# Statistics Toolbox™ Release Notes

**How to Contact MathWorks**

| | |
|---|---|
| www.mathworks.com | Web |
| comp.soft-sys.matlab | Newsgroup |
| www.mathworks.com/contact_TS.html | Technical Support |

| | |
|---|---|
| suggest@mathworks.com | Product enhancement suggestions |
| bugs@mathworks.com | Bug reports |
| doc@mathworks.com | Documentation error reports |
| service@mathworks.com | Order status, license renewals, passcodes |
| info@mathworks.com | Sales, pricing, and general information |

508-647-7000 (Phone)

508-647-7001 (Fax)

The MathWorks, Inc.
3 Apple Hill Drive
Natick, MA 01760-2098

For contact information about worldwide offices, see the MathWorks Web site.

*Statistics Toolbox™ Release Notes*

© COPYRIGHT 2005–2011 by The MathWorks, Inc.

**Trademarks**

MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See www.mathworks.com/trademarks for a list of additional trademarks. Other product or brand names may be trademarks or registered trademarks of their respective holders.

**Patents**

MathWorks products are protected by one or more U.S. patents. Please see www.mathworks.com/patents for more information.

# Contents

# Summary by Version

This table provides quick access to what's new in each version. For clarification, see "Using Release Notes" on page 2.

| Version (Release) | New Features and Changes | Version Compatibility Considerations | Fixed Bugs and Known Problems |
|---|---|---|---|
| **Latest Version V7.5 (R2011a)** | Yes Details | No | Bug Reports Includes fixes |
| V7.4 (R2010b) | Yes Details | Yes Summary | Bug Reports Includes fixes |
| V7.3 (R2010a) | Yes Details | No | Bug Reports Includes fixes |
| V7.2 (R2009b) | Yes Details | No | Bug Reports Includes fixes |
| V7.1 (R2009a) | Yes Details | No | Bug Reports Includes fixes |
| V7.0 (R2008b) | Yes Details | Yes Summary | No |
| V6.2 (R2008a) | Yes Details | Yes Summary | Bug Reports Includes fixes |
| V6.1 (R2007b) | Yes Details | Yes Summary | Bug Reports Includes fixes |
| V6.0 (R2007a) | Yes Details | Yes Summary | Bug Reports Includes fixes |
| V5.3 (R2006b) | Yes Details | Yes Summary | Bug Reports Includes fixes |
| V5.2 (R2006a) | Yes Details | No | Bug Reports Includes fixes |
| V5.1 (R14SP3) | Yes Details | No | No |
| V5.0.2 (R14SP2) | Yes Details | No | Bug Reports Includes fixes |

## Using Release Notes

Use release notes when upgrading to a newer version to learn about:

- New features

- Changes

- Potential impact on your existing files and practices

Review the release notes for other MathWorks® products required for this product (for example, MATLAB® or Simulink®). Determine if enhancements, bugs, or compatibility considerations in other products impact you.

If you are upgrading from a software version other than the most recent one, review the current release notes and all interim versions. For example, when you upgrade from V1.0 to V1.2, review the release notes for V1.1 and V1.2.

## What Is in the Release Notes

### New Features and Changes

- New functionality

- Changes to existing functionality

### Version Compatibility Considerations

When a new feature or change introduces a reported incompatibility between versions, the **Compatibility Considerations** subsection explains the impact.

Compatibility issues reported after the product release appear under Bug Reports at the MathWorks Web site. Bug fixes can sometimes result in incompatibilities, so review the fixed bugs in Bug Reports for any compatibility impact.

### Fixed Bugs and Known Problems

MathWorks offers a user-searchable Bug Reports database so you can view Bug Reports. The development team updates this database at release time

and as more information becomes available. Bug Reports include provisions for any known workarounds or file replacements. Information is available for bugs existing in or fixed in Release 14SP2 or later. Information is not available for all bugs in earlier releases.

Access Bug Reports using your MathWorks Account.

## Documentation on the MathWorks Web Site

Related documentation is available on mathworks.com for the latest release and for previous releases:

- Latest product documentation
- Archived documentation

# Version 7.5 (R2011a) Statistics Toolbox Software

This table summarizes what's new in Version 7.5 (R2011a):

| New Features and Changes | Version Compatibility Considerations | Fixed Bugs and Known Problems |
|---|---|---|
| Yes<br>Details below | No | Bug Reports<br>Includes fixes |

New features and changes introduced in this version are:

- "Boosted Decision Trees for Classification and Regression" on page 4

- "Memory and Performance Improvements in Linkage Methods" on page 5

- "Conditional Weighted Residuals and Derivative Step Control in nlmefit and nlmefitsa" on page 5

- "Detecting Ties in k-Nearest Neighbor Search" on page 5

- "Distribution Fitting Tool Uses fitdist Function" on page 5

- "Speed and Accuracy Improvements in Noncentral Chi-Square CDF" on page 6

- "Perfect Separation in Binomial Regression" on page 6

- "Sign Convention in mdscale" on page 6

- "Demo of Credit Rating Classification Via Bagged Decision Trees" on page 6

## Boosted Decision Trees for Classification and Regression

The new `fitensemble` function constructs ensembles of decision trees. It provides:

- Several popular boosting algorithms (`AdaBoostM1`, `AdaBoostM2`, `GentleBoost`, `LogitBoost`, and `RobustBoost`) for classification

- Least-squares boosting (`LSBoost`) for regression

- Most `TreeBagger` functionality for ensembles of bagged decision trees

There is also an improved interface for classification trees (`ClassificationTree`) and regression trees (`RegressionTree`), encompassing the functionality of `classregtree`.

For details, see "Ensemble Methods".

## Memory and Performance Improvements in Linkage Methods

The `linkage` and `clusterdata` functions have a new `savememory` option that can use less memory than before. With `savememory` set to `'on'`, the functions do not build a pairwise distance matrix, so use less memory and, depending on problem size, can use less time. You can use the `savememory` option when:

- The linkage `method` is `'ward'`, `'centroid'`, or `'median'`
- The linkage distance `metric` is `'euclidean'` (default)

For details, see the `linkage` and `clusterdata` function reference pages.

## Conditional Weighted Residuals and Derivative Step Control in nlmefit and nlmefitsa

The `nlmefit` and `nlmefitsa` functions now provide the conditional weighted residuals of the fit. Use this information to assess the quality of the model; see "Example: Examining Residuals for Model Verification".

The `statset` Options structure now includes `'DerivStep'`, which enables you to set finite differences for gradient estimation.

## Detecting Ties in k-Nearest Neighbor Search

`knnsearch` now optionally returns all $k$th nearest neighbors of points, instead of just one. The `knnsearch` methods for `ExhaustiveSearcher` and `KDTreeSearcher` also have this option.

## Distribution Fitting Tool Uses fitdist Function

MATLAB functions generated with the Distribution Fitting Tool now use the `fitdist` function to create fitted probability distribution objects.

The generated functions return probability distribution objects as output arguments.

## Speed and Accuracy Improvements in Noncentral Chi-Square CDF

`ncx2cdf` is now faster and more accurate for large values of the noncentrality parameter.

## Perfect Separation in Binomial Regression

If the two categories in a binomial regression model (such as `logit` or `probit`) are perfectly separated, the best-fitting model is degenerate with infinite coefficients. In this case, the `glmfit` function is likely to exceed its iteration limit. `glmfit` now tries to detect this perfect separation and display a diagnostic message.

## Sign Convention in mdscale

`mdscale` now enforces that, in each column of the output Y, the value with the largest magnitude has a positive sign. This change makes results consistent across releases and platforms—small changes used to lead to sign reversals.

## Demo of Credit Rating Classification Via Bagged Decision Trees

The credit rating demo that used to be exclusively in Financial Toolbox™ is now available in Statistics Toolbox™. The demo uses bagged decision trees for classifying creditworthiness.

To view the demo at the MATLAB command line, enter:

```
showdemo creditratingdemo
```

# Version 7.4 (R2010b) Statistics Toolbox Software

This table summarizes what's new in Version 7.4 (R2010b):

| New Features and Changes | Version Compatibility Considerations | Fixed Bugs and Known Problems |
|---|---|---|
| Yes<br>Details below | Yes<br>Summary | Bug Reports<br>Includes fixes |

New features and changes introduced in this version are:

- "Parallel Computing Support for More Functions" on page 7
- "Algorithm to Rank Features in Classification and Regression" on page 8
- "nlmefit Support for Error Models, and nlmefitsa changes" on page 8
- "Surrogate Splits for Decision Trees" on page 9
- "New Bagged Decision Tree Properties" on page 9
- "Enhanced Cluster Analysis Performance" on page 9
- "Export Probability Objects with dfittool" on page 9
- "Compute Partial Correlation of Two Variables Correcting for All Other Variables" on page 10
- "Specify Number of Evenly Spaced Quantiles" on page 10
- "Control Location and Orientation of Marginal Histograms with scatterhist" on page 10
- "Return Bootstrapped Statistics with bootci" on page 10

## Parallel Computing Support for More Functions

Statistics Toolbox now supports parallel execution for the following functions:

- candexch
- cordexch
- daugment
- dcovary

- nnmf

- plsregress

- rowexch

- sequentialfs

For more information, see the "Parallel Statistics" chapter in the User's Guide.

## Algorithm to Rank Features in Classification and Regression

New filter algorithm, relieff, is based on nearest neighbors. The ReliefF algorithm accounts for correlations among predictors by computing the effect of every predictor on the class label (or true response for regression) locally and then integrates these local estimates over the entire predictor space.

## nlmefit Support for Error Models, and nlmefitsa changes

nlmefit now supports the following error models:

- combined

- constant

- exponential

- proportional

You can specify an error model with both nlmefitsa and nlmefit.

The nlmefit bic calculation has changed. Now the degrees of freedom value is based on the number of groups rather than the number of observations. This conforms with the bic definition used by the nlmefitsa function.

Both nlmefit and nlmefitsa now store the estimated error parameters in the errorparm field of the output stats structure. The rmse field of the structure now contains the root mean squared residual for all error models; this value is computed on the log scale for the exponential model.

### Compatibility Considerations

In the previous release, the `rmse` field was used by `nlmefitsa` for both mean squared residual and the estimated error parameter. Change your code, if necessary, to address the appropriate field in the `stats` structure.

As described in "nlmefit Support for Error Models, and nlmefitsa changes" on page 8, `nlmefit` now calculates different `bic` values than in previous releases.

## Surrogate Splits for Decision Trees

The new surrogate splits feature in `classregtree` allows for better handling of missing values, more accurate estimation of variable importance, and calculation of the predictive measure of association between variables.

## New Bagged Decision Tree Properties

`TreeBagger` and `CompactTreeBagger` classes have two new properties:

- `NVarSplit` provides the number of decision splits for each predictor variable.

- `VarAssoc` provides a measure of association between pairs of predictor variables.

## Enhanced Cluster Analysis Performance

The `linkage` function has improved performance for the `centroid`, `median`, and `single` linkage methods.

The `linkage` and `pdist` hierarchical cluster analysis functions support larger array dimensions with 64-bit platforms, so can handle larger problems.

## Export Probability Objects with dfittool

The distribution fitting GUI (`dfittool`) now allows you to export fits to the MATLAB workspace as probability distribution fit objects. For more information, see "Modeling Data Using the Distribution Fitting Tool".

**Compatibility Considerations**

If you load a distribution fitting session that was created with previous versions of Statistics Toolbox, you cannot save an existing fit. Fit the distribution again to enable saving.

## Compute Partial Correlation of Two Variables Correcting for All Other Variables

`partialcorr` now accepts a new syntax, `RHO = partialcorr(X)`, which returns the sample linear partial correlation coefficients between pairs of variables in `X`, controlling for the remaining variables in `X`. For more information, see the function reference page.

## Specify Number of Evenly Spaced Quantiles

`quantile` now accepts a new syntax, `Y = quantile(X,N,...)`, which returns quantiles at the cumulative probabilities (1:N)/(N+1) where `N` is a scalar positive integer value.

## Control Location and Orientation of Marginal Histograms with scatterhist

`scatterhist` now accepts three parameter name/value pairs that control where and how the histogram plots appear. The new parameter names are `NBins`, `Location`, and `Direction`. For more information, see the function reference page.

## Return Bootstrapped Statistics with bootci

`bootci` has a new output option which returns the bootstrapped statistic computed for each of the `NBoot` bootstrap replicate samples. For more information, see the function reference page.

# Version 7.3 (R2010a) Statistics Toolbox Software

This table summarizes what's new in Version 7.3 (R2010a):

| New Features and Changes | Version Compatibility Considerations | Fixed Bugs and Known Problems |
|---|---|---|
| Yes Details below | No | Bug Reports Includes fixes |

New features and changes introduced in this version are:

- "Stochastic Algorithm Functionality in NLME Models" on page 11
- "$k$-Nearest Neighbor Searching" on page 11
- "Confidence Intervals Option in perfcurve" on page 11
- "Observation Weights Options in Resampling Functions" on page 12

## Stochastic Algorithm Functionality in NLME Models

New stochastic algorithm for fitting NLME models is more robust with respect to starting values, enables parameter transformations, and relaxes assumption of constant error variance. See nlmefitsa.

## *k*-Nearest Neighbor Searching

New functions for $k$-Nearest Neighbor ($k$NN) search efficiently to find the closest points to any query point. For information, see "$k$-Nearest Neighbor Search".

## Confidence Intervals Option in perfcurve

A new option in the perfcurve function computes confidence intervals for classifier performance curves.

## Observation Weights Options in Resampling Functions

New options to weight resampling probabilities broaden the range of models supported by `bootstrp`, `bootci`, and `perfcurve` functions.

# Version 7.2 (R2009b) Statistics Toolbox Software

This table summarizes what's new in Version 7.2 (R2009b):

| New Features and Changes | Version Compatibility Considerations | Fixed Bugs and Known Problems |
|---|---|---|
| Yes<br>Details below | No | Bug Reports<br>Includes fixes |

New features and changes introduced in this version are:

- "New Parallel Computing Support for Certain Functions" on page 13
- "New Stack and Unstack Methods for Dataset Arrays" on page 13
- "New Support for SAS Transport (.xpt) Files" on page 14
- "New Output Function in nlmefit for Monitoring or Canceling Calculations" on page 14

## New Parallel Computing Support for Certain Functions

Statistics Toolbox now supports parallel execution for the following functions:

- `bootci`
- `bootstrp`
- `crossval`
- `jackknife`
- `TreeBagger`

For more information on parallel computing in the Statistics Toolbox, see "Parallel Computing Support for Resampling Methods".

## New Stack and Unstack Methods for Dataset Arrays

`dataset.unstack` converts a "tall" dataset array to an equivalent dataset array that is in "wide format", by "unstacking" a single variable in the tall

dataset array into multiple variables in wide. `dataset.stack` reverses this manipulation by converting a "wide" dataset array to an equivalent dataset array that is in "tall format", by "stacking up" multiple variables in the wide dataset array into a single variable in tall.

## New Support for SAS Transport (.xpt) Files

Statistics Toolbox now supports importing and exporting files in SAS Transport (.xpt) format. For more information, see the `xptread` and `dataset.export` reference pages.

## New Output Function in nlmefit for Monitoring or Canceling Calculations

The `nlmefit` function now supports using an output function to monitor or cancel calculations. For more information, see the `nlmefit` reference page.

# Version 7.1 (R2009a) Statistics Toolbox Software

This table summarizes what's new in Version 7.1 (R2009a):

| New Features and Changes | Version Compatibility Considerations | Fixed Bugs and Known Problems |
|---|---|---|
| Yes<br>Details below | No | Bug Reports<br>Includes fixes |

New features and changes introduced in this version are:

- "Enhanced Dataset Functionality" on page 15
- "New Naïve Bayes Classification" on page 16
- "New Ensemble Methods for Classification and Regression Trees" on page 16
- "New Performance Curve Function" on page 16
- "New Probability Distribution Objects" on page 16

## Enhanced Dataset Functionality

- An enhanced `dataset.join` method provides additional types of join operations:

  - `join` can now perform more complicated inner and outer join operations that allow a many-to-many correspondence between dataset arrays `A` and `B`, and allow unmatched observations in either `A` or `B`.

  - `join` can be of `Type` `'inner'`, `'leftouter'`, `'rightouter'`, `'fullouter'`, or `'outer'` (which is a synonym for `'fullouter'`). For an inner join, the dataset array, `C`, only contains observations corresponding to a combination of key values that occurred in both `A` and `B`. For a left (or right) outer join, `C` also contains observations corresponding to keys in `A` (or `B`) that did not match any in `B` (or `A`).

  - `join` can now return index vectors indicating the correspondence between observations in `C` and those in `A` and `B`.

  - `join` now supports using multiple keys.

- join now supports an optional parameter for specifying missing key behavior rather than raising an error.

- An enhanced `dataset.export` method now supports exporting directly to Microsoft® Excel® files.

## New Naïve Bayes Classification

- The `NaiveBayes` classification object is suitable for data sets that contain many predictors or features.

- It supports normal, kernel, multinomial, and multivariate multinomial distributions.

## New Ensemble Methods for Classification and Regression Trees

- New classification objects, `TreeBagger` and `CompactTreeBagger`, provide improved performance through bootstrap aggregation (bagging).

- Includes Breiman's "random forest" method.

- Enhanced `classregtree` has more options for growing and pruning trees.

## New Performance Curve Function

- New `perfcurve` function provides graphical method to evaluate classification results.

- Includes ROC (receiver operating characteristic) and other curves.

## New Probability Distribution Objects

- Provides a consistent interface for working with probability distributions.

- Can be created directly using the `ProbDistUnivParam` constructor, or fit to data using the `fitdist` function.

- Option to fit distributions by group.

- Includes kernel object methods and parametric object methods that you can use to analyze the distribution represented by the object.

- Includes kernel object properties and parametric object properties that you can access to determine the fit results and evaluate their accuracy.

- Related enhancements in the `chi2gof`, `histfit`, `kstest`, `probplot`, and `qqplot` functions.

# Version 7.0 (R2008b) Statistics Toolbox Software

This table summarizes what's new in Version 7.0 (R2008b):

| New Features and Changes | Version Compatibility Considerations | Fixed Bugs and Known Problems |
|---|---|---|
| Yes<br>Details below | Yes<br>Summary | No |

New features and changes introduced in this version are organized by these topics:

## Classification

The new confusionmat function tabulates misclassifications by comparing known and predicted classes of observations.

## Data Organization

Dataset arrays constructed by the dataset function can now be written to an external text file using the new export function.

When reading external text files into a dataset array, dataset has a new 'TreatAsEmpty' parameter for specifying strings to be treated as empty.

## Compatibility Considerations

In previous versions, `dataset` used `eval` to evaluate strings in external text files before writing them into a dataset array. As a result, strings such as `'1/1/2008'` were treated as numerical expressions with two divides. Now, `dataset` treats such expressions as strings, and writes a string variable into the dataset array whenever a column in the external file contains a string that does not represent a valid scalar value.

# Model Assessment

The cross-validation function, `crossval`, has new options for directly specifying loss functions for mean-squared error or misclassification rate, without having to provide a separate function M-file.

# Multivariate Methods

The `procrustes` function has new options for computing linear transformations without scale or reflection components.

# Probability Distributions

The multivariate normal functions `mvnpdf`, `mvncdf`, and `mvnrnd` now accept vector specification of diagonal covariance matrices, with corresponding gains in computational efficiency.

The hypergeometric distribution has been added to both the `disttool` and `randtool` graphical user interfaces.

## Compatibility Considerations

The `ksdensity` function may give different answers for the case where there are censoring times beyond the last observed value. In this case, `ksdensity` tries to reduce the bias in its density estimate by folding kernel functions across a folding point so that they do not extend into the area that is completely censored. Two things have changed for this release:

**1** In previous releases the folding point was the last observed value. In this release it is the first censoring time after the last observed value.

**2** The folding procedure is applied not just when the `'function'` parameter is `'pdf'`, but for all `'function'` values.

## Regression Analysis

The new `nlmefit` function fits nonlinear mixed-effects models to data with both fixed and random sources of variation. Mixed-effects models are commonly used with data over multiple groups, where measurements are correlated within groups but independent between groups.

## Statistical Visualization

The `boxplot` function has new options for handling multiple grouping variables and extreme outliers.

The `lsline`, `gline`, `refline`, and `refcurve` functions now work with scatter plots produced by the `scatter` function. In previous versions, these functions worked only with scatter plots produced by the `plot` function.

The following visualization functions now have custom data cursors, displaying information such as observation numbers, group numbers, and the values of related variables:

- `andrewsplot`
- `biplot`
- `ecdf`
- `glyphplot`
- `gplotmatrix`
- `gscatter`
- `normplot`
- `parallelcoords`
- `probplot`
- `qqplot`
- `scatterhist`
- `wblplot`

### Compatibility Considerations

Changes to `boxplot` have altered a number of default behaviors:

- Box labels are now drawn as text objects rather than tick labels. Any code that customizes the box labels by changing tick marks should now set the tick locations as well as the tick labels.

- The function no longer returns a handles array with a fixed number handles, and the order and meaning of the handles now depends on which options are selected. To locate a handle of interest, search for its `'Tag'` property using `findobj`. `'Tag'` values for box plot components are listed on the `boxplot` reference page.

- There are now valid handles for outliers, even when boxes have no outliers. In previous releases, the handles array returned by the function had `NaN` values in place of handles when boxes had no outliers. Now the `'xdata'` and `'ydata'` for outliers are `NaN` when there are no outliers.

- For small groups, the `'notch'` parameter sometimes produces notches that extend outside of the box. In previous releases, the notch was truncated to the extent of the box, which could produce a misleading display. A new value of `'markers'` for this parameter avoids the display issue.

As a consequence, the `anova1` function, which displays notched box plots for grouped data, may show notches that extend outside the boxes.

## Utility Functions

The statistics options structure created by `statset` now includes a `Jacobian` field to specify whether or not an objective function can return the Jacobian as a second output.

# Version 6.2 (R2008a) Statistics Toolbox Software

This table summarizes what's new in Version 6.2 (R2008a):

| New Features and Changes | Version Compatibility Considerations | Fixed Bugs and Known Problems |
|---|---|---|
| Yes<br>Details below | Yes<br>Summary | Bug Reports<br>Includes fixes |

New features and changes introduced in this version are organized by these topics:

- "Descriptive Statistics" on page 22

- "Model Assessment" on page 23

- "Multivariate Methods" on page 23

- "Probability Distributions" on page 23

- "Regression Analysis" on page 23

- "Statistical Visualization" on page 23

- "Utility Functions" on page 23

## Descriptive Statistics

Bootstrap confidence intervals computed by `bootci` are now more accurate for lumpy data.

### Compatibility Considerations

The formula for `bootci` confidence intervals of type `'bca'` or `'cper'` involves the proportion of bootstrap statistics less than the observed statistic. The formula now takes into account cases where there are many bootstrap statistics exactly equal to the observed statistic.

## Model Assessment

Two new cross-validation functions, `cvpartition` and `crossval`, partition data and assess models in regression, classification, and clustering applications.

## Multivariate Methods

A new sequential feature selection function, `sequentialfs`, selects predictor subsets that optimize user-defined prediction criteria.

The new `nnmf` function performs nonnegative matrix factorization (NMF) for dimension reduction.

## Probability Distributions

The new `sobolset` and `haltonset` functions produce quasi-random point sets for applications in Monte Carlo integration, space-filling experimental designs, and global optimization. Options allow you to skip, leap over, and scramble the points. The `qrandstream` function provides corresponding quasi-random number streams for intermittent sampling.

## Regression Analysis

The new `plsregress` function performs partial least-squares regression for data with correlated predictors.

## Statistical Visualization

The `normspec` function now shades regions of a normal density curve that are either inside or outside specification limits.

## Utility Functions

The statistics options structure created by `statset` now includes fields for `TolTypeFun` and `TolTypeX`, to specify tolerances on objective functions and parameter values, respectively.

# Version 6.1 (R2007b) Statistics Toolbox Software

This table summarizes what's new in Version 6.1 (R2007b):

| New Features and Changes | Version Compatibility Considerations | Fixed Bugs and Known Problems |
|---|---|---|
| Yes<br>Details below | Yes<br>Summary | Bug Reports<br>Includes fixes |

New features and changes introduced in this version are organized by these topics:

- "Cluster Analysis" on page 24
- "Design of Experiments" on page 25
- "Hypothesis Tests" on page 25
- "Probability Distributions" on page 25
- "Regression Analysis" on page 26
- "Statistical Visualization" on page 27

## Cluster Analysis

The new `gmdistribution` class represents Gaussian mixture distributions, where random points come from different multivariate normal distributions with certain probabilities. The `gmdistribution` constructor creates mixture models with specified means, covariances, and mixture proportions, or by fitting a mixture model with a specified number of components to data. Methods for the class include:

- `fit` — Distribution fitting function
- `pdf` — Probability density function
- `cdf` — Cumulative distribution function
- `random` — Random number generator
- `cluster` — Data clustering
- `posterior` — Cluster posterior probabilities

- `mahal` — Mahalanobis distance

The `cluster` function for hierarchical clustering now accepts a vector of cutoff values, and returns a matrix of cluster assignments, with one column per cutoff value.

### Compatibility Considerations

The `kmeans` function now returns a vector of cluster indices of length $n$, where $n$ is the number of rows in the input data matrix X, even when X contains `NaN` values. In the past, rows of X with `NaN` values were ignored, and the vector of cluster indices was correspondingly reduced in size. Now the vector of cluster indices contains `NaN` values where rows have been ignored, consistent with other toolbox functions.

## Design of Experiments

A new option in the *D*-optimal design function `candexch` specifies fixed design points in the row-exchange algorithm. A similar feature is already available for the `daugment` function, which uses the coordinate-exchange algorithm.

## Hypothesis Tests

The `kstest` function now uses a more accurate method to calculate the *p*-value for a single-sample Kolmogorov-Smirnov test.

### Compatibility Considerations

`kstest` now compares the computed *p*-value to the desired cutoff, rather than comparing the test statistic to a table of values. Results may differ from those in previous releases, especially for small samples in two-sided tests where an asymptotic formula was used in the past.

## Probability Distributions

A new fitting function, `copulafit`, has been added to the family of functions that describe dependencies among variables using copulas. The function fits parametric copulas to data, providing a link between models of marginal distributions and models of data correlations.

A number of probability functions now have improved accuracy, especially for extreme parameter values. The functions are:

- `betainv` — More accurate for probabilities in `P` near 1.
- `binocdf` — More efficient and less likely to run out of memory for large values in `X`.
- `binopdf` — More accurate when the probabilities in `P` are on the order of `eps`.
- `fcdf` — More accurate when the parameter ratios `V2./V1` are much less than the values in `X`.
- `ncx2cdf` — More accurate in some extreme cases that previously returned `0`.
- `poisscdf` — More efficient and less likely to run out of memory for large values in `X`.
- `tcdf` — More accurate when the squares of the values in `X` are much less than the parameters in `V`.
- `tinv` — More accurate when the probabilities in `P` are very close to 0.5 and the outputs are very small in magnitude.

Function-style syntax for `paretotails` objects has been removed.

### Compatibility Considerations

The changes to the probability functions listed above may lead to different, but more accurate, outputs than in previous releases.

In previous releases, syntax of the form `obj(x)` for a `paretotails` objects `obj` invoked the `cdf` method. This syntax now produces a warning. To evaluate the cumulative distribution function, use the syntax `cdf(obj,x)`.

## Regression Analysis

The new `corrcov` function converts a covariance matrix to the corresponding correlation matrix.

The `mvregress` function now supports an option to force the estimated covariance matrix to be diagonal.

**Compatibility Considerations**

In previous releases the `mvregress` function, when using the `'cwls'` algorithm, estimated the covariance of coefficients `COVB` using the estimated, rather than the initial, covariance of the responses `SIGMA`. The initial `SIGMA` is now used, and `COVB` differs to a degree dependent on the difference between the initial and final estimates of `SIGMA`.

## Statistical Visualization

The `boxplot` function has a new `'compact'` plot style suitable for displaying large numbers of groups.

# Version 6.0 (R2007a) Statistics Toolbox Software

This table summarizes what's new in Version 6.0 (R2007a):

| New Features and Changes | Version Compatibility Considerations | Fixed Bugs and Known Problems |
| --- | --- | --- |
| Yes<br>Details below | Yes<br>Summary | Bug Reports<br>Includes fixes |

New features and changes introduced in this version are organized by these topics:

## Data Organization

New categorical and dataset arrays are available for organizing and processing statistical data.

- Categorical arrays facilitate the use of nominal and ordinal categorical data.

- Dataset arrays provide a natural way to encapsulate heterogeneous statistical data and metadata, so that it can be accessed and manipulated using familiar methods analogous to those for numerical matrices.

- Categorical and dataset arrays are supported by a variety of new functions for manipulating the encapsulated data.

- Categorical arrays are now accepted as input arguments in all Statistics Toolbox functions that make use of grouping variables.

## Hypothesis Testing

Expanded options are available for linear hypothesis testing.

- The new `linhyptest` function performs linear hypothesis tests on parameters such as regression coefficients. These tests have the form `H*b = c` for specified values of `H` and `c`, where `b` is a vector of unknown parameters.

- The `covb` output from `regstats` and the `SIGMA` output from `nlinfit` are suitable for use as the covariance matrix input argument required by `linhyptest`. The following functions have been modified to return a `covb` output for use with `linhyptest`: `coxphfit`, `glmfit`, `mnrfit`, `robustfit`.

## Multivariate Statistics

The new `cholcov` function computes a Cholesky-like decomposition of a covariance matrix, even if the matrix is not positive definite. Factors are useful in many of the same ways as Cholesky factors, such as imposing correlation on random number generators.

The `classify` function for discriminant analysis has been improved.

- The function now computes the coefficients of the discriminant functions that define boundaries between classification regions.

- The output of the function is now of the same type as the input grouping variable `group`.

### Compatibility Considerations

The `classify` function now returns outputs of different type than it did in the past. If the input argument `group` is a logical vector, output is now converted to a logical vector. In the past, output was returned as a cell array of `0`s and `1`s. If `group` is numeric, the output is now converted to the same type. For example, if `group` is of type `uint8`, the output will be of type `uint8`.

## Probability Distributions

New `paretotails` objects are available for modeling distributions with an empirical cdf or similar distribution in the center and generalized Pareto distributions in the tails.

- The `paretotails` function converts a data sample to a `paretotails` object. The objects are useful for generating random samples from a distribution similar to the data, but with tail behavior that is less discrete than the empirical distribution.

- Objects from the `paretotails` class are supported by a variety of new methods for working with the piecewise distribution.

- The `paretotails` class provides function-like behavior, so that `p(x)` evaluates the cdf of `p` at values `x`.

## Regression Analysis

The new `mvregresslike` function is a utility related to the `mvregress` function for fitting regression models to multivariate data with missing values. The new function computes the objective (log likelihood) function, and can also compute the estimated covariance matrix for the parameter estimates.

New `classregtree` objects are available for creating and analyzing classification and regression trees.

- The `classregtree` function fits a classification or regression tree to training data. The objects are useful for predicting response values from new predictors.

- Objects from the `classregtree` class are supported by a variety of new methods for accessing information about the tree.

- The `classregtree` class provides function-like behavior, so that `t(X)` evaluates the tree `t` at predictor values in `X`.

- The following functions now create or operate on objects from the new `classregtree` class: `treefit`, `treedisp`, `treeval`, `treefit`, `treeprune`, `treetest`.

### Compatibility Considerations

Objects from the `classregtree` class are intended to be compatible with the structure arrays that were produced in previous versions by the classification and regression tree functions listed above. In particular, `classregtree` supports dot indexing of the form `t.property` to obtain properties of the object `t`. The class also provides function-like behavior through parenthesis indexing, so that `t(x)` uses the tree `t` to classify or compute fitted values for

predictors x, rather than index into t as a structure array as it did in the past. As a result, cell arrays should now be used to aggregate `classregtree` objects.

## Statistical Visualization

The new `scatterhist` function produces a scatterplot of 2D data and illustrates the marginal distributions of the variables by drawing histograms along the two axes. The function is also useful for viewing properties of random samples produced by functions such as `copularnd`, `mvnrnd`, and `lhsdesign`.

## Other Improvements

- The `mvtrnd` function now produces a single random sample from the multivariate *t* distribution if the `cases` input argument is absent.

- The `zscore` function, which centers and scales input data by mean and standard deviation, now returns the means and standard deviations as additional outputs.

# Version 5.3 (R2006b) Statistics Toolbox Software

This table summarizes what's new in Version 5.3 (R2006b):

| New Features and Changes | Version Compatibility Considerations | Fixed Bugs and Known Problems |
|---|---|---|
| Yes<br>Details below | Yes<br>Summary | Bug Reports<br>Includes fixes |

New features and changes introduced in this version are organized by these topics:

- "Demos" on page 32
- "Design of Experiments" on page 32
- "Hypothesis Tests" on page 33
- "Multinomial Distribution" on page 33
- "Regression Analysis" on page 34
- "Statistical Process Control" on page 34

## Demos

The following demo has been updated:

- Selecting a Sample Size — Modified to highlight the new `sampsizepwr` function

## Design of Experiments

The following visualization functions, commonly used in the design of experiments, have been added:

- `interactionplot` — Two-factor interaction plot for the mean
- `maineffectsplot` — Main effects plot for the mean
- `multivarichart` — Multivari chart for the mean

## Hypothesis Tests

The following functions for hypothesis testing have been added or improved:

- jbtest — Replaces the chi-square approximation of the test statistic, which is asymptotic, with a more accurate algorithm that interpolates $p$-values from a table of quantiles. A new option allows you to run Monte Carlo simulations to compute $p$-values outside of the table.

- lillietest — Uses an improved version of Lilliefors' table of quantiles, covering a wider range of sample sizes and significance levels, with more accurate values. New options allow you to test for exponential and extreme value distributions, as well as normal distributions, and to run Monte Carlo simulations to compute $p$-values outside of the tables.

- runstest — Adds a test for runs up and down to the existing test for runs above or below a specified value.

- sampsizepwr — New function to compute the sample size necessary for a test to have a specified power. Options are available for choosing a variety of test types.

### Compatibility Considerations

If the significance level for a test lies outside the range of tabulated values, [0.001, 0.5], then both jbtest and lillietest now return an error. In previous versions, jbtest returned an approximate $p$-value and lillietest returned an error outside a smaller range, [0.01, 0.2]. Error messages suggest using the new Monte Carlo option for computing values outside the range of tabulated values.

If the data sample for a test leads to a $p$-value outside the range of tabulated values, then both jbtest and lillietest now return, with a warning, either the smallest or largest tabulated value. In previous versions, jbtest returned an approximate $p$-value and lillietest returned NaN.

## Multinomial Distribution

The multinomial distribution has been added to the list of almost 50 probability distributions supported by the toolbox.

- mnpdf — Multinomial probability density function

- `mnrnd` — Multinomial random number generator

# Regression Analysis

### Multinomial Regression

Support has been added for multinomial regression modeling of discrete multi-category response data, including multinomial logistic regression. The following new functions supplement the regression models in `glmfit` and `glmval` by providing for a wider range of response values:

- `mnrfit` — Fits a multinomial regression model to data

- `mnrval` — Computes predicted probabilities for the multinomial regression model

### Multivariate Regression

The new `mvregress` function carries out multivariate regression on data with missing response values. An option allows you to specify how missing data is handled.

### Survival Analysis

`coxphfit` — A new option allows you to specify the values at which the baseline hazard is computed.

# Statistical Process Control

The following new functions consolidate and expand upon existing functions for statistical process control:

- `capability` — Computes a wider range of probabilities and capability indices than the `capable` function found in previous releases

- `controlchart` — Displays a wider range of control charts than the `ewmaplot`, `schart`, and `xbarplot` functions found in previous releases

- `controlrules` — Supplements the new `controlchart` function by providing for a wider range of control rules (Western Electric and Nelson)

- `gagerr` — Performs a gage repeatability and reproducibility study on measurements grouped by operator and part

## Compatibility Considerations

The `capability` function subsumes the `capable` function that appeared in previous versions of Statistics Toolbox software, and the `controlchart` function subsumes the functions `ewmaplot`, `schart`, and `xbarplot`. The older functions remain in the toolbox for backwards compatibility, but they are no longer documented or supported.

# Version 5.2 (R2006a) Statistics Toolbox Software

This table summarizes what's new in Version 5.2 (R2006a):

| New Features and Changes | Version Compatibility Considerations | Fixed Bugs and Known Problems |
|---|---|---|
| Yes<br>Details below | No | Bug Reports<br>Includes fixes |

New features and changes introduced in this version are organized by these topics:

- "Analysis of Variance" on page 36
- "Bootstrapping" on page 36
- "Demos" on page 37
- "Design of Experiments" on page 37
- "Hypothesis Tests" on page 37
- "Multivariate Distributions" on page 38
- "Random Number Generation" on page 38
- "Robust Regression" on page 39
- "Statistical Process Control" on page 39

## Analysis of Variance

Support for nested and continuous factors has been added to the `anovan` function for *N*-way analysis of variance.

## Bootstrapping

The following functions have been added to supplement the existing `bootstrp` function for bootstrap estimation:

- `bootci` — Computes confidence intervals of a bootstrapped statistic. An option allows you to choose the type of the bootstrap confidence interval.

- `jackknife` — Draws jackknife samples from a data set and computes statistics on each sample

## Demos

The following demos have been added to the toolbox:

- Bayesian Analysis for a Logistic Regression Model

- Time Series Regression of Airline Passenger Data

The following demo has been updated to demonstrate new features:

- Random Number Generation

## Design of Experiments

The new `fracfactgen` function finds a set of fractional factorial design generators suitable for fitting a specified model.

The following functions for *D*-optimal designs have been enhanced:

- `cordexch`, `daugment`, `dcovary`, `rowexch` — New options specify the range of values and the number of levels for each factor, exclude factor combinations, treat factors as categorical rather than continuous, control the number of iterations, and repeat the design generation process from random starting points

- `candexch` — New options control the number of iterations and repeat the design generation process from random starting points

- `candgen` — New options specify the range of values and the number of levels for each factor, and treat factors as categorical rather than continuous

- `x2fx` — New option treats factors as categorical rather than continuous

## Hypothesis Tests

The new `dwtest` function performs a Durbin-Watson test for autocorrelation in linear regression.

## Multivariate Distributions

Two new functions have been added to compute multivariate cdfs. These supplement existing functions for pdfs and random number generators for the same distributions.

- `mvncdf` — Cumulative distribution function for the multivariate normal distribution

- `mvtcdf` — Cumulative distribution function for the multivariate $t$ distribution

## Random Number Generation

### Copulas

New functions have been added to the toolbox that allow you to use copulas to model correlated multivariate data and generate random numbers from multivariate distributions.

- `copulacdf` — Cumulative distribution function for a copula

- `copulaparam` — Copula parameters as a function of rank correlation

- `copulapdf` — Probability density function for a copula

- `copularnd` — Random numbers from a copula

- `copulastat` — Rank correlation for a copula

### Markov Chain Monte Carlo Methods

The following functions generate random numbers from nonstandard distributions using Markov Chain Monte Carlo methods:

- `mhsample` — Generate random numbers using the Metropolis-Hasting algorithm

- `slicesample` — Generate random numbers using a slice sampling algorithm

### Pearson and Johnson Systems of Distributions

Support has been added for random number generation from Pearson and Johnson systems of distributions.

- pearsrnd — Random numbers from a distribution in the Pearson system

- johnsrnd — Random numbers from a distribution in the Johnson system

## Robust Regression

To supplement the robustfit function, the following functions now have options for robust fitting:

- nlinfit — Nonlinear least-squares regression

- nlparci — Confidence intervals for parameters in nonlinear regression

- nlpredci — Confidence intervals for predictions in nonlinear regression

## Statistical Process Control

The following control chart functions now support time-series objects:

- xbarplot — Xbar plot

- schart — Standard deviation chart

- ewmaplot — Exponentially weighted moving average plot

# Version 5.1 (R14SP3) Statistics Toolbox Software

This table summarizes what's new in Version 5.1 (R14SP3):

| New Features and Changes | Version Compatibility Considerations | Fixed Bugs and Known Problems |
|---|---|---|
| Yes<br>Details below | No | No |

New features and changes introduced in this version are organized by these topics:

- "Demos" on page 40
- "Descriptive Statistics" on page 41
- "Hypothesis Tests" on page 41
- "Probability Distributions" on page 42
- "Regression Analysis" on page 43
- "Statistical Visualization" on page 43

## Demos

The following demos have been added to the toolbox:

- Curve Fitting and Distribution Fitting
- Fitting a Univariate Distribution Using Cumulative Probabilities
- Fitting an Orthogonal Regression Using Principal Components Analysis
- Modelling Tail Data with the Generalized Pareto Distribution
- Pitfalls in Fitting Nonlinear Models by Transforming to Linearity
- Weighted Nonlinear Regression

The following demo has been updated:

- Modelling Data with the Generalized Extreme Value Distribution

# Descriptive Statistics

The new `partialcorr` function computes the correlation of one set of variables while controlling for a second set of variables.

The `grpstats` function now computes a wider variety of descriptive statistics for grouped data. Choices include the mean, standard error of the mean, number of elements, group name, standard deviation, variance, confidence interval for the mean, and confidence interval for new observations. The function also supports the computation of user-defined statistics.

# Hypothesis Tests

### Chi-Square Goodness-of-Fit Test

The new `chi2gof` function tests if a sample comes from a specified distribution, against the alternative that it does not come from that distribution, using a chi-square test statistic.

### Variance Tests

Three functions have been added to test sample variances:

- `vartest` — One-sample chi-square variance test. Tests if a sample comes from a normal distribution with specified variance, against the alternative that it comes from a normal distribution with a different variance.

- `vartest2` — Two-sample $F$-test for equal variances. Tests if two independent samples come from normal distributions with the same variance, against the alternative that they come from normal distributions with different variances.

- `vartestn` — Bartlett multiple-sample test for equal variances. Tests if multiple samples come from normal distributions with the same variance, against the alternative that they come from normal distributions with different variances.

### Ansari-Bradley Test

The new `ansaribradley` function tests if two independent samples come from the same distribution, against the alternative that they come from distributions that have the same median and shape but different variances.

41

### Tests of Randomness

The new `runstest` function tests if a sequence of values comes in random order, against the alternative that the ordering is not random.

# Probability Distributions

Support has been added for two new distributions:

- "Generalized Extreme Value Distribution" on page 42
- "Generalized Pareto Distribution" on page 42

### Generalized Extreme Value Distribution

The Generalized Extreme Value distribution combines the Gumbel, Frechet, and Weibull distributions into a single distribution. It is used to model extreme values in data.

The following distribution functions have been added:

- `gevcdf` — Cumulative distribution function
- `gevfit` — Parameter estimation function
- `gevinv` — Inverse cumulative distribution function
- `gevlike` — Negative log-likelihood function
- `gevpdf` — Probability density function
- `gevrnd` — Random number generator
- `gevstat` — Distribution statistics

### Generalized Pareto Distribution

The Generalized Pareto distribution is used to model the tails of a data distribution.

The following distribution functions have been added:

- `gpcdf` — Cumulative distribution function
- `gpfit` — Parameter estimation function

- `gpinv` — Inverse cumulative distribution function
- `gplike` — Negative log-likelihood function
- `gppdf` — Probability density function
- `gprnd` — Random number generator
- `gpstat` — Distribution statistics

## Regression Analysis

- The new `coxphfit` function fits Cox's proportional hazards regression model to data.
- The new `invpred` function estimates the inverse prediction intervals for simple linear regression.
- The `polyconf` function has new options to let you specify the confidence interval computed.

## Statistical Visualization

Both the `ecdf` and `ksdensity` functions now produce plots when no output arguments are specified.

# Version 5.0.2 (R14SP2) Statistics Toolbox Software

This table summarizes what's new in Version 5.0.2 (R14SP2):

| New Features and Changes | Version Compatibility Considerations | Fixed Bugs and Known Problems |
|---|---|---|
| Yes<br>Details below | No | Bug Reports<br>Includes fixes |

New features and changes introduced in this version are organized by this topic:

## Multivariate Statistics

The cophenet function now returns cophenetic distances as well as the cophenetic correlation coefficient.

# Compatibility Summary for Statistics Toolbox Software

This table summarizes new features and changes that might cause incompatibilities when you upgrade from an earlier version, or when you use files on multiple versions. Details are provided in the description of the new feature or change.

| Version (Release) | New Features and Changes with Version Compatibility Impact |
|---|---|
| **Latest Version V7.5 (R2011a)** | None |
| V7.4 (R2010b) | See the **Compatibility Considerations** subheading for each of these new features and changes: <br><br> • "nlmefit Support for Error Models, and nlmefitsa changes" on page 8 <br><br> • "Export Probability Objects with dfittool" on page 9 |
| V7.3 (R2010a) | None |
| V7.2 (R2009b) | None |
| V7.1 (R2009a) | None |
| V7.0 (R2008b) | See the **Compatibility Considerations** subheading for each of these new features and changes: <br><br> • "Data Organization" on page 18 <br><br> • "Statistical Visualization" on page 20 |

| Version (Release) | New Features and Changes with Version Compatibility Impact |
|---|---|
| V6.2 (R2008a) | See the **Compatibility Considerations** subheading for this change:<br><br>• "Descriptive Statistics" on page 22 |
| V6.1 (R2007b) | See the **Compatibility Considerations** subheading for each of these new features and changes:<br><br>• "Cluster Analysis" on page 24<br>• "Hypothesis Tests" on page 25<br>• "Probability Distributions" on page 25<br>• "Regression Analysis" on page 26 |
| V6.0 (R2007a) | See the **Compatibility Considerations** subheading for each of these new features and changes:<br><br>• "Multivariate Statistics" on page 29<br>• "Regression Analysis" on page 30 |
| V5.3 (R2006b) | See the **Compatibility Considerations** subheading for each of these new features and changes:<br><br>• "Hypothesis Tests" on page 33<br>• "Statistical Process Control" on page 34 |
| V5.2 (R2006a) | None |

| Version (Release) | New Features and Changes with Version Compatibility Impact |
|---|---|
| V5.1 (R14SP3) | None |
| V5.0.2 (R14SP2) | None |