonal or seasonal ARIMA models, enter a value from 0 to 5. The maximum is 5. At least one of these parameters must be nonzero. The total for all parameters must not exceed 10. For most data, no more than two autoregressive parameters or two moving average parameters are required in ARIMA models.

  Suppose you enter 2 in the box for **Moving Average** under **Seasonal**, the model will include first and second order moving average terms.

- To specify the number of nonseasonal and/or seasonal differences to take, enter a number in the appropriate box. If you request one seasonal difference with k as the seasonal period, the kth difference will be taken.

- To include the constant in the model, check **Include constant term in model**.

- You may want to specify starting values for the parameter estimates. You must first enter the starting values in a worksheet column in the following order: AR's (autoregressive parameters), seasonal AR's, MA's (moving average parameters), seasonal MA's, and if you checked **Include constant term in model** enter the starting value for the constant in the last row of the column. This is the same order in which the parameters appear on the output. Check **Starting values for coefficients**, and enter the column containing the starting values for each parameter included in the model. Default starting values are 0.1 except for the constant.

Fitting an ARIMA model

Box and Jenkins [2] present an interactive approach for fitting ARIMA models to time series. This iterative approach involves identifying the model, estimating the parameters, checking model adequacy, and forecasting, if desired. The model identification step generally requires judgment from the analyst.

1 First, decide if the data are stationary. That is, do the data possess constant mean and variance.

- Examine a time series plot to see if a transformation is required to give constant variance.
- Examine the ACF to see if large autocorrelations do not die out, indicating that differencing may be required to give a constant mean.

A seasonal pattern that repeats every k^{th} time interval suggests taking the k^{th} difference to remove a portion of the pattern. Most series should not require more than two difference operations or orders. Be careful not to overdifference. If spikes in the ACF die out rapidly, there is no need for further differencing. A sign of an overdifferenced series is the first autocorrelation close to -0.5 and small values elsewhere [10].

Use **Stat > Time Series > Differences** to take and store differences. Then, to examine the ACF and PACF of the differenced series, use **Stat > Time Series > Autocorrelation** and **Stat > Time Series > Partial Autocorrelation**.

2 Next, examine the ACF and PACF of your stationary data in order to identify what autoregressive or moving average models terms are suggested.

- An ACF with large spikes at initial lags that decay to zero or a PACF with a large spike at the first and possibly at the second lag indicates an autoregressive process.
- An ACF with a large spike at the first and possibly at the second lag and a PACF with large spikes at initial lags that decay to zero indicates a moving average process.
- The ACF and the PACF both exhibiting large spikes that gradually die out indicates that both autoregressive and moving averages processes are present.

For most data, no more than two autoregressive parameters or two moving average parameters are required in ARIMA models. See [10] for more details on identifying ARIMA models.

3 Once you have identified one or more likely models, you are ready to use the ARIMA procedure.

- Fit the likely models and examine the significance of parameters and select one model that gives the best fit. See Entering the ARIMA model.
- Check that the ACF and PACF of residuals indicate a random process, signified when there are no large spikes. You can easily obtain an ACF and a PACF of residual using ARIMA's Graphs subdialog box. If large spikes remain, consider changing the model.
- You may perform several iterations in finding the best model. When you are satisfied with the fit, go ahead and make forecasts.

The ARIMA algorithm will perform up to 25 iterations to fit a given model. If the solution does not converge, store the estimated parameters and use them as starting values for a second fit. You can store the estimated parameters and use them as starting values for a subsequent fit as often as necessary.

Diagnostic checking

The graphs for the ACF and PACF of the ARIMA residuals include lines representing two standard errors to either side of zero. Values that extend beyond two standard errors are statistically significant at approximately $\alpha = 0.05$, and show evidence that the model has not explained all autocorrelation in the data. For ACF the distance between the lines and zero for the ith autocorrelation are determined by the following formula:

$$2\sqrt{1+2\sum_{k=1}^{i-1} r_k^2} / \sqrt{n}$$

where n = the number of observations in the series, and Γ_k = the k$_i$ autocorrelation.

The distance between the lines and zero for all partial autocorrelations is $2/\sqrt{n}$.

ARIMA – Graphs

Stat > Time Series > ARIMA > Graphs

Displays a time series plot and various residual plots. You do not have to store the residuals in order to produce these plots.

Dialog box items

Time series plot (including optional forecasts): Check to display a time series plot of the series. When you use the ARIMA - Forecasts subdialog box to generate forecasts, Minitab displays the forecasts and their 95% confidence limits on the plot.

Residual Plots

ACF of residuals: Check to display an autocorrelation function for the standard or raw residuals. The graph includes critical value lines at ± two standard errors.

PACF of residuals: Check to display a partial autocorrelation function for the standard or raw residuals. The graph includes critical value lines at ± two standard errors.

Individual plots: Choose to display one or more plots.

Histogram of residuals: Check to display a histogram of the residuals.

Normal plot of residuals: Check to display a normal probability plot of the residuals.

Residuals versus fits: Check to plot the residuals versus the fitted values.

Residuals versus order: Check to plot the residuals versus the order of the data. The row number for each data point is shown on the x-axis--for example, 1 2 3 4... n.

Four in one: Choose to display a layout of a histogram of residuals, a normal plot of residuals, a plot of residuals versus fits, and a plot of residuals versus order.

Residuals versus the variables: Enter one or more columns containing the variables against which you want to plot the residuals. Minitab displays a separate graph for each column.

ARIMA – Forecasts

Stat > Time Series > ARIMA > Forecasts

Predicts up to 150 values for specified time series.

Dialog box items

Lead: Specify the number of forecasts that you want Minitab to generate.

Origin: Specify the origin for the forecasts to begin. If the origin is not specified, it is set to the end of the series and the forecasts are for the future.

Storage: Controls the storage of ARIMA results.

Forecasts: Specify a storage column for the forecasted values (possibly for later plotting).

Lower limits: Specify a storage column for the lower confidence limits for the forecasts.

Upper limits: Specify a storage column for the upper confidence limits for the forecasts.

ARIMA – Results

Stat > Time Series > ARIMA > Results

Control the display of results to the Session window.

Dialog box items

Control the Display of Results

Display nothing: Display nothing in the Session window.

Table of final estimates, differencing information, residual sums of squares, and number of observations: Display a table of parameter estimates, differencing information, the number of observations, sums of squares statistics, and autocorrelation statistics.

In addition, table of estimates at each iteration (and back forecasts, if they are not dying out rapidly): In addition to the above, display a table of parameter estimates at each iteration and back forecasts if they are not dying out rapidly.

In addition, correlation matrix of estimated parameters: in addition to the above, display a correlation matrix of parameter estimates.

In addition, the back forecasts: In addition to the above, display the back forecasts.

ARIMA – Storage

Stat > Time Series > ARIMA > Storage

Stores residuals, fits, and coefficients in the worksheet.

Dialog box items

Residuals: Check to store the residuals.

Fits: Check to store the fitted values.

Coefficients: Check to store estimated coefficients.

Example of ARIMA

The ACF and PACF of the food employment data (see Example of autocorrelation and Example of partial autocorrelation) suggest an autoregressive model of order 1, or AR(1), after taking a difference of order 12. You fit that model here, examine diagnostic plots, and examine the goodness of fit. To take a seasonal difference of order 12, you specify the seasonal period to be 12, and the order of the difference to be 1. In the subsequent example, you perform forecasting.

1 Open the worksheet EMPLOY.MTW.

2 Choose **Stat > Time Series > ARIMA**.

3 In **Series**, enter *Food*.

4 Check **Fit seasonal model**. In **Period**, enter *12*. Under **Nonseasonal**, enter *1* in **Autoregressive**. Under **Seasonal**, enter *1* in **Difference**.

5 Click **Graphs**. Check **ACF of residuals** and **PACF of residuals**.

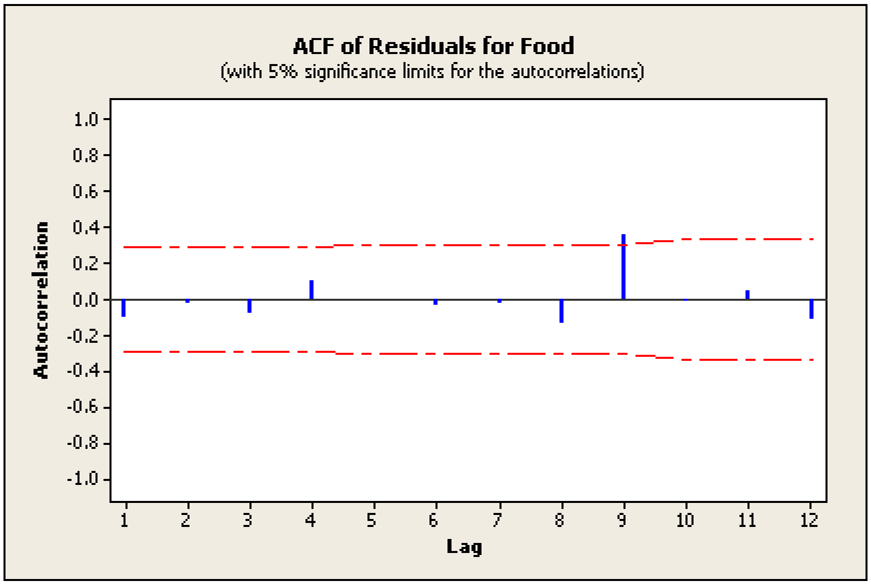
6 Click **OK** in each dialog box.

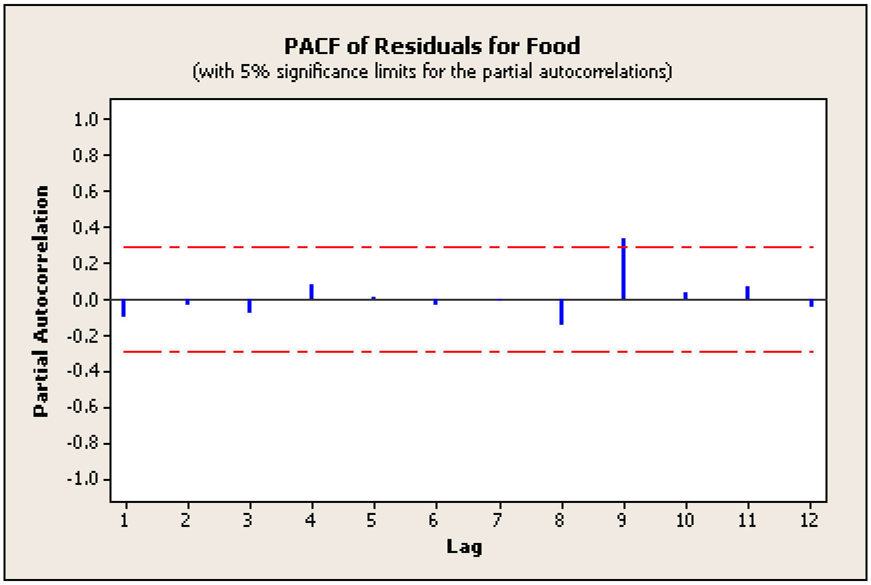
Session window output

ARIMA Model: Food

```
Estimates at each iteration

Iteration      SSE   Parameters
        0  95.2343  0.100  0.847
        1  77.5568  0.250  0.702
        2  64.5317  0.400  0.556
        3  56.1578  0.550  0.410
        4  52.4345  0.700  0.261
        5  52.2226  0.733  0.216
        6  52.2100  0.741  0.203
        7  52.2092  0.743  0.201
        8  52.2092  0.743  0.200
        9  52.2092  0.743  0.200

Relative change in each estimate less than 0.0010


Final Estimates of Parameters

Type         Coef  SE Coef     T      P
AR   1     0.7434   0.1001  7.42  0.000
Constant   0.1996   0.1520  1.31  0.196


Differencing: 0 regular, 1 seasonal of order 12
Number of observations:  Original series 60, after differencing 48
Residuals:    SS =  51.0364 (backforecasts excluded)
              MS =  1.1095  DF = 46


Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36  48
Chi-Square   11.3   19.1   27.7   *
DF             10     22     34   *
P-Value     0.338  0.641  0.768   *
```

Graph window output





Interpreting the results

The ARIMA model converged after nine iterations. The AR(1) parameter had a t-value of 7.42. As a rule of thumb, you can consider values over two as indicating that the associated parameter can be judged as significantly different from zero. The MSE (1.1095) can be used to compare fits of different ARIMA models.

The Ljung-Box statistics give nonsignificant p-values, indicating that the residuals appeared to uncorrelated. The ACF and PACF of the residuals corroborate this. You assume that the spikes in the ACF and PACF at lag 9 are the result of random events. The AR(1) model appears to fit well so you use it to forecast employment in the Example of Forecasting with ARIMA.

Example of Forecasting with ARIMA

In the example of fitting an ARIMA model, you found that an AR(1) model with a twelfth seasonal difference gave a good fit to the food sector employment data. You now use this fit to predict employment for the next 12 months.

Step 1: Refit the ARIMA model without displaying the acf and pacf of the residuals

1 Perform steps 1-4 of Example of ARIMA.

Step 2: Display a time series plot

1 Click Graphs. Check **Time series plot**. Click **OK**.

Step 3: Generate the forecasts

1 Click **Forecast**. In **Lead**, enter *12*. Click **OK** in each dialog box.

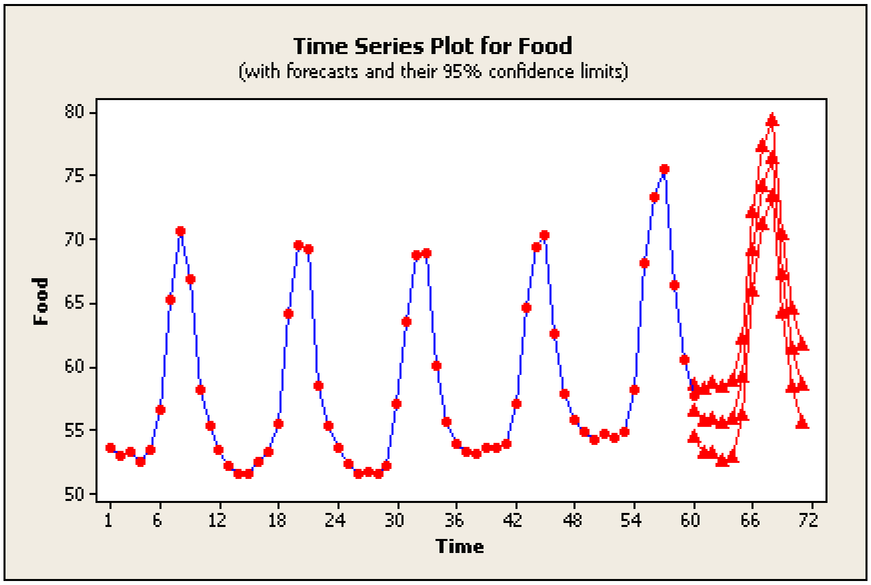
Session window output

ARIMA Model: Food

```
Estimates at each iteration

Iteration      SSE   Parameters
        0  95.2343  0.100  0.847
        1  77.5568  0.250  0.702
        2  64.5317  0.400  0.556
        3  56.1578  0.550  0.410
        4  52.4345  0.700  0.261
        5  52.2226  0.733  0.216
        6  52.2100  0.741  0.203
        7  52.2092  0.743  0.201
        8  52.2092  0.743  0.200
        9  52.2092  0.743  0.200
```

```
Relative change in each estimate less than 0.0010


Final Estimates of Parameters

Type         Coef  SE Coef     T      P
AR   1     0.7434   0.1001  7.42  0.000
Constant   0.1996   0.1520  1.31  0.196


Differencing: 0 regular, 1 seasonal of order 12
Number of observations:  Original series 60, after differencing 48
Residuals:    SS =  51.0364 (backforecasts excluded)
              MS =  1.1095  DF = 46


Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12      24      36  48
Chi-Square   11.3    19.1    27.7   *
DF             10      22      34   *
P-Value     0.338   0.641   0.768   *


Forecasts from period 60

                     95 Percent
                       Limits
Period  Forecast    Lower    Upper  Actual
    61   56.4121  54.3472  58.4770
    62   55.5981  53.0251  58.1711
    63   55.8390  53.0243  58.6537

    64   55.4207  52.4809  58.3605
    65   55.8328  52.8261  58.8394
    66   59.0674  56.0244  62.1104
    67   69.0188  65.9559  72.0817
    68   74.1827  71.1089  77.2565
    69   76.3558  73.2760  79.4357
    70   67.2359  64.1527  70.3191
    71   61.3210  58.2360  64.4060
    72   58.5100  55.4240  61.5960
```

Graph window output



Interpreting the results

ARIMA gives forecasts, with 95% confidence limits, using the AR(1) model in both the Session window and a Graph window. The seasonality dominates the forecast profile for the next 12 months with the forecast values being slightly higher than for the previous 12 months.

Tables

Overview

Tables

Stat > Tables

Tally Individual Variables – prints one-way tables of counts and percents from raw data

Cross Tabulation and Chi-Square – generates tables containing count statistics. The chi-square option tests dependence among characteristics in a two-way classification.

Chi-Square Test – does a χ^2 analysis of a contingency table

Descriptive Statistics – displays one-way, two-way, and multi-way tables containing descriptive statistic summaries of data for categorical variables and associated variables.

Tables Overview

The following table procedures summarize data into tables and perform analyses on the tabled data:

- Tally Individual Variables displays a table of counts, cumulative counts, percents, and cumulative percents for each specified variable.
- Cross Tabulation and Chi-Square displays one-way, two-way, and multi-way tables containing count data. The chi-square option tests dependence among characteristics in a two-way classification. Use this procedure to test if the probabilities of items or subjects being classified for one variable depend upon the classification of the other variable. To use this procedure, your data must be in raw form or frequency form.
- Chi-Square Test (Table in Worksheet) - tests for dependence among characteristics in a two-way classification. Your data must be in contingency table form.
- Descriptive Statistics -displays one-way, two-way, and multi-way tables containing descriptive statistic summaries of data for categorical variables and associated variables.

Acknowledgment

We are grateful for the collaboration of James R. Allen of Allen Data Systems, Cross Plains, Wisconsin in the development of the cross tabulation procedure.

Table Examples

These examples show how to use Minitab's table commands:

Tally Individual Variables

Cross Tabulation and Chi-square: with Three Categorical Variables

Cross Tabulation and Chi-square: Chi-Square Analysis

Cross Tabulation and Chi-square: with Missing Data

Cross Tabulation and Chi-square: Changing Table Layout

Cross Tabulation and Chi-square: Measures of Association for Ordinal Data

Cross Tabulation and Chi-square: Mantel-Haenszel-Cochran Test

Chi-Square Test

Descriptive Statistics: Display Data in a Table

Descriptive Statistics: with Additional Statistical Summaries

Descriptive Statistics: with Missing Data

Descriptive Statistics: with Categorical and Associated Variable Summaries

Arrangement of Input Data

Data used for input to the Tables procedures can be arranged in four different ways in your worksheet:

Raw data – one row for each observation:

C1	C2
1	1
2	1
2	3
1	2

C1= gender, C2 = politics

Contingency table – each cell contains counts:

C1	C2
17	18
10	19
19	17

C1= males, C2 = females

Rows 1-3 represent the three levels for politics

Frequency data – each row represents a unique combination of group codes:

C1	C2	C3
1	1	17
1	2	10
1	3	19
2	1	18

C1= gender, C2 = politics

Indicator variables – one row for each observation

C1	C2	C3	C4	C5
1	0	1	0	0
0	1	1	0	0
0	1	0	0	1
1	0	0	1	0

C1 = 1 if male, 0 if female

C2 = 1 if female, 0 if male

C3 = 1 if Democrat, 0 = otherwise

C4 = 1 if Republican, 0 = otherwise

C5 = 1 if Other, 0 = otherwise

To create:

- frequency data from raw data, see Store Descriptive Statistics.
- contingency table from raw data or frequency data, see Cross Tabulation and Chi-Square, then copy and paste the table output into your worksheet.
- indicator variables from raw data, see Calc > Make Indicator Variables.

Tally Individual Variables

Tally Individual Variables

Stat > Tables > Tally Individual Variables

Use to display counts, cumulative counts, percents, and cumulative percents for each specified variable.

Dialog box items

Variables: Enter the columns for tallying.

Display:

Counts: Check to display the number of times each distinct value occurs.

Percents: Check to display the percent contribution for each category.

Cumulative counts: Check to display cumulative counts for each category.

Cumulative percents: Check to display cumulative percent for each category.

If you do not choose a percent option, Minitab displays counts. If you choose a percent option and want counts, you must check **Counts**.

Data – Tally Individual Variables

Your data should be in the following format:

- Data must be arranged as columns of raw data. See Arrangement of Input Data.
- Data may be numeric, text, or date/time.
- Column lengths may vary.

By default, Minitab excludes missing data from all calculations.

To change the order in which text categories are processed from their default alphabetized order, see Ordering Text Categories.

To tally

1 Choose **Stat > Tables > Tally Individual Variables**.

2 In **Variables**, enter the columns for tallying. click **OK**.

Example of tallying individual variables

Suppose you are studying the influence of patient activity on the performance of a new drug. After collecting your data, you would like to examine the distribution of patient activity.

1 Open the worksheet EXH_TABL.MTW. If you have not already done so, set the value order for the variable Activity.

2 Choose **Stat > Tables > Tally Individual Variables**.

3 In **Variables**, enter *Activity*.

4 Under **Display**, check **Counts, Percents, Cumulative counts**, and **Cumulative percents**. Click **OK**.

Session window output

Tally for Discrete Variables: Activity

```
Activity  Count  CumCnt  Percent  CumPct
  Slight      9       9     9.89    9.89
Moderate     61      70    67.03   76.92
   A lot     21      91    23.08  100.00
      N=     91
```

Interpreting the results

You have 9 subjects with slight activity level, representing 9.89% of all the individuals in the study. You have 61 subjects with moderate activity level.

The cumulative count of 70 represents individuals in the slight and moderate activity category The cumulative percentage indicates that 76.92% (70/91) of the subjects in your study had slight or moderate activity levels. The cumulative percentage sums to 100% (9.89+67.03+23.08).

Cross Tabulation and Chi-Square

Cross Tabulation and Chi-Square

Stat > Tables > Cross Tabulation and Chi-Square

Use to generate tables of counts and percents. You can also perform a chi-square test and choose your table layout.

Dialog box items

Categorical Variables

For rows: Enter the columns containing the categories that define the rows of the table.

For columns: Enter up to two columns containing the categories that define the columns of the table.

For layers: Enter the columns containing the categories that define the layers of the two-way tables.

Frequencies are in: If you have frequency data, enter the column containing the frequencies.

Display

Counts: Check to display the total number of values in each cell and for the margins.

Row percents: Check to display the percentage each cell represents of the total observations in the table row.

Column percents: Check to display the percentage each cell represents of the total observations in the table column.

Total percents: Check to display the percentage each cell represents of all the observations in the table.

<Chi-Square>

<Other Stats>

<Options>

Data – Cross Tabulation and Chi-Square

To use Cross Tabulation and Chi-Square, your data should be in the following format:

- Data may be arranged in raw form or frequency form. See Arrangement of Input Data.
- Data may be numeric, text, or date/time. Frequency data must be non-negative integers.
- Column lengths must be the same.

By default, Minitab excludes missing data from all calculations. See Cross Tabulation Options.

To change the order in which text categories are processed from their default alphabetized order, see Ordering Text Categories.

Note For the Chi-square test, data must be in raw or frequency format. If data are in contingency table form, use Chi-square (Table in worksheet).

Arrangement of Input Data

Data used for input to the Tables procedures can be arranged in four different ways in your worksheet:

Raw data – one row for each observation:

C1	C2
1	1
2	1
2	3
1	2

C1= gender, C2 = politics

Contingency table – each cell contains counts:

C1	C2
17	18
10	19
19	17

C1= males, C2 = females

Rows 1-3 represent the three levels for politics

Frequency data – each row represents a unique combination of group codes:

C1	C2	C3
1	1	17
1	2	10
1	3	19
2	1	18

C1= gender, C2 = politics

Indicator variables – one row for each observation

C1	C2	C3	C4	C5
1	0	1	0	0
0	1	1	0	0
0	1	0	0	1
1	0	0	1	0

C1 = 1 if male, 0 if female

C2 = 1 if female, 0 if male

C3 = 1 if Democrat, 0 = otherwise

C4 = 1 if Republican, 0 = otherwise

C5 = 1 if Other, 0 = otherwise

To create:

- frequency data from raw data, see Store Descriptive Statistics.
- contingency table from raw data or frequency data, see Cross Tabulation and Chi-Square, then copy and paste the table output into your worksheet.
- indicator variables from raw data, see Calc > Make Indicator Variables.

To cross tabulate data

1 Choose **Stat > Tables > Cross Tabulation and Chi-Square**.

2 Do one or more of the following, depending on your desired table layout:

- In **For rows**, enter the columns containing the categories that define the rows of the table.
- In **For columns**, enter the up to two columns containing the categories that define the columns of the table.
- In **For layers**, enter the columns containing the categories that define the layers of the two-way tables.

3 If you have frequency data, check **Frequencies are in** and enter the column containing the frequencies.

4 If you like, use any of the dialog box options, then click **OK**.

To perform a chi-square test with raw data or frequency data

1 Choose **Stat > Tables > Cross Tabulation and Chi-Square**.

2 To input your data, do one of the following:

- For raw data, enter the columns in **For rows** and **For columns**.
- For frequency or collapsed data,
  - enter the columns in **For rows** and **For columns**
  - check **Frequencies are in** and enter the column containing the frequencies

3 Click **Chi-Square**.

4 If you like, use any dialog box options, then click **OK** in each dialog box.

Note You must use a two-way table layout for this analysis.

Table Layout

Table layout is dependent upon:

- The number of categorical variables
- Designation of categorical variables into rows and columns
- Layering of categorical variables

Each table can have up to ten categorical variables, and you can choose which variables will be arranged into rows or columns. You can organize variables to emphasize a relationship, or create a more compact display for a report.

Table layout	Example
One-way table one categorical variable	Rows: Gender Count Female 35 Male 56 All 91
Two-way table two categorical variables	Rows: Gender Columns: Smokes No Yes All Female 27 8 35 Male 37 19 56 All 64 27 91 Cell Contents: Count
Multi-way table three to ten categorical variables	Rows: Gender / Smokes Columns: Activity A lot Moderate Slight All Female No 4 20 3 27 Yes 1 6 1 8 Male No 12 22 3 37 Yes 4 13 2 19 All All 21 61 9 91 Cell Contents: Count

Layering Variables
each level of the layering variable is a separate two-way table

```
Results for Activity = A lot
Rows: Gender   Columns: Smokes
           No  Yes  All
Female    4     1     5
Male     12     4    16
All      16     5    21
Cell Contents:     Count

Results for Activity = Moderate
Rows: Gender   Columns: Smokes
            No  Yes  All
Female    20     6    26
Male      22    13    35
All       42    19    61
Cell Contents:     Count
```

Note To perform a χ^2 test for association, you must use a two-way table layout.

Cross Tabulation – Chi-Square

Stat > Tables > Cross Tabulation and Chi-Square > Chi-square

You can performs a chi-square test of association between variables when your data are in raw or frequency form. To perform this χ^2 test for association, you must use a two-way table layout.

When you enter...	Minitab tabulates...
one row variable one column variable	a two-way table and performs a χ^2 test for association
one row variable one column variable one or more layer variables	multiple two-way tables and performs a χ^2 test for association on each table

Note If your data are in table form, use Chi-square (Table in worksheet).

Dialog box items

Display

Chi-square Analysis: Check to perform a χ^2 test for association.

Expected cell counts: Check to display each cell's expected count.

Residuals: Check to display the difference between the observed and the expected count in each cell.

Standardized residuals: Check to display the standardized residuals of each cell.

Adjusted residuals: Check to display the adjusted residuals of each cell.

Each cell's contribution to the chi-square statistic: Check to display each cell's contribution to the overall chi-square statistic. The contribution to the chi-square statistic is the (standardized residual)^2 for each cell.

Cross Tabulation and Chi-Square – Other Stats

Stat > Tables > Cross Tabulation and Chi-Square > Other Stats

You can perform tests of independence and control the display of statistics for measures of association.

Dialog box items

Tests

Fisher's exact test for 2x2 tables: Check to perform Fisher's exact test.

Mantel-Haenszel-Cochran test for multiple 2x2 tables: Check to perform Mantel–Haenszel–Cochran test.

Other Measures of Association

Cramer's V-square statistic: Check to display Cramer's V–square statistics.

Kappa for inter-rater reliability: Check to display Kappa statistics for inter-rater reliability.

Goodman-Kruskal lambda and tau: Check to display Goodman–Kruskal's lambda and tau statistics for each dependent variable.

Measures of concordance for ordinal categories Check to display measures of concordance and discordant pairs, gamma, Somer's d, and Kendall's tau–b statistics for the ordinal categories.

Correlation coefficients for ordinal categories: Check to display Pearson's and Spearman's correlation coefficients.

Cross Tabulation and Chi-Square – Options

Stat > Tables > Cross Tabulation and Chi-Square > Options

You can control the display of marginal statistics and include missing data into table display and calculations.

Dialog box items

Display marginal statistics for

Rows and columns: Choose to display marginal statistics for all categorical variables. By default, Minitab displays marginal statistics for all categorical variables.

No variables: Choose to exclude marginal statistics from the display.

Selected variables: Choose to include marginal statistics for specific categorical variables. Enter the columns.

Display missing values for

All variables: Choose to display missing values for all variables. By default, Minitab displays missing values for all categorical variables.

No variables: Choose to suppress missing values.

Selected variables: Choose to include missing values in the display for specific categorical variables, then enter the columns to display.

Include displayed missing values in calculations Check to include missing values in calculations. Minitab uses the displayed missing values that you selected above in calculations, so make sure that **Display Missing Values** is selected appropriately. By default, Minitab does not include missing values in calculations.

To set the value order:

1 Click on any cell in the variable **Activity** of the worksheet.

2 Choose **Editor > Column > Value Order**.

3 Choose **User-specified order**.

4 In the right box change **A lot**, **Moderate**, **Slight** to **Slight**, **Moderate**, **A lot**. You can use the right-mouse button to cut and paste. Click **OK**.

Example of Measures of Association for Ordinal Data

You want to assess the relationship between Gender and activity level. Two samples, 35 females and 56 males were taken and asked to evaluate their activity level as: Slight, Moderate, and A lot. You choose to obtain measures of association for ordinal data.

1 Open the worksheet EXH_TABL.MTW. If you have not already done so, set the value order for the variable Activity.

2 Choose **Stat > Tables > Cross Tabulation and Chi-Square**.

3 In **For rows**, enter *Activity*.

4 In **For columns**, enter *Gender*.

5 Under **Display**, check **Counts**.

6 Click **Other Stats**.

7 Under **Other Measures of Association**, check **Measures of concordance for ordinal categories** and **Correlation coefficient for ordinal categories**.

8 Click **OK** in each dialog box.

Session window output

Tabulated statistics: Activity, Gender

```
Rows: Activity   Columns: Gender

            Female  Male  All

Slight           4     5    9
Moderate        26    35   61
A lot            5    16   21
All             35    56   91

Cell Contents:      Count


Pearson's r     0.146135
Spearman's rho  0.150582


Measures of Concordance for Ordinal Categories

Pairs       Number  Summary Measures
Concordant     620  Somers' D (Activity dependent)  0.147959
Discordant     330  Somers' D (Gender dependent)    0.143635
Ties          3145  Goodman and Kruskal's Gamma     0.305263
Total         4095  Kendall's Tau-b                 0.145781

Test of Concordance: P-Value = 0.0754352
```

Interpreting the results

The cells in the table contain the counts for Activity (row) and Gender (column). You can draw the following conclusions:

- Pearson's r (0.146135) and Spearman's rho (0.150582) values suggests a weak association between Gender and Activity.
- Somers' D (Activity dependent) is 0.147959 and (Gender dependent) 0.143635 suggests that there are more discordant pairs than concordant pairs, excluding ties on the independent variables.
- Kendall's tau-b (0.145781) and Goodman-Kruskal gamma (0.305263), suggesting that there are more discordant pairs than concordant pairs.

The null hypothesis for the test of concordance is that the probability of concordance equals the probability of discordance. Assuming α-level as 0.05 for the test, you can conclude that there is not enough evidence suggesting that women are less active than men.

Example of Cross Tabulation with Three Categorical Variables

You would like to classify the individuals in your study by gender, smoking status and weight as the associated variable. To present the same information in a three-way table, see Example of changing table layout.

1 Open the worksheet EXH_TABL.MTW. If you have not already done so, set the value order for the variable Activity.

2 Choose **Stat > Tables > Cross Tabulation and Chi-Square**.

3 In **For rows**, enter *Gender*. In **For columns**, enter **Activity**. In **For layers**, enter **Smokes**.

4 Under **Display**, check **Counts**. Click **OK**.

Session window output

Tabulated statistics: Gender, Activity, Smokes

```
Results for Smokes = No


Rows: Gender   Columns: Activity

          Slight  Moderate  A lot  All

Female         3        20       4   27
Male           3        22      12   37
All            6        42      16   64

Cell Contents:      Count


Results for Smokes = Yes


Rows: Gender   Columns: Activity

          Slight  Moderate  A lot  All

Female         1         6       1    8
Male           2        13       4   19
All            3        19       5   27

Cell Contents:      Count
```

Interpreting the results

In this layout, Minitab creates a two-way table for each level of the layering variable, Smokes. The row variable is Gender and the column variable is Activity. It may be easier for you to compare your data in a different layout. You can change your table layout by designating variables to be across rows, down columns, or as layers.

Example of Cross Tabulation with Missing Data

To classify the individuals in your study by gender and eye color. You create a two-way table.

1 Open the worksheet EYECOLOR.MTW.

2 Choose **Stat > Tables > Cross Tabulation and Chi-Square**.

3 In **For rows**, enter *Gender*. In **For columns**, enter *Eyecolor*.

4 Under **Display**, check **Counts, Row percents, Column percents, and Total percents**. Click **OK**.

Session window output

Tabulated statistics: Gender, Eyecolor

```
Rows: Gender   Columns: Eyecolor

         Blue   Brown   Green   Hazel  Missing     All

F           7       5       7       5        1      24
        29.17   20.83   29.17   20.83        *  100.00
        46.67   38.46   63.64   62.50        *   51.06
        14.89   10.64   14.89   10.64        *   51.06

M           8       8       4       3        2      23
        34.78   34.78   17.39   13.04        *  100.00
        53.33   61.54   36.36   37.50        *   48.94
        17.02   17.02    8.51    6.38        *   48.94

All        15      13      11       8        *      47
        31.91   27.66   23.40   17.02        *  100.00
       100.00  100.00  100.00  100.00        *  100.00
```

```
        31.91   27.66   23.40   17.02        *  100.00

Cell Contents:      Count
                    % of Row
                    % of Column
                    % of Total
```

You realize there are missing data. You decide to include the missing data into your table calculations to see how the marginal statistics are affected.

1 Choose **Stat > Tables > Cross Tabulation and Chi-Square**.

2 In **For rows**, enter *Gender*. In **For columns**, enter *Eyecolor*.

3 Under **Display**, check **Counts, Row percents, Column percents, and Total percents**.

4 Click **Options**, check **Include displayed missing values in calculations**. Click **OK** in each dialog box.

Session window output

Tabulated statistics: Gender, Eyecolor

```
Rows: Gender   Columns: Eyecolor

         Blue   Brown   Green   Hazel  Missing     All

F           7       5       7       5        1      25
           28      20      28      20        4     100
        46.67   38.46   63.64   62.50    33.33   50.00
           14      10      14      10        2      50

M           8       8       4       3        2      25
           32      32      16      12        8     100
        53.33   61.54   36.36   37.50    66.67   50.00
           16      16       8       6        4      50

All        15      13      11       8        3      50
           30      26      22      16        6     100
       100.00  100.00  100.00  100.00   100.00  100.00
           30      26      22      16        6     100

Cell Contents:      Count
                    % of Row
                    % of Column
                    % of Total
```

Interpreting the results

By default, Minitab displays all missing values in your tables, but does not include them in calculations unless you check **Include displayed missing values in calculations** in the Options sub dialog box. When a data point is missing, by default, the entire row (observation) is omitted from the calculation.

For example, in your data, there are 50 rows of data. You are missing 3 observations, 1 observation representing female and 2 observations representing male. The default table ignores the missing values in the calculations and indicates that you have 47 total counts. The table including missing values in the calculations indicates that you have 50 total counts (47 + 1 + 2).

Also, there are 7 females with blue eye color, 5 with brown eye color, 7 with green eye color, and 5 with hazel eye color. The eye color of one female is missing. When missing data are not included in the calculations, the total number of females is 24. When missing data are included in the calculations, the total number of females is 50.

You have to decide which total is more meaningful for your application. If any part of the data for an individual is missing, some researchers would eliminate the entire point from the study, while others would use the data they can.

Example of Chi-Square Analysis using Cross Tabulation and Chi-Square

You are interested in determining whether there is an association between gender and activity level for the individuals in your study. Perform a χ^2 test for association using raw data.

1 Open the worksheet EXH_TABL.MTW. If you have not already done so, set the value order for the variable Activity.

2 Choose **Stat > Tables > Cross Tabulation and Chi-Square**.

3 In **For rows**, enter *Gender*. In **For columns**, enter *Activity*.

4 Under **Display**, check **Counts**.

5 Click **Chi-Square.** Check **Chi-square analysis**, **Expected cell counts**, and **Standardized residuals**. Click **OK** in each dialog box.

Session window output

Tabulated statistics: Gender, Activity

```
Rows: Gender   Columns: Activity

           Slight  Moderate    A lot    All

Female          4        26        5     35
             3.46     23.46     8.08  35.00
           0.2894    0.5241  -1.0827      *

Male            5        35       16     56
             5.54     37.54    12.92  56.00
          -0.2288   -0.4143   0.8559      *

All             9        61       21     91
             9.00     61.00    21.00  91.00
                *         *        *      *

Cell Contents:      Count
                    Expected count
                    Standardized residual


Pearson Chi-Square = 2.487, DF = 2, P-Value = 0.288
Likelihood Ratio Chi-Square = 2.613, DF = 2, P-Value = 0.271

* NOTE * 1 cells with expected counts less than 5
```

Interpreting the results

The cells in the table contain the counts, the expected counts, and the standardized residual. There is no evidence of association (p=0.288, 0.271) between Gender and Activity. Because slightly less than 20% (one of six) have expected counts less than five, you may want to interpret the results with caution.

Example of Changing Table Layout with Cross Tabulation and Chi-Square

You would like to classify the individuals in your study by gender, smoking status, and activity level in a three-way table. To present the same information in a layered two-way table, see Example of cross tabulation with three categorical variables.

1 Open the worksheet EXH_TABL.MTW. If you have not already done so, set the value order for the variable Activity.

2 Choose **Stat > Tables > Cross Tabulation and Chi-Square**.

3 In **For rows**, enter *Gender*. In **For columns**, enter *Activity* and *Smokes*.

4 Under **Display**, check **Counts**. Click **OK**.

Session window output

Tabulated statistics: Gender, Activity, Smokes

```
Rows: Gender   Columns: Activity / Smokes

           Slight    Moderate     A lot      All
          No  Yes    No   Yes    No  Yes     All

Female     3    1    20     6     4    1      35
Male       3    2    22    13    12    4      56
All        6    3    42    19    16    5      91

Cell Contents:      Count
```

Interpreting the results

In this layout, the row variable is Gender, the uppermost column variable is Activity, and innermost column variable is Smokes. It may be easier for you to compare your data in a different layout. You can change your table layout by designating variables to be across rows, down columns, or as layers.

Chi Square Test (Table in Worksheet)

Chi-Square Test (Table in Worksheet)

Stat > Tables > Chi-Square Test (Table in Worksheet)

You can perform a chi-square test of association between variables if your data are in table form.

If your data are in raw or frequency form, use Cross Tabulation and Chi-Square.

Dialog box items

Columns containing the table: Enter up to seven columns containing the contingency table data. Delete rows with missing data before using this procedure.

Data – Chi-Square (Table in Worksheet)

Your data should be in the following format:

- Data must be in contingency table form. See Arrangement of Input Data.
- You can have up to seven columns in your contingency table.
- You must delete rows of missing data from the worksheet.

If you wish to change the order in which text categories are processed from their default alphabetized order, see Ordering Text Categories.

Note If data are in raw or frequency form, use Cross Tabulation - Chi-square.

Arrangement of Input Data

Data used for input to the Tables procedures can be arranged in four different ways in your worksheet:

Raw data – one row for each observation:

C1	C2
1	1
2	1
2	3
1	2

C1= gender, C2 = politics

Contingency table – each cell contains counts:

C1	C2
17	18
10	19
19	17

C1= males, C2 = females

Rows 1-3 represent the three levels for politics

Frequency data – each row represents a unique combination of group codes:

C1	C2	C3
1	1	17
1	2	10
1	3	19
2	1	18

C1= gender, C2 = politics

Indicator variables – one row for each observation

C1	C2	C3	C4	C5
1	0	1	0	0
0	1	1	0	0
0	1	0	0	1
1	0	0	1	0

C1 = 1 if male, 0 if female

C2 = 1 if female, 0 if male

C3 = 1 if Democrat, 0 = otherwise

C4 = 1 if Republican, 0 = otherwise

C5 = 1 if Other, 0 = otherwise

To create:

- frequency data from raw data, see Store Descriptive Statistics.

- contingency table from raw data or frequency data, see Cross Tabulation and Chi-Square, then copy and paste the table output into your worksheet.
- indicator variables from raw data, see Calc > Make Indicator Variables.

To perform a chi-square test for association with contingency table data

1 Choose **Stat** > **Tables** > **Chi-Square Test (Table in Worksheet)**.

2 In **Columns containing the table**, enter the columns containing the contingency table data. Click **OK**.

Cells with Small Expected Frequencies

Minitab displays the number of cells that have expected frequencies less than 5. In Example of Chi-Square, there were 2 such cells. Some statisticians hesitate to use the χ^2 test if more than 20% of the cells have expected frequencies below five, especially if the p-value is small and these cells give a large contribution to the total χ^2 value.

If any cell has an expected frequency less than one, the total χ^2 is not displayed, because the results may not be valid. If some cells have small expected frequencies, consider combining or omitting row and/or column categories.

Chi-Square Goodness-of-Fit Test

Use chi-square goodness-of-fit test to determine if the data can be adequately modeled by the selected distribution. Consider these common probability models:

- Integer distribution – all possible outcomes are equally likely, such as rolling dice.
- Discrete distribution – all possible outcomes have unequal probabilities, such as defining A to have probability = 0.1, probability of B = 0.6, and probability of C = 0.3.
- Binomial distribution – only two possible outcomes, say "success" and "failure", such as tossing a coin.
- Poisson distribution – events occur randomly in time, such as flight arrival times.

The test statistics has a chi-square distribution when the following assumptions are met:

- The data are obtained from a random sample.
- The expected frequency of each category must be at least 5.

Example of calculating the expected number of outcomes, the test statistics, and the p-value for a binomial case.

Suppose you count how many times heads appears in 5 coin tosses, and you repeat the set of 5 tosses 1,000 time. There can be from 0 to 5 heads in any set of tosses. You observed the following outcomes:

Number of Heads	0	1	2	3	4	5
Observed	39	166	298	305	144	48

To test whether the coin toss is fair (heads appears half the time), you need to calculate the expected number of outcomes, then calculate the test statistics and the associated p-value.

Step 1: Calculate the expected number of outcomes for a binomial case

1 Enter the possible outcomes *0*, *1*, *2*, *3*, *4*, and *5* in a worksheet column. Name the column by typing *Outcomes* in the name cell.

2 Choose **Calc** > **Probability Distributions** > **Binomial**.

3 Choose **Probability**.

4 In **Number of trials**, enter *5*. In **Probability of success**, enter *0.5*.

5 Choose **Input column**, then enter *Outcomes*. In **Optional storage**, type *Probs* to name the storage column. Click **OK**.

6 Choose **Calc** > **Calculator**.

7 In **Store result in variable**, type *Expected* to name the storage column.

8 In **Expression**, enter **Probs * 1000**. Click **OK**. Numbered

Step 2: Calculate the test statistics and the p-value

1 Enter the observed values in a worksheet column. Name the column by typing *Observed* in the name cell.

2 Choose **Calc** > **Calculato**r.

3 In **Store result in variable**, type *Chisquare*.

4 In **Expression**, enter **SUM((Observed – Expected)**2 / Expected)**. Click **OK**.

5 Choose **Calc** > **Probability Distributions** > **Chi-Square.**

6 Choose **Cumulative probability**, and in **Degrees of freedom**, type *5*. (the number of outcomes minus one, 6 – 1 = 5).

7 Choose **Input column**, and enter *Chisquare*. In **Optional storage**, type *CumProb* . Click **OK**.

8 Choose **Calc** > **Calculator**.

9 In **Store result in variable**, type *Pvalue*.

10 In **Expression**, enter **1** – **CumProb**. Click **OK**.

Data window output

Note The expected number of outcomes, χ^2 statistic and p-value are stored in columns in the worksheet.

Outcomes	Probs	Expected	Observed	Chisquare	CumProb	Pvalue
0	0.0313	31.25	39	13.3216	0.9795	0.0205
1	0.1563	156.25	166			
2	0.3125	312.50	298			
3	0.3125	312.50	305			
4	0.1563	156.25	144			
5	0.0313	31.25	48			

Interpreting the results

The column, Probs, contains the expected probability of obtaining from 0 to 5 heads in 5 coin tosses. The expected probability of obtaining 2 heads with 5 coin tosses is 0.3125. The column, Expected, contains the expected number of outcomes out of 1000 that you would obtain from 0 to 5 heads in 5 tosses. For example, if you repeated the coin toss 1000 times, you would expect to have 31 trials that had five heads. The p-value of 0.0205 associated with the chi-square statistic of 13.3216 indicates that the binomial probability model with p = 0.5 may not be a good model for this experiment.

The observed number of outcomes are not consistent with the expected number of outcomes using a binomial model. This is not what you would expect with a fair coin; however, it will happen occasionally. (If $\alpha = 0.05$, 5 out of 100 experiments may give false results.)

Example of Chi-Square Test (Table in Worksheet)

You are interested in the relationship between gender and political party affiliation. You query 100 people about their political affiliation and record the number of males (row 1) and females (row 2) for each political party. The worksheet data appears as follows:

```
C1          C2            C3
Democrat    Republican    Other
28          18            4
22          27            1
```

1 Open the worksheet EXH_TABL.MTW.

2 Choose **Stat > Tables > Chi-Square Test (Table in Worksheet)**.

3 In **Columns containing the table**, enter *Democrat*, *Republican* and *Other*. Click **OK**.

Session window output

Chi-Square Test: Democrat, Republican, Other

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

       Democrat  Republican  Other  Total
    1        28          18      4     50
          25.00       22.50   2.50
          0.360       0.900  0.900

    2        22          27      1     50
          25.00       22.50   2.50
          0.360       0.900  0.900

Total        50          45      5    100
```

```
Chi-Sq = 4.320, DF = 2, P-Value = 0.115
2 cells with expected counts less than 5.
```

Interpreting the results

No evidence exists for association (p = 0.115) between gender and political party affiliation. Of the 6 cells, 2 have expected counts less than five (33%). Therefore, even if you had a significant p-value for these data, you should interpret the results with caution. To be more confident of the results, repeat the test, omitting the Other category.

Descriptive Statistics

Descriptive Statistics

Stat > Tables > Descriptive Statistics

Use to generate tables containing count statistics for categorical variables and summary statistics for associated, numeric variables.

For summary statistics on a specific column, use Stat > Basic Statistics > Display Descriptive Statistics.

Dialog box items

Categorical Variables

For rows: Enter the columns containing the categories that define the rows of the table.

For columns: Enter up to two columns containing the categories that define the columns of the table.

For layers: Enter the columns containing the categories that define the layers of the two-way tables.

Frequencies are in: If you have frequency data, enter the column containing the frequencies.

Display summaries for

Categorical Variables: Click to specify count statistics for categorical variables.

Associated Variables: Click to specify summary statistics for variables associated with the categorical variables.

<Options>

Data – Descriptive Statistics

To use Descriptive Statistics, your data should be in the following format:

- Data may be arranged in raw form or frequency form. See Arrangement of Input Data.
- Categorical variables may be numeric, text, or date/time. Associated variables must be numeric or date/time. Frequency data must be non-negative integers.
- Column lengths must be the same.

By default, Minitab excludes missing data from all calculations. See Descriptive Statistics Options.

To change the order in which text categories are processed from their default alphabetized order, see Ordering Text Categories.

To create a table with descriptive statistics

1 Choose **Stat > Tables > Descriptive Statistics**.

2 Do one or more of the following, depending on your desired table layout:

- In **For rows**, enter the columns containing the categories that define the rows of the table.
- In **For columns**, enter the up to two columns containing the categories that define the columns of the table.
- In **For layers**, enter the columns containing the categories that define the layers of the two-way tables.

3 If you have frequency data, check **Frequencies are in** and enter the column containing the frequencies.

4 In **Display summaries for**

- Click **Categorical variables** to specify count statistics for categorical variables.
- Click **Associated variables** to specify summary statistics for variables associated with the categorical variables.

5 If you like, use any of the dialog box options, then click **OK**.

Table Layout

Table layout is dependent upon:

- The number of categorical variables
- Designation of categorical variables into rows and columns
- Layering of categorical variables

Each table can have up to ten categorical variables, and you can choose which variables will be arranged into rows or columns. You can organize variables to emphasize a relationship, or create a more compact display for a report.

Table layout	Example
One-way table one categorical variable	Rows: Gender Count Female 35 Male 56 All 91
Two-way table two categorical variables	Rows: Gender Columns: Smokes No Yes All Female 27 8 35 Male 37 19 56 All 64 27 91 Cell Contents: Count
Multi-way table three to ten categorical variables	Rows: Gender / Smokes Columns: Activity A lot Moderate Slight All Female No 4 20 3 27 Yes 1 6 1 8 Male No 12 22 3 37 Yes 4 13 2 19 All All 21 61 9 91 Cell Contents: Count
Layering Variables each level of the layering variable is a separate two-way table	Results for Activity = A lot Rows: Gender Columns: Smokes No Yes All Female 4 1 5 Male 12 4 16 All 16 5 21 Cell Contents: Count Results for Activity = Moderate Rows: Gender Columns: Smokes No Yes All Female 20 6 26 Male 22 13 35 All 42 19 61 Cell Contents: Count

Note To perform a χ^2 test for association, you must use a two-way table layout.

To set the value order:

1 Click on any cell in the variable **Activity** of the worksheet.

2 Choose **Editor > Column > Value Order**.

3 Choose **User-specified order**.

4 In the right box change **A lot**, **Moderate**, **Slight** to **Slight**, **Moderate**, **A lot**. You can use the right-mouse button to cut and paste. Click **OK**.

Descriptive Statistics – Categorical Variables

Stat > Tables > Descriptive Statistics > Categorical Variables

Displays a table with summary count statistics for categorical variables.

Dialog box items

Display

Counts: Check to display the number of values in each cell and for the margins.

Row percents: Check to display the percentage of each cell represents of the total observations in the table row.

Column percents: Check to display the percentage of each cell represents of the total observations in the table column.

Total percents: Check to display the percentage of each cell represents of all the observations in the table.

Descriptive Statistics – Associated Variables

Stat > Tables > Descriptive Statistics > Associated Variables

Displays summary statistics for associated variables.

Dialog box items

Associated variables: Enter the columns containing the variables to be summarized.

Display

Means: Check to display cell means.

Medians: Check to display cell medians.

Minimums: Check to display cell minimums.

Maximums: Check to display cell maximums.

Sums: Check to display cell sums.

Standard deviations: Check to display cell standard deviations.

Data: Check to display all of the data in each cell.

N nonmissing: Check to display the number of nonmissing values.

N missing: Check to display the number of missing values.

Proportion equal to: Check to display the proportion of observations with the specified value for each cell in the table, then specify the desired value. You must delete all missing values before using this option.

Proportion between___and___: Check to display the proportion of observations which fall in the specified range for each cell in the table, then specify the low and high values. You must delete all missing values before using this option.

Descriptive Statistics – Options

Stat > Tables > Descriptive Statistics > Options

You can display marginal statistics and missing data in table display and calculations.

Dialog box items

Display marginal statistics for

Rows and columns: Choose to display marginal statistics for all categorical variables. By default, Minitab displays marginal statistics for all categorical variables.

No variables: Choose to exclude marginal statistics from the table.

Selected variables: Choose to display marginal statistics for specific categorical variables, then enter one or more columns.

Display missing values for

All variables: Choose to display missing values for all variables. By default, Minitab displays missing values for all categorical variables.

No variables: Choose to suppress missing values.

Selected variables: Choose to display missing values for specific categorical variables, then enter one or more columns.

Include displayed missing values in calculations Check to include missing values in calculations. Minitab uses the displayed missing values that you selected above in calculations, so make sure that **Display Missing Values** is selected appropriately. By default, Minitab does not include missing values in calculations.

Example of Data Display in a Table

Suppose you are interested in displaying pulse data for individuals in a smoking study sorted by smoking status.

1 Open the worksheet PULSE1.MTW.

2 Choose **Stat > Tables > Descriptive Statistics**.

3 In **For rows**, enter *Smokes*.

4 Click **Associated variables**, enter *Pulse1* and *Pulse2*.

5 Under **Display**, check **Data**. Click **OK** in each dialog box.

Session window output

Tabulated statistics: Smokes

```
Rows: Smokes

       DATA:    DATA:
      Pulse1   Pulse2

1         78      104
         100      115
          88      110
          62       98

2         96      140
          62      100
          82      100
          68      112
          96      116
          78      118
          80      128
```

Interpreting the results

The numbers in the table are the Pulse data corresponding to the individual's smoking status.

Example of Tables with Missing Data

You would like to summarize weight data for individuals in your study by gender and activity level.

1 Open the worksheet EYECOLOR.MTW.

2 Choose **Stat > Tables > Descriptive Statistics**.

3 In **For rows**, enter *Gender*. In **For columns**, enter *Eyecolor*.

4 Click **Associated variables**, enter *Weight*.

5 Under **Display**, check **Means** and **Standard deviations**. Click **OK** in each dialog box.

Session window output

Tabulated statistics: Gender, Eyecolor

```
Rows: Gender   Columns: Eyecolor

        Blue  Brown  Green  Hazel  Missing    All

F      124.2  123.3  118.0  125.0    120.0  122.2
       19.25  10.63  13.40  18.03        *  15.06

M      156.3  159.1  156.8  164.7    170.0  158.4
       28.47  13.91  20.77  10.02        *  20.12

All    142.5  146.1  132.1  139.9        *  140.3
       29.16  21.86  24.89  25.22        *  25.36

Cell Contents:  Weight  :  Mean
                Weight  :  Standard deviation
```

You realize there are missing data. You decide to include the missing data into your table calculations to see how the marginal statistics are affected.

1 Choose **Stat > Tables > Descriptive Statistics**.

2 In **For rows**, enter *Gender*. In **For columns**, enter *Eyecolor*.

3 Click **Associated variables**, enter *Weight*.

4 Under **Display**, check **Means** and **Standard deviations**. Click **OK**.

5 Click **Options**, check **Include displayed missing values in calculations**. Click **OK** in each dialog box.

Session window output

Tabulated statistics: Gender, Eyecolor

```
Rows: Gender   Columns: Eyecolor

        Blue  Brown  Green  Hazel  Missing    All

F      124.2  123.3  118.0  125.0    120.0  122.1
       19.25  10.63  13.40  18.03        *  14.72

M      156.3  159.1  156.8  164.7    170.0  158.9
       28.47  13.91  20.77  10.02        *  19.80

All    142.5  146.1  132.1  139.9    145.0  140.5
       29.16  21.86  24.89  25.22    35.36  25.36

Cell Contents:  Weight  :  Mean
                Weight  :  Standard deviation
```

Interpreting the results

By default, Minitab displays all missing values in your tables, but does not include them in calculations unless you check **Include displayed missing values in calculations** in the Options sub dialog box. When a data point is missing, by default, the entire worksheet row (observation) is omitted from the calculation.

In this example, you are missing activity data for 2 females and 2 males. If you do not include missing data in your calculations, the mean weight of females in your study is 122.2 pounds. If you include the weight data for females even when the activity level is missing, the mean weight is 122.1 pounds.

You have to decide which statistic is more meaningful for your application. If any part of the data for an individual is missing, some researchers would eliminate the entire point from the study, while others would use the data they can.

Example of Tables with Additional Statistical Summaries

You would like to summarize the heights and weights of the individuals in your study classified by gender and activity level.

1 Open the worksheet EXH_TABL.MTW. If you have not already done so, set the value order for the variable Activity.

2 Choose **Stat > Tables > Descriptive Statistics**.

3 In **For rows**, enter *Gender*. In **For columns**, enter *Activity*.

4 Click **Associated variables**, enter *Height* and *Weight*.

5 Under **Display**, check **Means**, **Standard deviations**, and **N nonmissing**. Click **OK** in each dialog box.

Session window output

Tabulated statistics: Gender, Activity

```
Rows: Gender   Columns: Activity

          Slight  Moderate  A lot    All

Female     65.00     65.62  64.60  65.40
           123.0     124.5  121.0  123.8
           2.160     2.735  2.074  2.563
            7.70     12.78  21.02  13.37
               4        26      5     35
               4        26      5     35

Male       72.40     70.43  71.13  70.80
           170.0     158.1  155.5  158.4
           2.510     2.521  2.649  2.579
           19.69     20.58  13.21  18.77
               5        35     16     56
               5        35     16     56

All        69.11     68.38  69.57  68.73
           149.1     143.8  147.3  145.1
           4.485     3.532  3.773  3.679
           28.80     24.27  21.12  23.87
               9        61     21     91
               9        61     21     91

Cell Contents:  Height  :  Mean
                Weight  :  Mean
                Height  :  Standard deviation
                Weight  :  Standard deviation
                Height  :  Nonmissing
                Weight  :  Nonmissing
```

Interpreting the results

Minitab displays the mean, standard deviation, and sample size of Height and Weight, classified by Gender and Activity. For example, the men with moderate activity level have a mean weight of 158.1 lbs, with standard deviation of 20.58 lbs. These values are based on 35 nonmissing observations. The column margins are the overall statistics for all the data. For example, the mean height of all the participants in your study is 68.73 inches.

Example of Tables with Categorical and Associated Variable Summaries

You would like to summarize count statistics and pulse data for the individuals in your study classified by gender and activity level.

1 Open the worksheet EXH_TABL.MTW.

2 Choose **Stat > Tables > Descriptive Statistics**.

3 In **For rows**, enter *Gender*. In **For columns**, enter *Smokes*.

4 Click **Categorical variables**, check **Counts** and **Row Percents**. Click **OK**.

5 Click **Associated variables**, enter *Pulse*.

6 Under **Display**, check **Means**. Click **OK** in each dialog box.

Session window output

Tabulated statistics: Gender, Smokes

```
Rows: Gender   Columns: Smokes

             No    Yes     All

Female    74.59  84.50   76.86
             27      8      35
          77.14  22.86  100.00

Male      70.00  72.42   70.82
             37     19      56
          66.07  33.93  100.00

All       71.94  76.00   73.14
             64     27      91
          70.33  29.67  100.00

Cell Contents:  Pulse  :  Mean
                          Count
                          % of Row
```

Interpreting the results

This example shows a table of both categorical count summaries and associated variable statistical summaries. Minitab displays the mean pulse value, the counts, and the row percents classified by gender and smoking status. Of the 56 men in your study, 19 are smokers. Their mean pulse rate is 72.42. The corresponding row percent is 33.93.

EDA

Overview

Exploratory Data Analysis Overview

Exploratory data analysis (EDA) methods are used primarily to explore data before using more traditional methods, or to examine residuals from a model. These methods are particularly useful for identifying extraordinary observations and noting violations of traditional assumptions, such as nonlinearity or nonconstant variance.

- Stem-and-Leaf displays a character-based stem-and-leaf plot.
- Boxplot displays a box-and-whiskers plot.
- Letter Values generates a letter-value display. Use this procedure to describe the location and spread of sample distributions.
- Median Polish fits an additive model to a two-way design and identifies data patterns not explained by row and column effects. This procedure is similar to analysis of variance except medians are used instead of means, thus adding robustness against the effect of outliers.
- Resistant Line uses a method that is resistant to outliers to fit a straight line to your data. You can fit a resistant line before using a least squares regression to see if the relationship is linear, to find re-expressions to linearize the relationship if necessary, and to identify outliers.
- Resistant Smooth smoothes an ordered sequence of data, usually collected over time, to remove random fluctuations. Smoothing is useful for discovering and summarizing both data trends and outliers.
- Rootogram displays a suspended rootogram for your data. A suspended rootogram is a histogram with a normal distribution fit to it, which displays the deviations from the fitted normal distribution.

EDA

Stat > EDA

Exploratory data analysis methods are used primarily to explore data before using more traditional methods, or to examine residuals from a model. They are particularly useful for identifying extraordinary observations and noting violations of traditional assumptions such as nonlinearity or nonconstant variance.

Select one of the following commands:

Stem-and-Leaf - does a stem-and-leaf plot

Boxplot - does a box-and-whiskers plot

Letter Values - prints a letter-value display

Median Polish - uses median polish to analyze a two-way layout

Resistant Line - fits a line to data using a procedure that is resistant to outliers

Resistant Smooth - smoothes data (usually a time series)

Rootogram - prints a suspended rootogram

Exploratory Data Analysis Examples

Minitab provides the following examples:

Letter-Value Display

Median Polish

Rootogram

References for EDA

[1] P.F. Velleman (1980). "Definition and Comparison of Robust Nonlinear Data Smoothing Algorithms," Journal of the American Statistical Association, Volume 75, Number 371, pp.609–615.

[2] P.F. Velleman and D.C. Hoaglin (1981). ABC's of EDA, Duxbury Press.

Stem-and-Leaf

Stem-and-Leaf

Stem-and-Leaf

Boxplot

Boxplot

Gallery

Data

One Y, Simple

To display a simple boxplot

Example, one y - simple

One Y, With Groups

To display a boxplot with groups

Example, one y - with groups

Multiple Y's, Simple

To display a simple boxplot with multiple y's

Example, multiple y's - simple

Multiple Y's, With Groups

To display a boxplot with multiple y's and groups

Example, multiple y's - with groups

Letter Values

Letter Values

Stat > EDA > Letter Values

Use letter-value displays to describe the location and spread of sample distributions. The statistics given depend on the sample size, and include median, hinges, eighths, and more.

Dialog box items

Variable: Select the column containing the variable for which you want to obtain letter values.

Store letter values: Check to store the letter values. The storage column will contain all the numbers on the output listed under LOWER (starting from the bottom and going up), then the median, and then the numbers listed under UPPER (starting from the top and going down).

Store mids: Check to store the midpoint for each letter value.

Store spreads: Check to store the spreads for each letter value.

Data – Letter Values

You need one column that contains numeric or date/time data, but no missing values. Delete any missing values from the worksheet before displaying letter values.

To display letter values

1 Choose **Stat > EDA > Letter Values**.

2 In **Variable**, enter the column that contains the data for which you want to obtain letter values.

3 If you like, use any of the dialog box options, then click **OK**.

Example of displaying letter values

Students in an introductory statistics course participated in a simple experiment. Before beginning the experiment each student recorded their resting pulse rate. We will use a letter-value display to describe the location and spread of the resting pulse rate data.

1 Open the worksheet PULSE.MTW.

2 Choose **Stat > EDA > Letter Values**.

3 In **Variable**, enter **Pulse1**. Click **OK**.

Session window output

Letter Value Display: Pulse1

```
    Depth   Lower    Upper     Mid  Spread
N=     92
M    46.5  71.000   71.000  71.000
H    23.5  64.000   80.000  72.000  16.000
E    12.0  62.000   88.000  75.000  26.000
D     6.5  59.000   91.000  75.000  32.000
C     3.5  56.000   95.000  75.500  39.000
B     2.0  54.000   96.000  75.000  42.000
        1  48.000  100.000  74.000  52.000
```

Interpreting the results

Pulse1 contains 92 observations (N = 92). The letter values displayed are found by moving in from each end of the ordered observations to a given depth.

If the depth does not coincide with a data value, the average of the nearest neighbors is taken.

- The median for this data is the average of the forty-sixth and forty-seventh ordered observations and is 71.
- The hinges are the average of the twenty-second and twenty-third observations from either end, with values of 64 and 80, the average of these being the Mid, or 72. The difference between the upper and lower hinges is the Spread, or 16.
- The eighths (E), sixteenths (D), and other letter values are calculated in a similar fashion.

Median Polish

Median Polish

Stat > EDA > Median Polish

Median Polish fits an additive model to a two-way design and identifies data patterns not explained by row and column effects. This procedure is similar to analysis of variance except medians are used instead of means, thus adding robustness against the effect of outliers. For a complete discussion, see [1] and [2].

Median Polish does not print results. Use Stat > Tables > Descriptive Statistics to display the data, stored fits, or residuals.

Dialog box items

Response: Select the column containing the response variable.

Row factor: Select the column containing the row factor. The levels of the row factor must be consecutive integers starting at 1.

Column factor: Select the column containing the column factor. The levels of the column factor must be consecutive integers starting at 1.

Number of iterations: Enter the number of iterations to find the solution. The default is four.

> **Columns first:** Check to use column medians rather than row medians for the first iteration. Starting with rows and starting with columns does not necessarily yield the same fits, even if many iterations are done.

Storage

Common effect: Specify a column to store the common effects.

Row effects: Specify a column to store the row effects.

Column effects: Specify a column to store the column effects.

Comparison values: Specify a column to store the comparison values.

Residuals: Check to store the residuals.

Fits: Check to store fitted values.

Data – Median Polish

Arrange your data in three numeric columns in the worksheet–a response, a row factor, and a column factor. Each row represents one observation. Row levels and column levels must be consecutive integers starting at one. The table may be unbalanced and may have empty cells, but you cannot have any missing values. Delete any missing values from the worksheet before performing a median polish.

To perform a median polish

1 Choose **Stat > EDA > Median Polish**.

2 In **Response**, enter the column that contains the measurement data.

3 In **Row factor**, enter the column that contains the row factor levels.

4 In **Column factor**, enter the column that contains the column factor levels.

5 If you like, use any of the dialog box options, then click **OK**.

Improving the Fit of an Additive Model

Data that is not well described by an additive model may be made more additive by re-expressing or transforming the data. You can use comparison values to help you choose an appropriate data transformation.

1 Calculate the comparison values for each observation. For an observation in row i and column j:

comparison value +[(row effect i) x (column effect j)] / common effect

2 Plot each residual against its comparison value for visual inspection of the data.

3 Fit a straight line to the data using Resistant Line.

4 Determine whether or not a transformation will improve the fit of the additive model.
Let p = 1 - (slope of the resistant line).

- If p = 1 (the line is horizontal), no simple transformation will improve the model.
- If p = ½, SQRT(Y), where Y is the data, is likely to be more nearly additive (and thus better analyzed by median polish).
- If p = 0, log Y will be more nearly additive.
- If p is between 0 and 1, then Y^p will be more nearly additive.

The exploratory technique described above is similar to Tukey's one degree of freedom for non-additivity method.

Example of Median Polish

Suppose you want to fit a model to experimental data in a two-way design. The experiment involved three types of helmets where a force was applied to the front and the back of the helmet. The two factors of interest are helmet type and location of force applied; whereas, the response measure is impact. The impact was measured to determine whether or not any identifiable data patterns exist that would indicate a difference between the three helmet types and the front and back portion of the helmet, with the level of protection (as measured by Impact) provided. Here, we fit an additive model to a two-way design using a median polish.

Since Median Polish does not display any results, use Display Data and Descriptive Statistics to display results in the Session window.

Step 1: Perform the median polish

1 Open the worksheet EXH_STAT.MTW. Choose **Stat > EDA > Median Polish**.

2 In **Response**, enter **Impact**.

3 In **Row factor**, enter **HelmetType**. In **Column factor**, enter **Location**.

4 In **Common effect**, enter **CommonEffect**. In **Row effects**, enter **RowEffect**. In **Column effects**, enter **ColumnEffect**.

5 Check **Residuals**. Click **OK**.

Session window output

```
This version of MPOLISH does not display any results.
```

Store the results and use Display Data and Cross Tabulation.

Step 2: Display the common, row, and column effects

1 Choose **Data > Display Data**.

2 In **Columns, constants, and matrices to display**, enter **CommonEffect**, **RowEffect**, and **ColumnEffect**. Click **OK**.

Session window output

Data Display

```
CommonEffect    44.5000


Row  RowEffect  ColumnEffect
  1          0            -1
  2         23             1
  3         -3
```

Step 3: Display the data and residuals

1 Choose **Stat > Tables > Descriptive Statistics**.

2 Under **Categorical variables**, in **For rows**, enter **HelmetType.** In **For columns,** enter **Location**.

3 Under **Display summaries for**, click **Associated variables.**

4 Under **Associated variables**, enter **Resi1**. Under **Display**, check **Data**. Click **OK** in each dialog box.

Session window output

Tabulated statistics: HelmetType, Location

```
Rows: HelmetType   Columns: Location

        1     2

1     3.5   0.5
     -0.5  -5.5


2    -4.5  -1.5
      1.5   2.5


3     0.5  -0.5
     -1.5   3.5

Cell Contents:  RESI1  :  DATA
```

Interpreting the results

This section is based on the output from both steps 2 and 3. The common effect, which summarizes the general level of Impact, is 44.5.

The row effects account for changes in Impact from row to row relative to the common value. The row effects are 0, 23, –3 for helmet type 1, 2, and 3, respectively, indicating that the impact for helmet type 2 was much higher than the common level; whereas, helmet type 3 was slightly lower.

The column effects account for changes in Impact from column to column relative to the common value. The column effects are –1, 1 for locations 1 and 2, respectively, indicating that the impact was slightly lower than the common effect for the front of the helmet and slightly higher for the back of the helmet.

The residuals for the two observations per cell are shown in the printed table. These are 3.5 and –0.5 for cell 1,1, and so on. You can use the residuals to identify extraordinary values.

Resistant Line

Resistant Line

Stat > EDA > Resistant Line

Resistant line fits a straight line to your data using a method that is resistant to outliers. Velleman and Hoaglin [2] suggest fitting a resistant line before using least squares regression to see if the relationship is linear, to find re-experiences to linearize the relationship if necessary, and to identify outliers.

Dialog box items

Response: Select the column containing the response variable (Y) At least six, but preferably nine or more observations are needed.

Predictor: Select the column containing the predictor variable (X).

Maximum number of iterations: Specify the maximum number of iterations used to find a solution. The default is 10. This procedure will stop before the specified number of iterations if the value of the slope does not change very much.

Storage

Residuals: Check to store the residuals.

Fits: Check to store the fitted values.

Coefficients: Check to store the coefficients.

<Results>

Data – Resistant Line

You must have two numeric columns–a response variable column and predictor variable column–with at least six, but preferably nine or more, observations.

Minitab automatically omits missing data from the calculations.

To fit a resistant line

1 Choose **Stat > EDA > Resistant Line**.

2 In **Response**, enter the column that contains the measurement data (Y).

3 In **Predictor**, enter the column that contains the predictor variable data (X).

4 If you like, use any of the dialog box options, then click **OK**.

Resistant Line – Results

Stat > EDA > Resistant Line > Results

Control display of Session window output.

Dialog box items

Control the Display of Results

Display nothing: Choose to suppress display of all results.

Slope, level, and half-slope ratio: Choose to display the slope, level, and half-slope ratio (default).

In addition, the slope for each iteration: Choose to display the slope, level, half-slope ratio, and the slope for each iteration..

Resistant Smooth

Resistant Smooth

Stat > EDA > Resistant Smooth

Resistant Smooth smoothes an ordered series of data, usually collected over time, to remove random fluctuations. Smoothing is useful for discovering and summarizing both data trends and outliers. Resistant Smooth offers two smoothing methods: **4253H, twice** and **3RSSH, twice**. See Method.

Dialog box items

Variable: Select the column containing the variable to be smoothed. Resistant Smooth requires at least seven observations.

Storage

Rough: Specify a column to store the rough data; rough data = raw data – smoothed data.

Smooth: Specify a column to store the smoothed data.

Method of Smoothing: Allows you to choose one of two methods to smooth the data.

4253H, twice: Choose to perform a 4253H smoother twice.

3RSSH, twice: Choose to perform a 3RSSH smoother twice.

Data – Resistant Smooth

You must have a numeric column with at least seven observations. You can have missing data at the beginning and end of the column, but not in the middle.

To perform a resistant smoothing

1 Choose **Stat > EDA > Resistant Smooth**.

2 In **Variable**, enter the column that contains the raw data to be smoothed.

3 In **Rough**, enter a column to store the rough data (rough data = raw data - smoothed data).

4 In **Smooth**, enter a column to store the smoothed data.

5 If you like, use any of the dialog box options, then click **OK**.

Rootogram

Rootogram

Stat > EDA > Rootogram

A suspended rootogram is a histogram with a normal distribution fit to it, which displays the deviations from the fitted normal distribution. Since a rootogram is fit using percentiles, it protects against outliers and extraordinary bin counts. For further details see [2].

Dialog box items

Source of Data

Variable: Choose to have raw data used to form the rootogram, then specify the column containing the raw data.

Frequencies: Choose to have counts (frequencies) used to form the rootogram, then specify the column containing the frequencies. The first count is for the half-open bin below the lowest bin boundary and must be zero if no observations fall below the first bin boundary. Similarly, the last count is for the half-open bin above the highest bin boundary. Therefore, the column of counts has one more entry than the column of bin boundaries.

Use bin boundaries in: Check to specify bin boundaries, then specify the column containing the bin boundaries. Input the bin boundaries from smallest to largest down the column. If no bin boundaries are given, the bins are taken to have width = 1.

Use mean___and **std. dev.___:** Check to override the automatic estimation of the mean and standard deviation used in fitting the Normal comparison curve. Then specify the mean and standard deviation you want to use.

Storage

Bin boundaries: Specify a storage column for the bin boundaries.

Counts: Specify a storage column for the frequencies.

Double root residuals: Specify a storage column for the double root residuals (DRRes). The suspended rootogram is a plot of the DRRes, using the sign of the DRRes for the plotting symbol.

Fits: Specify a column to store the fitted counts. The fitted count, fi, is N x (area under the normal curve with the specified mean and stdev, in bin i).

Data – Rootogram

Your data can be in one of two forms: raw or frequency. To use

- raw data, you need one column of numeric or date/time data. By default, the rootogram procedure will determine the bin boundaries.

- frequency data, you need one numeric column that contains the count (frequency) of observations for each bin. The frequencies need to be ordered down the column from the upper-most bin to the lower-most bin (equivalent to the left-most and right-most bins in a histogram, respectively). By default, the bins have a width of 1.

Optionally, you can specify the bin boundaries for both raw and frequency data in another column. In the bin boundary column, enter the bin boundaries down the column from the smallest to largest.

If you are using bin boundaries with frequency data, the first row of the frequency data column is the count for the number of observations that fall below the smallest bin boundary. If no observations fall below the first bin boundary, the count in the first row is zero. Similarly, the last row of the frequency data column contains the count for the number of observations that fall above the largest bin boundary. The frequency data column will have one more entry than the column of bin boundaries.

Minitab automatically omits missing data from the calculations.

To display a suspended rootogram

1 Choose **Stat > EDA > Rootogram**.

2 Do one of the following:

- under **Source of Data**, choose **Variable**, and enter the column that contains the raw data, or
- choose **Frequencies**, and enter the column that contains the counts

3 If you like, use one or more of the dialog box options, then click **OK**.

Example of Rootogram

Here, we use a rootogram to determine whether or not the weight measurements from 92 students follow a normal distribution.

1 Open the worksheet PULSE.MTW.

2 Choose **Stat > EDA > Rootogram**.

3 In **Variable**, enter **Weight**. Click **OK**.

Session window output

Rootogram: Weight

```
Bin  Count  RawRes  DRRes              Suspended Rootogram
  1    0.0    -0.7  -0.90                   .     -----          .
  2    0.0    -1.2  -1.44                   .  --------          .
  3    2.0    -0.8  -0.35                   .        --          .
  4    5.0    -0.5  -0.10                   .         -          .
  5   12.0     3.0   0.99                   .          +++++     .
  6   12.0    -0.5  -0.06                   .         -          .
  7   11.0    -3.7  -0.94                   .     -----          .
  8   17.0     2.4   0.66                   .          ++++      .
  9   16.0     3.7   1.03                   .          ++++++    .
 10    5.0    -3.8  -1.34                   .   -------          .
 11    5.0    -0.4  -0.04                   .         -          .
 12    5.0     2.2   1.23                   .          +++++++   .
 13    1.0    -0.2   0.04                   .          +         .
 14    0.0    -0.4  -0.66                   .      ----          .
 15    1.0     0.9   1.20                   .          +++++++   .
 16    0.0    -0.0  -0.09                   .         -          .

 In display, value of one character is .2             OO
```

Interpreting the results

The suspended rootogram plots the double root residuals (DRRes) using the sign of the DRRes as the plotting symbol. The DRRes indicate how closely the data follow the comparison (normal) distribution.

For the Weight variable, there is a slight concentration of negative signs in the lower bins, and the highest concentration of positive signs in the middle and higher bins, indicating where the sample distribution tends to depart from normal. However, there is no strong evidence that a normal distribution could not be used to describe these data because the double root residuals are all within the confidence limits.

Power and Sample Size

Overview

Power and Sample Size Overview

Use Minitab's power and sample size capabilities to evaluate power and sample size before you design and run an experiment (prospective) or after you perform an experiment (retrospective).

- A **prospective** study is used before collecting data to consider design sensitivity. You want to be sure that you have enough power to detect differences (effects) that you have determined to be important. For example, you can increase the design sensitivity by increasing the sample size or by taking measures to decrease the error variance.
- A **retrospective** study is used after collecting data to help understand the power of the tests that you have performed. For example, suppose you conduct an experiment and the data analysis does not reveal any statistically significant results. You can then calculate power based on the minimum difference (effect) you wish to detect. If the power to detect this difference is low, you may want to modify your experimental design to increase the power and continue to evaluate the same problem. However, if the power is high, you may want to conclude that there is no meaningful difference (effect) and discontinue experimentation.

Minitab provides power, sample size, and difference (effect) calculations (also the number of center points for factorial and Plackett-Burman designs) for the following procedures:

- one-sample Z
- one-sample t
- two-sample t
- one-sample proportion
- two-sample proportion
- one-way analysis of variance
- two-level factorial designs
- Plackett-Burman designs

What is Power?

There are four possible outcomes for a hypothesis test. The outcomes depend on whether the null hypothesis (H0) is true or false and whether you decide to "reject" or "fail to reject" H0. The **power** of a test is the probability of correctly rejecting H0 when it is false. In other words, power is the likelihood that you will identify a significant difference (effect) when one exists.

The four possible outcomes are summarized below:

	Null Hypothesis	
Decision	**True**	**False**
fail to reject H0	correct decision p = 1 – α	Type II error p = β
reject H0	Type I error p = α	**correct decision** **p = 1 – β**

When H0 is true and you reject it, you make a Type I error. The probability (p) of making a Type I error is called **alpha** (α) and is sometimes referred to as the **level of significance** for the test.

When H0 is false and you fail to reject it, you make a type II error. The probability (p) of making a type II error is called **beta** (β).

Choosing probability levels

When you are determining the α and β values for your test, you should consider the

- **severity of making an error**–The more serious the error, the less often you will be willing to allow it to occur. Therefore, you should assign smaller probability values to more serious errors.
- **magnitude of effect you want to detect**–**Power** is the probability (p = 1 - β) of correctly rejecting H0 when it is false. Ideally, you want to have high power to detect a difference that you care about, and low power for a meaningless difference.

  For example, suppose you want to claim that children in your school scored higher than the general population on a standardized achievement test. You need to decide how much higher than the general population your test scores need to be so you are not making claims that are misleading. If your mean test score is only 0.7 points higher than the general population on a 100 point test, do you really want to detect a difference? Probably not. Therefore, you should choose your sample size so that you only have power to detect differences that you consider meaningful.

Factors that influence power

A number of factors influence power:

- α, the probability of a Type I error (level of significance). As the probability of a Type I error (α) increases, the probability of a type II error (β) decreases. Hence, as α increases, power = 1 – β also increases.

- σ, the variability in the population. As σ increases, power decreases.
- the size of the population difference (effect). As the size of population difference (effect) decreases, power decreases.
- sample size. As sample size increases, power increases.

Power and Sample Size

Stat > Power and Sample Size

Allows you to calculate power and sample size for the following procedures:

1-Sample Z

1-Sample t

2-Sample t

1 Proportion

2 Proportions

One-Way ANOVA

2-Level Factorial Design

Plackett-Burman Design

Estimating σ

For power or minimum difference calculations, the estimate of σ depends on whether or not you have already collected data.

- Prospective studies are done before collecting data so σ has to be estimated. You can use related research, pilot studies, or subject-matter knowledge to estimate σ.
- Retrospective studies are done after data have been collected so you can use the data to estimate σ.
  - For 1-Sample Z or 1-Sample t, use the standard deviation of the sample.
  - For 2-Sample t, use the pooled standard deviation (Pooled StDev) if assuming equal variances.
  - For One-way ANOVA, 2-Level Factorial Design, and Plackett-Burman Design, use the square root of the mean square error (MS Error).

For sample size calculations, the data have not been collected yet so the population standard deviation (σ) has to be estimated. You can use related research, pilot studies, or subject-matter knowledge to estimate σ.

Note If you would like to specify the difference (effects) in standardized (sigma) units, enter 1 in **Standard deviation**.

Power and Sample Size examples

The following examples illustrate how to calculate power and sample size for hypothesis tests with Minitab.

sample size for a one-sample t-test

power for a two-sample test of proportion

sample size for a one-way ANOVA

power for a two-level fractional factorial

1-Sample Z

Power and Sample Size for 1-Sample Z

Stat > Power and Sample Size > 1-Sample Z

Use to calculate one of the following for a hypothesis test of the mean when the population standard deviation (σ) is known

- power
- sample size
- minimum difference (effect)

You need to determine what are acceptable values for any two of these parameters and Minitab will solve for the third.

For example, if you specify values for power and the minimum difference, Minitab will determine the sample size required to detect the specified difference at the specified level of power. See Defining the minimum difference.

Dialog box items

Specify values for any two of the following: Specify two values and Minitab will calculate the third.

Sample sizes: Enter one or more sample sizes.

Differences: Enter one or more differences in terms of the null hypothesis.

Power values: Enter one or more power values.

Standard deviation: Enter σ for your data.

Tip You can use shorthand notation to enter sample sizes. For example, 10:40/5 means sample sizes from 10 to 40 by 5.

<Options>

To calculate power, sample size, or minimum difference – Z-test and t-tests

1 Choose **Stat > Power and Sample Size > 1-Sample Z**, **1-Sample t**, **or 2-Sample t**.

2 Do one of the following:

- Solve for power

  1 In **Sample sizes**, enter one or more numbers. For a two-sample test, the number you enter is considered the sample size for each group. For example, if you want to determine power for an analysis with 10 observations in each group for a total of 20, you would enter **10**.

  2 In **Differences**, enter one or more numbers.

- Solve for sample size

  1 In **Differences**, enter one or more numbers.

  2 In **Power values**, enter one or more numbers.

- Solve for the minimum difference

  1 In **Sample sizes**, enter one or more numbers. For a two-sample test, the number you enter is considered the sample size for each group.

  2 In **Power values**, enter one or more numbers.

  Minitab will solve for all combinations of the specified values. For example, if you enter 3 values in **Sample sizes** and 2 values in **Differences**, Minitab will compute the power for all 6 combinations of sample sizes and differences.

  For a discussion of the value needed in **Differences**, see Defining the difference.

3 In **Sigma**, enter an estimate of the population standard deviation (σ) for your data. See Estimating σ.

4 If you like, use one or more of the available dialog box options, then click **OK**.

Defining the minimum difference – Z-test and t-tests

In the main dialog box, you need to specify the minimum difference you are interested in detecting. The manner in which you express this difference depends on whether you are performing a one- or two-sample test:

- For a one-sample Z- or t-test, express the difference in terms of the null hypothesis.

  For example, suppose you are testing whether or not your students' mean test score is different from the population mean. If you would like to detect a difference of three points, you would enter 3 in **Differences**.

- For a two-sample t-test, express the difference as the difference between the population means that you would like to be able to detect.

  For example, suppose you are investigating the effects of water acidity on the growth of two populations of tadpoles. If you are interested in differences of 4 mm or more, you would enter 4 in **Differences**.

If you choose **Less than** as your alternative hypothesis, then you must enter a negative value in **Differences**. If you choose **Greater than**, you must enter a positive value. This is because with a one-tailed test, you have no power to detect an effect in the opposite direction.

Estimating σ

For power or minimum difference calculations, the estimate of σ depends on whether or not you have already collected data.

- Prospective studies are done before collecting data so σ has to be estimated. You can use related research, pilot studies, or subject-matter knowledge to estimate σ.

- Retrospective studies are done after data have been collected so you can use the data to estimate σ.
  - For 1-Sample Z or 1-Sample t, use the standard deviation of the sample.
  - For 2-Sample t, use the pooled standard deviation (Pooled StDev) if assuming equal variances.
  - For One-way ANOVA, 2-Level Factorial Design, and Plackett-Burman Design, use the square root of the mean square error (MS Error).

For sample size calculations, the data have not been collected yet so the population standard deviation (σ) has to be estimated. You can use related research, pilot studies, or subject-matter knowledge to estimate σ.

Note If you would like to specify the difference (effects) in standardized (sigma) units, enter 1 in **Standard deviation**.

Power and Sample Size for 1-Sample Z – Options

Stat > Power and Sample Size > 1-Sample Z > Options

Use to define the alternative hypothesis, specify the significance level, and store the sample sizes, differences (effects), and power values. When calculating sample size, Minitab stores the power value that will generate the nearest integer sample size.

Dialog box items

Alternative Hypothesis Define the alternative hypothesis.

Less than: Choose to perform a lower-tailed test.

Not equal: Choose to perform two-tailed test.

Greater than: Choose to perform an upper-tailed test.

Significance level: Enter the significance level (α). The default is 0.05.

Store sample sizes in: Enter a storage column for sample sizes.

Store differences in: Enter a storage column for differences.

Store power values in: Enter a storage column for power values.

1-Sample t

Power and Sample Size for 1-Sample t

Stat > Power and Sample Size > 1-Sample t

Use to calculate one of the following for a one-sample t-test or paired t-test.

- power
- sample size
- minimum difference (effect)

You need to determine what are acceptable values for any two of these parameters and Minitab will solve for the third.

For example, if you specify values for power and the minimum difference, Minitab will determine the sample size required to detect the specified difference at the specified level of power. See Defining the minimum difference.

Dialog box items

Specify values for any two of the following: Specify two values and Minitab will calculate the third.

Sample sizes: Enter one or more sample sizes. For a paired t-test, this is the number of pairs.

Differences: Enter one or more differences in terms of the null hypothesis.

Power values: Enter one or more power values.

Standard deviation: Enter an estimate of the population standard deviation (σ). For a paired t-test, this is the standard deviation of the differences between pairs. See Estimating σ.

Tip You can use shorthand notation to enter sample sizes. For example, 10:40/5 means sample sizes from 10 to 40 by 5.

<Options>

To calculate power, sample size, or minimum difference – Z-test and t-tests

1 Choose **Stat > Power and Sample Size > 1-Sample Z**, **1-Sample t**, **or 2-Sample t**.

2 Do one of the following:

- Solve for power
  1 In **Sample sizes**, enter one or more numbers. For a two-sample test, the number you enter is considered the sample size for each group. For example, if you want to determine power for an analysis with 10 observations in each group for a total of 20, you would enter **10**.
  2 In **Differences**, enter one or more numbers.
- Solve for sample size
  1 In **Differences**, enter one or more numbers.
  2 In **Power values**, enter one or more numbers.
- Solve for the minimum difference
  1 In **Sample sizes**, enter one or more numbers. For a two-sample test, the number you enter is considered the sample size for each group.
  2 In **Power values**, enter one or more numbers.

Minitab will solve for all combinations of the specified values. For example, if you enter 3 values in **Sample sizes** and 2 values in **Differences**, Minitab will compute the power for all 6 combinations of sample sizes and differences.

For a discussion of the value needed in **Differences**, see Defining the difference.

3 In **Sigma**, enter an estimate of the population standard deviation (σ) for your data. See Estimating σ.

4 If you like, use one or more of the available dialog box options, then click **OK**.

Defining the minimum difference – Z-test and t-tests

In the main dialog box, you need to specify the minimum difference you are interested in detecting. The manner in which you express this difference depends on whether you are performing a one- or two-sample test:

- For a one-sample Z- or t-test, express the difference in terms of the null hypothesis.

  For example, suppose you are testing whether or not your students' mean test score is different from the population mean. If you would like to detect a difference of three points, you would enter 3 in **Differences**.

- For a two-sample t-test, express the difference as the difference between the population means that you would like to be able to detect.

  For example, suppose you are investigating the effects of water acidity on the growth of two populations of tadpoles. If you are interested in differences of 4 mm or more, you would enter 4 in **Differences**.

If you choose **Less than** as your alternative hypothesis, then you must enter a negative value in **Differences**. If you choose **Greater than**, you must enter a positive value. This is because with a one-tailed test, you have no power to detect an effect in the opposite direction.

Estimating σ

For power or minimum difference calculations, the estimate of σ depends on whether or not you have already collected data.

- Prospective studies are done before collecting data so σ has to be estimated. You can use related research, pilot studies, or subject-matter knowledge to estimate σ.
- Retrospective studies are done after data have been collected so you can use the data to estimate σ.
  - For 1-Sample Z or 1-Sample t, use the standard deviation of the sample.
  - For 2-Sample t, use the pooled standard deviation (Pooled StDev) if assuming equal variances.
  - For One-way ANOVA, 2-Level Factorial Design, and Plackett-Burman Design, use the square root of the mean square error (MS Error).

For sample size calculations, the data have not been collected yet so the population standard deviation (σ) has to be estimated. You can use related research, pilot studies, or subject-matter knowledge to estimate σ.

Note If you would like to specify the difference (effects) in standardized (sigma) units, enter 1 in **Standard deviation**.

Power and Sample Size for 1-Sample t – Options

Stat > Power and Sample Size > 1-Sample t > Options

Use to define the alternative hypothesis, specify the significance level, and store the sample sizes, differences (effects), and power values. When calculating sample size, Minitab stores the power value that will generate the nearest integer sample size.

Dialog box items

Alternative Hypothesis Define the alternative hypothesis.

- **Less than:** Choose to perform a lower-tailed test.
- **Not equal:** Choose to perform two-tailed test.
- **Greater than:** Choose to perform an upper-tailed test.

Significance level: Enter the significance level (α). The default is 0.05.

Store sample sizes in: Enter a storage column for sample sizes.

Store differences in: Enter a storage column for differences.

Store power values in: Enter a storage column for power values.

Example of calculating sample size for a one-sample t-test

Suppose you are the production manager at a dairy plant. In order to meet state requirements, you must maintain strict control over the packaging of ice cream. The volume cannot vary more than 3 oz for a half-gallon (64-oz) container. The packaging machine tolerances are set so the process σ is 1. How many samples must be taken to estimate the mean package volume at a confidence level of 99% (α = .01) for power values of 0.7, 0.8, and 0.9?

1 Choose **Stat > Power and Sample Size > 1-Sample t**.

2 In **Differences**, enter *3*. In **Power values**, enter *0.7 0.8 0.9*.

3 In **Standard deviation**, enter *1*.

4 Click **Options**. In **Significance level**, enter *0.01*. Click **OK** in each dialog box.

Session window output

Power and Sample Size

```
1-Sample t Test

Testing mean = null (versus not = null)
Calculating power for mean = null + difference
Alpha = 0.01  Assumed standard deviation = 1


            Sample  Target
Difference    Size   Power  Actual Power
         3       5     0.7      0.894714
         3       5     0.8      0.894714
         3       6     0.9      0.982651
```

Interpreting the results

Minitab displays the sample size required to obtain the requested power values. Because the target power values would result in non-integer sample sizes, Minitab displays the power (Actual Power) that you would have to detect differences in volume greater than three ounces using the nearest integer value for sample size. If you take a sample of five cartons, power for your test is 0.895; for a sample of six cartons, power is 0.983.

2-Sample t

Power and Sample Size for 2-Sample t

Stat > Power and Sample Size > 2-Sample t

Use to calculate one of the following for a hypothesis tests of the difference in means

- power
- sample size
- minimum difference (effect)

You need to determine what are acceptable values for any two of these parameters and Minitab will solve for the third.

For example, if you specify values for power and the minimum difference, Minitab will determine the sample size required to detect the specified difference at the specified level of power. See Defining the minimum difference.

Dialog box items

Specify values for any two of the following: Specify two values and Minitab will calculate the third.

Sample sizes: Enter one or more sample sizes. Each number you enter is considered to be the sample size for each group.

Differences: Enter one or more differences in terms of the null hypothesis.

Power values: Enter one or more power values.

Standard deviation: Enter an estimate of the population standard deviation (σ) for your data. See Estimating σ.

Tip You can use shorthand notation to enter sample sizes. For example, **10:40/5** means sample sizes from 10 to 40 by 5.

<Options>

To calculate power, sample size, or minimum difference – Z-test and t-tests

1. Choose **Stat > Power and Sample Size > 1-Sample Z**, **1-Sample t**, **or 2-Sample t**.
2. Do one of the following:
   - Solve for power
     1. In **Sample sizes**, enter one or more numbers. For a two-sample test, the number you enter is considered the sample size for each group. For example, if you want to determine power for an analysis with 10 observations in each group for a total of 20, you would enter **10**.
     2. In **Differences**, enter one or more numbers.
   - Solve for sample size
     1. In **Differences**, enter one or more numbers.
     2. In **Power values**, enter one or more numbers.
   - Solve for the minimum difference
     1. In **Sample sizes**, enter one or more numbers. For a two-sample test, the number you enter is considered the sample size for each group.
     2. In **Power values**, enter one or more numbers.

   Minitab will solve for all combinations of the specified values. For example, if you enter 3 values in **Sample sizes** and 2 values in **Differences**, Minitab will compute the power for all 6 combinations of sample sizes and differences.

   For a discussion of the value needed in **Differences**, see Defining the difference.
3. In **Sigma**, enter an estimate of the population standard deviation (σ) for your data. See Estimating σ.
4. If you like, use one or more of the available dialog box options, then click **OK**.

Defining the minimum difference – Z-test and t-tests

In the main dialog box, you need to specify the minimum difference you are interested in detecting. The manner in which you express this difference depends on whether you are performing a one- or two-sample test:

- For a one-sample Z- or t-test, express the difference in terms of the null hypothesis.

  For example, suppose you are testing whether or not your students' mean test score is different from the population mean. If you would like to detect a difference of three points, you would enter 3 in **Differences**.

- For a two-sample t-test, express the difference as the difference between the population means that you would like to be able to detect.

  For example, suppose you are investigating the effects of water acidity on the growth of two populations of tadpoles. If you are interested in differences of 4 mm or more, you would enter 4 in **Differences**.

If you choose **Less than** as your alternative hypothesis, then you must enter a negative value in **Differences**. If you choose **Greater than**, you must enter a positive value. This is because with a one-tailed test, you have no power to detect an effect in the opposite direction.

Estimating σ

For power or minimum difference calculations, the estimate of σ depends on whether or not you have already collected data.

- Prospective studies are done before collecting data so σ has to be estimated. You can use related research, pilot studies, or subject-matter knowledge to estimate σ.
- Retrospective studies are done after data have been collected so you can use the data to estimate σ.
  - For 1-Sample Z or 1-Sample t, use the standard deviation of the sample.
  - For 2-Sample t, use the pooled standard deviation (Pooled StDev) if assuming equal variances.
  - For One-way ANOVA, 2-Level Factorial Design, and Plackett-Burman Design, use the square root of the mean square error (MS Error).

For sample size calculations, the data have not been collected yet so the population standard deviation (σ) has to be estimated. You can use related research, pilot studies, or subject-matter knowledge to estimate σ.

Note If you would like to specify the difference (effects) in standardized (sigma) units, enter 1 in **Standard deviation**.

Power and Sample Size for 2-Sample t – Options

Stat > Power and Sample Size > 2-Sample t > Options

Use to define the alternative hypothesis, specify the significance level, and store the sample sizes, differences (effects), and power values. When calculating sample size, Minitab stores the power value that will generate the nearest integer sample size.

Dialog box items

Alternative Hypothesis Define the alternative hypothesis.

Less than: Choose to perform a lower-tailed test.

Not equal: Choose to perform two-tailed test.

Greater than: Choose to perform an upper-tailed test.

Significance level: Enter the significance level (α). The default significance level is 0.05.

Store sample sizes in: Enter a storage column for sample sizes.

Store differences in: Enter a storage column for differences.

Store power values in: Enter a storage column for power values.

1 Proportion

Power and Sample Size for 1 Proportion

Stat > Power and Sample Size > 1 Proportion

Use to calculate one of the following for a one proportion test

- power
- sample size
- minimum difference (displayed as an alternative proportion value)

You need to determine what are acceptable values for any two of these parameters and Minitab will solve for the third.

For example, if you specify values for power and the minimum difference, Minitab will determine the sample size required to detect the specified difference at the specified level of power.

Dialog box items

Specify values for any two of the following: Specify two values and Minitab will calculate the third.

Sample sizes: Enter one or more sample sizes.

Alternative values of p: Enter one or more proportions. See Defining the minimum difference.

Power values: Enter one or more power values.

Hypothesized p: Enter the expected proportion. The default is 0.5.

Tip You can use shorthand notation to enter sample sizes. For example, 10:40/5 means sample sizes from 10 to 40 by 5.

<Options>

To calculate power, sample size, or minimum difference – Tests of Proportions

1 Choose **Stat > Power and Sample Size > 1 Proportion** or **2 Proportions**.

2 Do one of the following:

- Solve for power

  1 In **Sample sizes**, enter one or more numbers. For a two proportion test, the number you enter is considered the sample size for each group. For example, if you want to determine power for an analysis with 10 observations in each group for a total of 20 observations, you would enter *10*.

  2 In **Alternative values of p** or **Proportion 1 values**, enter one or more proportions.

- Solve for sample size

  1 In **Alternative values of p** or **Proportion 1 values**, enter one or more proportions.

  2 In **Power values**, enter one or more numbers.

- Solve for the minimum difference

  1 In **Sample sizes**, enter one or more numbers. For a two proportion test, the number you enter is considered the sample size for each group, not the total number for the experiment.

  2 In **Power values**, enter one or more numbers.

Minitab will solve for all combinations of the specified values. For example, if you enter 3 values in **Sample sizes** and 2 values in **Alternative values of p**, Minitab will compute the power for all 6 combinations of sample sizes and alternative proportions.

For a discussion of the values needed in **Alternative values of p** and **Proportion 1 values**, see Defining the difference.

3 Do one of the following:

- For a one-sample test, enter the expected proportion under the null hypothesis in **Hypothesized p**. The default is 0.5.
- For a two-sample test, enter the second proportion in **Proportion 2**. The default is 0.5.

For a discussion of the values needed in **Hypothesized p** and **Proportion 2**, see Defining the difference.

4 If you like, use one or more of the dialog box options, then click **OK**.

Defining the Difference – Tests of Proportions

Minitab uses two proportions to determine the minimum difference. The manner in which you express these proportions depends on whether you are performing a one- or two-sample proportion test.

- For a one-sample test of proportion, enter the expected proportion under the null hypothesis for **Hypothesized p** in the dialog box.

  Suppose you are testing whether the data are consistent with the following null hypothesis and would like to detect any differences where the true proportion is greater than 0.73.

  H_0: p = 0.7 H_1: p > 0.7 where p is the population proportion

  In Minitab, enter 0.73 in **Alternative values of p**; enter 0.7 in **Hypothesized p**. (The alternative proportion is not the value of the alternative hypothesis, but the value at which you want to evaluate power.)

- For a two-sample test of proportion, enter the expected proportions under the null hypothesis for **Proportion 2** in the dialog box.

  Suppose a biologist wants to test for a difference in the proportion of fish affected by pollution in two lakes. The biologist would like to detect a difference of 0.03 and suspects that about one quarter (0.25) of fish in Lake A have been affected.

  H_0: $p_1 = p_2$ H_1: $p_1 \neq p_2$

  In Minitab, enter 0.22 and 0.28 in **Proportion 1 values**; enter 0.25 in **Proportion 2**.

Power and Sample Size for 1 Proportion – Options

Stat > Power and Sample Size > 1 Proportion > Options

Use to define the alternative hypothesis, specify the significance level of the test, and store the sample sizes, alternate values of p, and power values. When calculating sample size, Minitab stores the power value that will generate the nearest integer sample size.

Dialog box items

Alternative Hypothesis Define the alternative hypothesis.

Less than: Choose to perform a lower-tailed test.

Not equal: Choose to perform two-tailed test.

Greater than: Choose to perform an upper-tailed test.

Significance level: Enter the significance level (α). The default is 0.05.

Store sample sizes in: Enter a storage column for sample sizes.

Store alternatives in: Enter a storage column for alternative values of p.

Store power values in: Enter a storage column for power values.

2 Proportions

Power and Sample Size for 2 Proportions

Stat > Power and Sample Size > 2 Proportions

Use to calculate one of the following for a hypothesis tests of the difference in proportions

- power
- sample size
- minimum difference (displayed as a proportion 1 value)

You need to determine what are acceptable values for any two of these parameters and Minitab will solve for the third.

For example, if you specify values for power and the minimum difference, Minitab will determine the sample size required to detect the specified difference at the specified level of power.

Dialog box items

Specify values for any two of the following: Specify two values and Minitab will calculate the third.

Sample sizes: Enter one or more sample sizes. For a two proportion test, the number you enter is considered to be the sample size for each group.

Proportion 1 values: Enter one or more values for proportion 1. See Defining the minimum difference.

Power values: Enter one or more power values.

Proportion 2: Enter the second proportion. The default is 0.5.

Tip You can use shorthand notation to enter sample sizes. For example, 10:40/5 means sample sizes from 10 to 40 by 5.

<Options>

To calculate power, sample size, or minimum difference – Tests of Proportions

1 Choose **Stat > Power and Sample Size > 1 Proportion** or **2 Proportions**.

2 Do one of the following:

- Solve for power

  1 In **Sample sizes**, enter one or more numbers. For a two proportion test, the number you enter is considered the sample size for each group. For example, if you want to determine power for an analysis with 10 observations in each group for a total of 20 observations, you would enter *10*.

  2 In **Alternative values of p** or **Proportion 1 values**, enter one or more proportions.

- Solve for sample size

  1 In **Alternative values of p** or **Proportion 1 values**, enter one or more proportions.

  2 In **Power values**, enter one or more numbers.

- Solve for the minimum difference

  1 In **Sample sizes**, enter one or more numbers. For a two proportion test, the number you enter is considered the sample size for each group, not the total number for the experiment.

  2 In **Power values**, enter one or more numbers.

Minitab will solve for all combinations of the specified values. For example, if you enter 3 values in **Sample sizes** and 2 values in **Alternative values of p**, Minitab will compute the power for all 6 combinations of sample sizes and alternative proportions.

For a discussion of the values needed in **Alternative values of p** and **Proportion 1 values**, see Defining the difference.

3 Do one of the following:

- For a one-sample test, enter the expected proportion under the null hypothesis in **Hypothesized p**. The default is 0.5.
- For a two-sample test, enter the second proportion in **Proportion 2**. The default is 0.5.

For a discussion of the values needed in **Hypothesized p** and **Proportion 2**, see Defining the difference.

4 If you like, use one or more of the dialog box options, then click **OK**.

Defining the Difference – Tests of Proportions

Minitab uses two proportions to determine the minimum difference. The manner in which you express these proportions depends on whether you are performing a one- or two-sample proportion test.

- For a one-sample test of proportion, enter the expected proportion under the null hypothesis for **Hypothesized p** in the dialog box.

  Suppose you are testing whether the data are consistent with the following null hypothesis and would like to detect any differences where the true proportion is greater than 0.73.

  H_0: p = 0.7 H_1: p > 0.7   where p is the population proportion

  In Minitab, enter 0.73 in **Alternative values of p**; enter 0.7 in **Hypothesized p**. (The alternative proportion is not the value of the alternative hypothesis, but the value at which you want to evaluate power.)

- For a two-sample test of proportion, enter the expected proportions under the null hypothesis for **Proportion 2** in the dialog box.

  Suppose a biologist wants to test for a difference in the proportion of fish affected by pollution in two lakes. The biologist would like to detect a difference of 0.03 and suspects that about one quarter (0.25) of fish in Lake A have been affected.

  H_0: $p_1 = p_2$ H_1: $p_1 \neq p_2$

  In Minitab, enter 0.22 and 0.28 in **Proportion 1 values**; enter 0.25 in **Proportion 2**.

Power and Sample Size for 2 Proportions – Options

Stat > Power and Sample Size > 2 Proportion > Options

Use to define the alternative hypothesis, specify the significance level of the test, and store the sample sizes or alternate p values and power values.

Dialog box items

Alternative Hypothesis Define the alternative hypothesis.

Less than: Choose to perform a lower-tailed test.

Not equal: Choose to perform two-tailed test.

Greater than: Choose to perform an upper-tailed test.

Significance level: Enter the significance level (α) of the test. The default is 0.05.

Store sample sizes in: Enter a storage column for the sample sizes.

Store proportion 1 in: Enter a storage column for the proportion 1 values.

Store power values in: Enter a storage column for the power values.

Example of calculating power for a two-sample test of proportion

As a political advisor, you want to determine whether there is a difference between the proportions of men and women who support a tax reform bill. Results of a previous survey suggest that 30% (p = 0.30) of the voters in general support the bill. If you mail 1000 surveys to voters of each gender, what is the power to detect the difference if men and women in the general population differ in support for the bill by 5% (0.05) or more?

1 Choose **Stat > Power and Sample Size > 2 Proportions**.

2 In **Sample sizes**, enter *1000*.

3 In **Proportion 1 values**, enter *0.25 0.35*.

4 In **Proportion 2**, enter *0.30*. Click **OK**.

Session window output

Power and Sample Size

```
Test for Two Proportions

Testing proportion 1 = proportion 2 (versus not =)
Calculating power for proportion 2 = 0.3
Alpha = 0.05


              Sample
Proportion 1    Size     Power
        0.25    1000  0.707060
        0.35    1000  0.665570

The sample size is for each group.
```

Interpreting the results

If 30% (0.30) of one gender support the bill and only 25% (0.25) of the other does, you'll have a 71% chance of detecting a difference if you send out 1000 surveys to each. If the population proportions are actually 0.30 and 0.35, you'll have a 67% chance of detecting a difference.

One-way ANOVA

Power and Sample Size for One-Way ANOVA

Stat > Power and Sample Size > One-Way ANOVA

Use to calculate one of the following for a test of the equality of population means

- power
- sample size
- minimum detectable difference between the smallest and largest factor means (maximum difference)

You need to determine what are acceptable values for any two of these parameters and Minitab will solve for the third.

For example, if you specify values for power and the maximum difference between the factor level means, Minitab will determine the sample size required to detect the specified difference at the specified level of power. See Defining the maximum difference.

Dialog box items

Number of levels: Enter the number of factor levels (treatment conditions).

Specify values for any two of the following: Specify two values and Minitab will calculate the third.

> **Sample sizes:** Enter one or more sample sizes. Each number you enter is considered the number of observations for every factor level. For example, if you have 3 factor levels with 5 observations each, you would enter 5.
>
> **Values of the maximum difference between means:** Enter one or more values for the maximum difference between means.
>
> **Power values:** Enter one or more power values.

Standard deviation: Enter an estimate of the population standard deviation (σ) for your data. See Estimating σ.

Tip You can use shorthand notation to enter sample sizes. For example, 10:40/5 means sample sizes from 10 to 40 by 5.

<Options>

Power and Sample Size for One-Way ANOVA – Options

Stat > Power and Sample Size > One-Way ANOVA > Options

Use to specify the significance level of the test and store the sample sizes, sums of squares, and power values.

Dialog box items

Significance level: Specify the significance level of the test. The default significance level is $\alpha = 0.05$.

Store sample sizes in: Enter a storage column for sample sizes.

Store sums of squares in: Enter a storage column for sums of squares.

Store power values in: Enter a storage column for power values. When calculating sample size, Minitab stores the power value associated with the nearest integer sample size.

To calculate power, sample size, or maximum difference – One-Way ANOVA

1 Choose **Stat > Power and Sample Size > One-Way ANOVA**.

2 In **Number of levels**, enter the number of factor levels (treatment conditions).

3 Do one of the following:

- Solve for power

  1 In **Sample sizes**, enter one or more numbers. Each number you enter is considered the number of observations in every factor level. For example, if you have 3 factor levels with 5 observations each, you would enter 5.

  2 In **Values of the maximum difference between means**, enter one or more numbers.

- Solve for sample size

  1 In **Values of the maximum difference between means**, enter one or more numbers.

  2 In **Power values**, enter one or more numbers.

- Solve for the maximum difference

  1 In **Sample sizes**, enter one or more numbers. Each number you enter is considered the number of observations in every factor level.

  2 In **Power values**, enter one or more numbers.

Minitab will solve for all combinations of the specified values. For example, if you enter 3 values in **Sample sizes** and 2 values in **Values of the maximum difference between means**, Minitab will compute the power for all 6 combinations of sample sizes and maximum differences. See Defining the maximum difference.

4 In **Standard deviation**, enter an estimate of the population standard deviation (σ) for your data. See Estimating σ.

5 If you like, use one or more of the available dialog box options, then click **OK**.

Defining the maximum difference – One-Way Analysis of Variance

In order to calculate power or sample size, you need to estimate the maximum difference between the smallest and largest actual factor level means. For example, suppose you are planning an experiment with four treatment conditions (four factor levels). You want to find a difference between a control group mean of 10 and a level mean that is 15. In this case, the maximum difference between the means is 5.

Estimating σ

For power or minimum difference calculations, the estimate of σ depends on whether or not you have already collected data.

- Prospective studies are done before collecting data so σ has to be estimated. You can use related research, pilot studies, or subject-matter knowledge to estimate σ.
- Retrospective studies are done after data have been collected so you can use the data to estimate σ.
  - For 1-Sample Z or 1-Sample t, use the standard deviation of the sample.
  - For 2-Sample t, use the pooled standard deviation (Pooled StDev) if assuming equal variances.
  - For One-way ANOVA, 2-Level Factorial Design, and Plackett-Burman Design, use the square root of the mean square error (MS Error).

For sample size calculations, the data have not been collected yet so the population standard deviation (σ) has to be estimated. You can use related research, pilot studies, or subject-matter knowledge to estimate σ.

Note If you would like to specify the difference (effects) in standardized (sigma) units, enter 1 in **Standard deviation**.

Example of calculating power for a one-way ANOVA

Suppose you are about to undertake an investigation to determine whether or not 4 treatments affect the yield of a product using 5 observations per treatment. You know that the mean of the control group should be around 8, and you would like to find significant differences of +4. Thus, the maximum difference you are considering is 4 units. Previous research suggests the population σ is 1.64.

1 Choose **Stat > Power and Sample Size > One-way ANOVA**.

2. In **Number of levels**, enter *4*.
3. In **Sample sizes**, enter *5*.
4. In **Values of the maximum difference between means**, enter *4*.
5. In **Standard deviation**, enter *1.64*. Click **OK**.

Session window output

Power and Sample Size

```
One-way ANOVA

Alpha = 0.05  Assumed standard deviation = 1.64  Number of Levels = 4


   SS  Sample               Maximum
Means    Size      Power  Difference
    8       5   0.826860           4

The sample size is for each level.
```

Interpreting the results

If you assign five observations to each treatment level, you have a power of 0.83 to detect a difference of 4 units or more between the treatment means.

2-Level Factorial Design

Power and Sample Size for 2-Level Factorial Design

Stat > Power and Sample Size >2-Level Factorial Design

Use to calculate one of the following for two-level full and fractional factorial designs and Plackett-Burman designs

- number of replicates
- power
- minimum effect
- number of center points

You need to determine what are acceptable values for any three of these parameters and Minitab will solve for the fourth.

For example, if you specify values for power, minimum effect, and number of center points, Minitab will determine the number of replicates required to detect the specified effect at the specified level of power.

Dialog box items

Number of factors: Enter the number of factors (input variables).

Number of corner points: Enter the number of corner points.

Specify values for any three of the following: Specify three values and Minitab will calculate the fourth.

Replicates: Enter one or more numbers of replicates.

Effects: Enter the effect(s).

Power values: Enter one or more power values.

Number of center points (per block): Enter one or more values for the number of center points.

Standard deviation: Enter an estimate of the population standard deviation (σ) for your data. See Estimating σ.

<Design>

<Options>

To calculate power, replicates, minimum effect, or number of center points – Two-Level Factorials and Plackett-Burman Designs

1. Choose **Stat** > **Power and Sample Size** > **2-Level Factorial Design** or **Plackett-Burman Design**.
2. In **Number of factors**, enter the number of factors (input variables).
3. In **Number of corner points**, enter a number. See Determining the number of corner points.

4 Do one of the following:

- Solve for power
  1 In **Replicates**, enter one or more numbers.
  2 In **Effects**, enter one or more numbers.
  3 In **Number of center points**, enter one or more numbers.
- Solve for the number of replicates
  1 In **Effects**, enter one or more numbers.
  2 In **Power values**, enter one or more numbers.
  3 In **Number of center points**, enter one or more numbers.
- Solve for the minimum effect
  1 In **Replicates**, enter one or more numbers.
  2 In **Power values**, enter one or more numbers.
  3 In **Number of center points**, enter one or more numbers.
- Solve the number of center points
  1 In **Replicates**, enter one or more numbers.
  2 In **Effects**, enter one or more numbers.
  3 In **Power values**, enter one or more numbers.

For information on the value needed in **Effects**, see Defining the effect. For information on the value needed in **Replicates**, see Determining the number of replicates.

5 In **Standard deviation**, enter an estimate of the population standard deviation (σ) for your data. See Estimating σ.

6 If you like, use one or more of the available dialog box options, then click **OK**.

Defining the effect – Two-Level Factorial and Plackett-Burman Designs

When calculating power or number of replicates, you need to specify the minimum effect you are interested in detecting. You express this effect as the difference between the low and high factor level means. For example, suppose you are trying to determine the effect of column temperature on the purity of your product. You are only interested in detecting a difference in purity that is greater than 0.007 between the low and high levels of temperature. In the dialog box, enter 0.007 in **Effects**.

Determining the Number of Corner Points – Two-Level Factorial and Plackett-Burman Designs

For all designs, you need to specify the appropriate number of corner points given the number of factors. For example, for a 6 factor full factorial design you would have 64 corner points. However, for a 6 factor fractional factorial design, you can have either 8, 16, or 32 corner points. Use the information provided in Summary of Two-Level Designs to determine the correct number of corner points for your design.

Determining the Number of Replicates – Two-Level Factorial and Plackett-Burman Designs

Rather than using sample size to indicate the number of observations you need, factorial designs are expressed in terms of the number of replicates. A **replicate** is a repeat of each of the design points (experimental conditions) in the base design. For example, if you are doing a full factorial with three factors, one replicate would require eight corner points. The base design would include all combinations of the low and high levels for all factors. Each time you replicate the design eight runs are added; these runs are duplicates of the original eight corner points.

For a discussion of replication, see Replicating the design. For a discussion of two-level factorial and Plackett-Burman designs, see Factorial Designs Overview.

Estimating σ

For power or minimum difference calculations, the estimate of σ depends on whether or not you have already collected data.

- Prospective studies are done before collecting data so σ has to be estimated. You can use related research, pilot studies, or subject-matter knowledge to estimate σ.

- Retrospective studies are done after data have been collected so you can use the data to estimate σ.
  - For 1-Sample Z or 1-Sample t, use the standard deviation of the sample.
  - For 2-Sample t, use the pooled standard deviation (Pooled StDev) if assuming equal variances.
  - For One-way ANOVA, 2-Level Factorial Design, and Plackett-Burman Design, use the square root of the mean square error (MS Error).

For sample size calculations, the data have not been collected yet so the population standard deviation (σ) has to be estimated. You can use related research, pilot studies, or subject-matter knowledge to estimate σ.

Note If you would like to specify the difference (effects) in standardized (sigma) units, enter 1 in **Standard deviation**.

Number of terms in the saturated model – 2-level factorial designs

Unless you choose to omit terms from your model, Minitab's power calculations are based on a model that uses all of the available degrees of freedom, otherwise known as a saturated model. Use the table below to determine the number of terms in the saturated model for a 2-level factorial design with no blocks.

Model	Number of terms in saturated model
Without center points or blocks	= number of unique corner points in base design – 1
Includes center points but no blocks	= number of unique corner points in base design

When the model contains blocks, the number of terms in the saturated model depends on the number of degrees of freedom (df) remaining after accounting for blocks (df = number of blocks - 1) and center points (1 df).

The simplest way to determine the number of terms in the saturated model is to create the design and count the number of terms in the resulting model as follows:

1 Choose **Stat > DOE > Factorial > Create Factorial Design** and create your design.

2 Go to the Session window and count the number of terms displayed in the Alias Structure. Do not count the identity term (I).
   - If you have no center points, then this is the number of terms in the saturated model for the design you created.
   - If you have one or more center points, then the number of terms in the saturated model is 1 greater than this.

Determining the Number of Blocks – Two-level Factorial Designs

The number of blocks in your design must divide evenly into the number of corner points times the number of replicates. If the number of blocks in your model does not meet this criteria, Minitab does one of the following:

- If you are solving for power or effect size, Minitab displays an error message about the wrong number of blocks.
- If you are solving for number of replicates, Minitab increases the number of replicates until the number of corner points times the number of replicates is a multiple of the number of blocks and you may end up with more runs than expected. A better approach is to start with one block and add additional blocks based on the number of replicates you need to achieve the desired power.

Use the following approach to determine the number of blocks when you are solving for number of replicates.

1 Start with one block and calculate how many replicates of the base design you need to get the desired power.

2 Add blocks based on the number of runs you have. You must be able to evenly divide the number of corner points times the number of replicates by the number of blocks.

3 Rerun the power analysis to see how the additional blocks affect your power. Including additional blocks in your model may reduce your power by taking away degrees of freedom from the error term. You can increase power by adding center points to your model.

Power and Sample Size for 2-Level Factorial Design – Design

Stat > Power and Sample Size > 2-Level Factorial Design > Design

Use to specify your design.

Dialog box items

Number of blocks: Enter the number of blocks in your design. (For more information, see Determining the number of blocks.)

Number of terms omitted from model: Enter the number of terms you are going to omit from the saturated model.

Include term for center points in model: Check if you are going to fit a term for the center points.

Include blocks in model: Check if you are going to include blocks in the model.

Power and Sample Size for 2-Level Factorial Design – Options

Stat > Power and Sample Size > 2-Level Factorial Design > Options

Use to specify the significance level of the test and store the replicates, effects, power values, and center points.

Dialog box items

Significance level: Enter the significance level of the test. The default is α = 0.05.

Store replicates in: Enter a storage column for number of replicates.

Store effects in: Enter a storage column for number of effects.

Store power values in: Enter a storage column for power values. When calculating the number of replicates, Minitab stores the power value associated with the nearest integer number of replicates.

Store center points in: Enter a storage column for number of replicates.

Example of Calculating Power for a Two-Level Fractional Factorial Design

As a quality engineer, you need to determine the "best" settings for 4 input variables (factors) to improve the transparency of a plastic part. You have determined that a 4 factor, 8 run design (1/2 fraction) with 3 center points will allow you to estimate the effects you are interested in. Although you would like to perform as few replicates as possible, you must be able to detect effects with magnitude of 5 or more. Previous experimentation suggests that 4.5 is a reasonable estimate of σ.

1 Choose **Stat > Power and Sample Size > 2-Level Factorial Design**.

2 In **Number of factors**, enter *4*.

3 In **Number of corner points**, enter *8*.

4 In **Replicates**, enter *1 2 3 4*.

5 In **Effects**, enter *5*.

6 In **Number of center points**, enter *3*.

7 In **Standard deviation**, enter *4.5*. Click **OK**.

Session window output

Power and Sample Size

```
2-Level Factorial Design

Alpha = 0.05  Assumed standard deviation = 4.5

Factors:     4   Base Design: 4, 8
Blocks:   none

Including a term for center points in model.


Center                Total
Points  Effect  Reps   Runs     Power
     3       5     1     11  0.157738
     3       5     2     19  0.518929
     3       5     3     27  0.730495
     3       5     4     35  0.856508
```

Interpreting the results

If you do not replicate your design (Reps = 1), you will only have a 16% chance of detecting effects that you have determined are important. If you use 4 replicates of your 1/2 fraction design for a total 35 runs (32 corner points and 3 center points), you will have an 86% chance of finding important effects.

Plackett-Burman Design

Power and Sample Size for Plackett-Burman Design

Stat > Power and Sample Size > Plackett-Burman Design

Use to calculate power and number of replicates for two-level full and fractional Plackett-Burman designs. To calculate power or replicates for fractional designs, you need to specify the appropriate number of corner points for the number of factors. Use the information provided in Summary of Two-Level Designs to determine the correct number of corner points for these fractional designs.

To calculate power or replicates for these tests, you need to determine the minimum **effect** that you consider to be important. Then, you can determine the power or the number of replicates you need to be able to reject the null hypothesis when the true value differs from the hypothesized value by this minimum effect.

Dialog box items

Number of factors: Enter the number of factors (input variables).

Number of corner points: Enter the number of corner points.

Specify values for any three of the following: Specify three values and Minitab will calculate the fourth.

Replicates: Enter one or more numbers of replicates.

Effects: Enter the effect(s).

Power values: Enter one or more power values.

Number of center points: Enter one or more values for the number of center points.

Standard deviation: Enter an estimate of the population standard deviation (σ) for your data. See Estimating σ.

<Design>

<Options>

To calculate power, replicates, minimum effect, or number of center points – Two-Level Factorials and Plackett-Burman Designs

1. Choose **Stat** > **Power and Sample Size** > **2-Level Factorial Design** or **Plackett-Burman Design**.
2. In **Number of factors**, enter the number of factors (input variables).
3. In **Number of corner points**, enter a number. See Determining the number of corner points.
4. Do one of the following:
   - Solve for power
     1. In **Replicates**, enter one or more numbers.
     2. In **Effects**, enter one or more numbers.
     3. In **Number of center points**, enter one or more numbers.
   - Solve for the number of replicates
     1. In **Effects**, enter one or more numbers.
     2. In **Power values**, enter one or more numbers.
     3. In **Number of center points**, enter one or more numbers.
   - Solve for the minimum effect
     1. In **Replicates**, enter one or more numbers.
     2. In **Power values**, enter one or more numbers.
     3. In **Number of center points**, enter one or more numbers.
   - Solve the number of center points
     1. In **Replicates**, enter one or more numbers.
     2. In **Effects**, enter one or more numbers.
     3. In **Power values**, enter one or more numbers.

   For information on the value needed in **Effects**, see Defining the effect. For information on the value needed in **Replicates**, see Determining the number of replicates.
5. In **Standard deviation**, enter an estimate of the population standard deviation (σ) for your data. See Estimating σ.
6. If you like, use one or more of the available dialog box options, then click **OK**.

Defining the effect – Two-Level Factorial and Plackett-Burman Designs

When calculating power or number of replicates, you need to specify the minimum effect you are interested in detecting. You express this effect as the difference between the low and high factor level means. For example, suppose you are trying to determine the effect of column temperature on the purity of your product. You are only interested in detecting a difference in purity that is greater than 0.007 between the low and high levels of temperature. In the dialog box, enter 0.007 in **Effects**.

Determining the Number of Corner Points – Two-Level Factorial and Plackett-Burman Designs

For all designs, you need to specify the appropriate number of corner points given the number of factors. For example, for a 6 factor full factorial design you would have 64 corner points. However, for a 6 factor fractional factorial design, you can have either 8, 16, or 32 corner points. Use the information provided in Summary of Two-Level Designs to determine the correct number of corner points for your design.

Determining the Number of Replicates – Two-Level Factorial and Plackett-Burman Designs

Rather than using sample size to indicate the number of observations you need, factorial designs are expressed in terms of the number of replicates. A **replicate** is a repeat of each of the design points (experimental conditions) in the base design. For example, if you are doing a full factorial with three factors, one replicate would require eight corner points. The base design would include all combinations of the low and high levels for all factors. Each time you replicate the design eight runs are added; these runs are duplicates of the original eight corner points.

For a discussion of replication, see Replicating the design. For a discussion of two-level factorial and Plackett-Burman designs, see Factorial Designs Overview.

Estimating σ

For power or minimum difference calculations, the estimate of σ depends on whether or not you have already collected data.

- Prospective studies are done before collecting data so σ has to be estimated. You can use related research, pilot studies, or subject-matter knowledge to estimate σ.
- Retrospective studies are done after data have been collected so you can use the data to estimate σ.
  - For 1-Sample Z or 1-Sample t, use the standard deviation of the sample.
  - For 2-Sample t, use the pooled standard deviation (Pooled StDev) if assuming equal variances.
  - For One-way ANOVA, 2-Level Factorial Design, and Plackett-Burman Design, use the square root of the mean square error (MS Error).

For sample size calculations, the data have not been collected yet so the population standard deviation (σ) has to be estimated. You can use related research, pilot studies, or subject-matter knowledge to estimate σ.

Note If you would like to specify the difference (effects) in standardized (sigma) units, enter 1 in **Standard deviation**.

Number of terms in the saturated model – Plackett-Burman designs

Unless you choose to omit terms from your model, Minitab's power calculations are based on a model that uses all of the available degrees of freedom, otherwise known as a saturated model. Use the table below to determine the number of terms in the saturated model for a Plackett-Burman design.

Model	Number of terms in saturated model
Without center points	= number of unique corner points in base design – 1
Includes center points	= number of unique corner points in base design

Power and Sample Size for Plackett-Burman Design – Design

Stat > Power and Sample Size >Plackett-Burman > Design

Use to specify your design.

Dialog box items

Number of terms omitted from model: Subtract the number of terms you are analyzing from the number of terms in the saturated model and enter this value.

Include term for center points in model: Check if you are going to fit a term for the center points.

Power and Sample Size for Plackett-Burman Design - Options

Stat > Power and Sample Size > Plackett-Burman Design > Options

Use to specify the significance level of the test and store the replicates, effects, power values, and center points.

Dialog box items

Significance level: Enter the significance level of the test. The default is $\alpha = 0.05$.

Store replicates in: Enter a storage column for number of replicates.

Store effects in: Enter a storage column for number of effects.

Store power values in: Enter a storage column for power values. When calculating the number of replicates, Minitab stores the power value associated with the nearest integer number of replicates.

Store center points in: Enter a storage column for number of replicates.

Index