Statistics Toolbox 7

Perform statistical analysis, modeling, and algorithm development
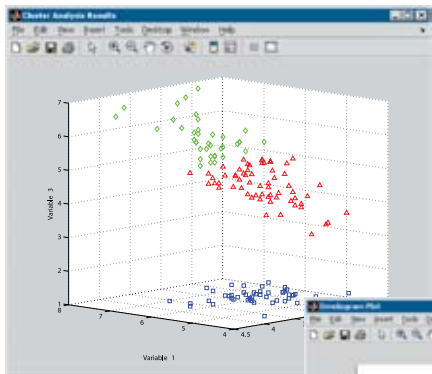
Statistics Toolbox™ provides a comprehensive set of tools to assess and understand data. Statistics Toolbox includes functions and interactive tools for modeling data, analyzing historical trends, simulating systems, developing statistical algorithms, and learning and teaching statistics.

The toolbox supports a wide range of tasks, from calculating basic descriptive statistics to developing and visualizing multidimensional nonlinear models. It offers a rich set of statistical plot types and interactive graphics, such as polynomial fitting and response surface modeling.
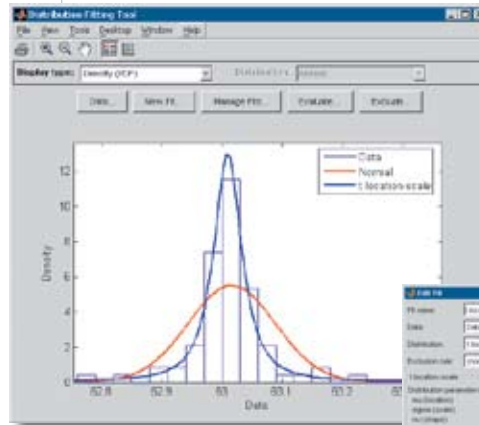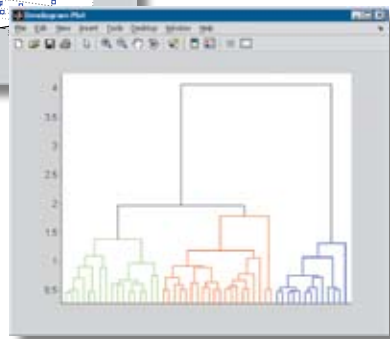
All toolbox functions are written in the open MATLAB® language so that you can inspect the algorithms, modify the source code, and create your own custom functions.

KEY FEATURES

- Data organization and management
- Descriptive statistics
- Statistical plotting and data visualization
- Probability distributions
- Analysis of variance (ANOVA)
- Linear and nonlinear modeling
- Multivariate statistics
- Design of Experiments (DOE)
- Hypothesis testing
- Statistical Process Control (SPC)

Fitting univariate distributions to data. The Distribution Fitting Tool lets you easily import, analyze, and plot your data.

Plots showing how data with similar characteristics can be clustered into groups. Statistics Toolbox includes functions for multivariate analysis and clustering.

Regression analysis to determine the most important ingredients for cement-mixture curing. Stepwise regression capabilities in Statistics Toolbox provide automated procedures for identifying models from several potential explanatory variables.

Data Organization and Management

Statistics Toolbox provides two specialized arrays for statistical data: dataset arrays and categorical arrays.
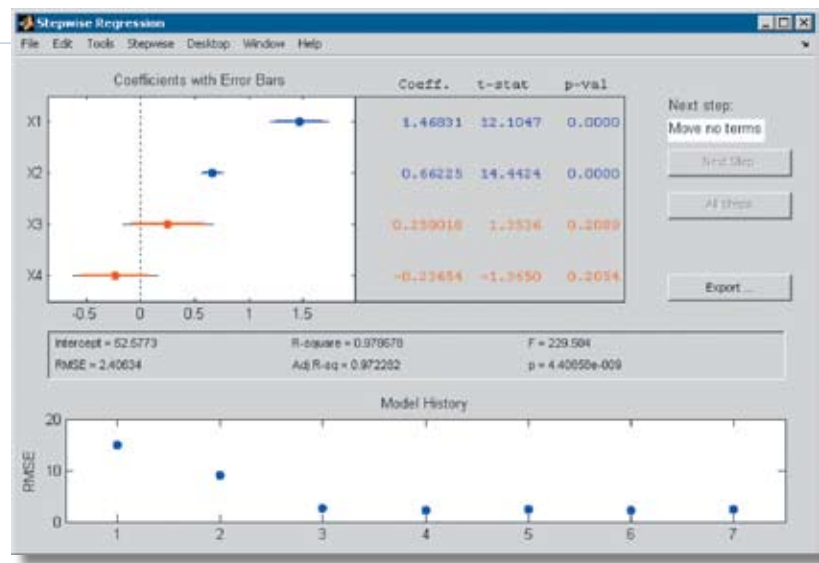
Dataset Arrays

Dataset arrays enable convenient organization and analysis of heterogeneous statistical data and metadata. Dataset arrays have columns that represent measured variables and rows that represent observations. With dataset arrays, you can:

- Collect variables of different data types and sizes in a single array

- Use metadata to describe variables and observations and to access them by name

- View summary statistics and display data in an intuitive tabular format

- Create, manage, and operate on dataset arrays using a variety of supporting methods

Categorical Arrays

Categorical arrays let you organize and process categorical data that takes on values from a finite set of discrete levels or categories. With categorical arrays, you can:

- Store nominal data using descriptive labels, such as "red," "green," and "blue" for an unordered set of colors

- Store ordinal data using descriptive labels, such as "cold," "warm," and "hot" for an ordered set of temperature measurements

- Manipulate categorical data using familiar array operations and indexing methods

- Index into other variables or create subsets of data based on the category of observation

- Group observations of the same category for computing statistics and creating visualizations

Descriptive Statistics

Descriptive statistics methods enable you to quickly understand and describe potentially large sets of data. Statistics Toolbox includes functions for calculating:

- Measures of central tendency (measures of location), including average, median, and various means

- Measures of dispersion (measures of spread), including range, variance, standard deviation, and mean or median absolute deviation

- Linear and rank correlation (partial and full)

- Results based on data with missing values

- Percentile and quartile estimates

- Density estimates (using a kernel-smoothing function)

These functions help you summarize the values in a data sample with a few highly relevant numbers. Although descriptive statistics are typically generated using parametric techniques, you can also use bootstrap methods to derive descriptive statistics.

Statistical Plotting and Interactive Graphics

Statistics Toolbox includes numerous functions that help you represent your data graphically. In addition to the standard set of MATLAB plot types, Statistics Toolbox includes box plots, probability plots, histograms and 3-D histograms, control charts, quantile-quantile plots, and several multivariate plots. It also provides interactive graphics that enhance analysis in areas such as:

- Nonlinear and polynomial fitting and prediction

- Exploration of distribution functions and distribution fitting and analysis

- Interactive random number generation

- Response surface modeling

- Interactive process experimentation and analysis

- Stepwise regression analysis
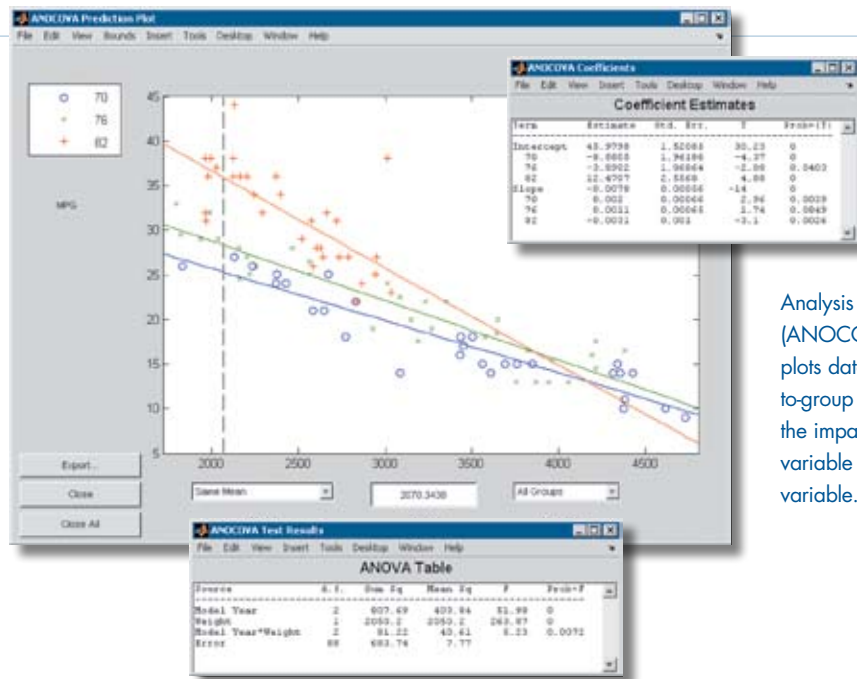
Analysis of covariance (ANOCOVA) tool, which plots data to assess group-to-group differences and the impact of a predictor variable on a response variable.

Probability Distributions

Statistics Toolbox includes an extensive library of probability distribution functions that let you fit probability distributions to your data, compute functions from them, and generate random samples. From the command line you have access to more than 150 functions for:

- Generating random and quasi-random point sets

- Calculating the probability density function (pdf)

- Calculating the cumulative distribution function (cdf) and its inverse

- Computing mean and variance

- Estimating distribution parameters

Statistics Toolbox includes three interactive graphical user interfaces (GUIs) that simplify common analysis tasks. The Distribution Fitting Tool GUI lets you fit data using 23 predefined probability distributions, a non-parametric (kernel-smoothing) estimator, or a custom distribution that you define yourself. It supports both complete and censored (reliability) data and lets you exclude data, save and load sessions, and generate MATLAB code.

The Distribution Tool GUI lets you learn about a variety of probability distributions and explore how various parameters affect their position and shape.

The Random Number Tool GUI provides a random number generator to simulate behavior associated with particular distributions. You can use this random data to test hypotheses or models under different conditions, as well as perform Monte Carlo simulations.

Statistics Toolbox also includes functions for generating random samples from multivariate distributions, such as t, normal, copulas, and Wishart; sampling from finite populations; performing Latin hypercube sampling; and generating samples from Pearson and Johnson systems of distributions.
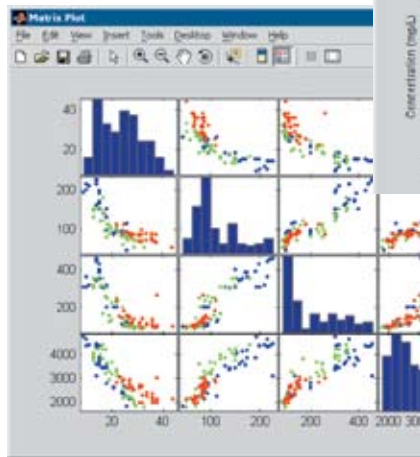
Statistics Toolbox can also fit parametric copulas to data, providing a link between models that describe marginal distribution and models that describe data correlations.
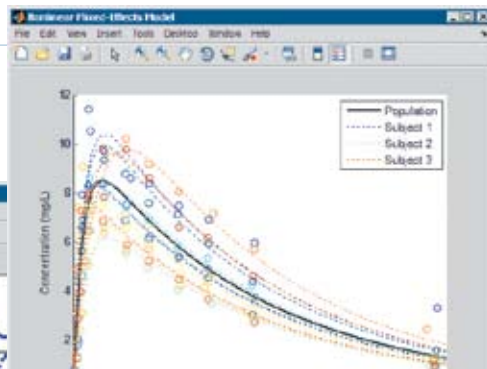
Analysis of Variance

Analysis of variance (ANOVA) lets you determine whether data sets from different groups have different characteristics. You can classify groups using discrete predictor variables. A follow-up multiple comparison analysis can pinpoint which pairs of groups differ from each other.

Statistics Toolbox includes algorithms for ANOVA and related techniques, including:

- One-way ANOVA with graphics

- Two-way ANOVA for balanced data

- Multiway ANOVA for balanced and unbalanced data (both fixed and random effects)

- Multivariate ANOVA

- Nonparametric one- and two-way ANOVA (Kruskal-Wallis, Friedman)

- Analysis of covariance (ANOCOVA)

- Multiple comparison of group means, slopes, and intercepts

Nonlinear mixed-effects model of drug absorption and elimination showing intrasubject concentration-versus-time profiles. The `nlmefit` function in Statistics Toolbox generates a population model using fixed and random effects.

Matrix of scatter plots and histograms comparing automobile performance over three model years. Statistics Toolbox makes it easy to plot multiple variables and compare data.

Linear and Nonlinear Modeling

The linear and nonlinear models provided in Statistics Toolbox let you model a response variable as a function of one or more predictor variables. These models make predictions, establish relationships between variables, or simplify a problem. For example, linear and nonlinear regression models help establish which variables have the most impact on a response. Robust regression methods can help you find outliers and reduce their effect on the fitted model.

The toolbox provides linear algorithms for:

• Polynomial, stepwise, ridge, robust, and multiple linear regression

• Generalized linear models, including multinomial (discrete-choice) models

• Response surface fitting

The toolbox also provides nonlinear fitting functions. Using nonlinear least squares functions, you can:

• Estimate parameters

• Interactively visualize and predict multidimensional nonlinear fitting

• Set confidence intervals for parameters and predicted values

Statistics Toolbox supports a variety of techniques for classification analysis, including:

• Discriminant analysis

• Classification and regression trees

Nonlinear mixed-effect (NLME) models provide a flexible mechanism to model correlations within subgroups contained within a larger data sample.
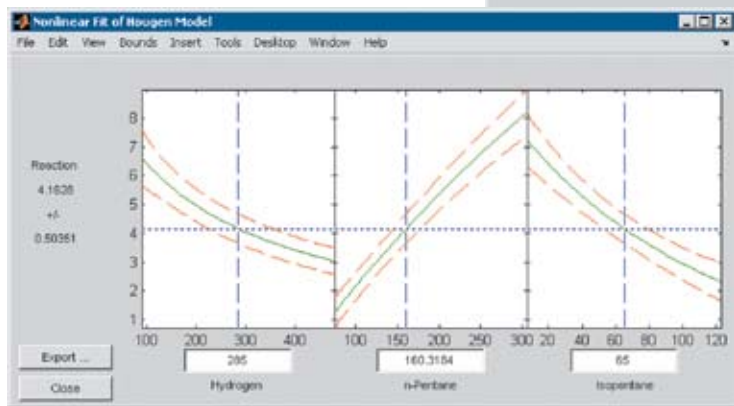
Statistical learning techniques such as cross-validation and sequential feature selection can be applied to both linear and nonlinear fitting methods.

You can also use the toolbox to work with Hidden Markov models. You can estimate the parameters of a model using the Baum-Welch algorithm, calculate the most likely path through a model using the Viterbi algorithm, and generate random sequences from a given model.

Multivariate Statistics

Multivariate statistics methods let you analyze your data by evaluating groups of variables together. You can:

• Segment data in clusters for further analysis

• Visualize and assess the group-to-group differences in a data set

• Reduce a large set of variables to a more manageable but still representative set

Multivariate statistics tasks supported by Statistics Toolbox include:

• Factor analysis

• Factor rotation and biplots

• Cluster analysis (both hierarchical and k-means)

• Multidimensional scaling (classical, metric, and nonmetric)

• Multivariate plotting (parallel coordinates, glyph plots, and Andrews plots)

Statistics Toolbox also supports data transformation and dimensionality reduction techniques like principal component analysis and partial least squares.

Design of Experiments

Functions for Design of Experiments (DOE) help you create and test practical plans to gather data for statistical modeling. These plans show you how to manipulate your data inputs in tandem to generate information about their effect on the outputs. Supported design types include:

- Full factorial

- Fractional factorial

- Response surface (central composite and Box-Behnken)

- D-optimal

- Latin hypercube

You can use the toolbox to define, analyze, and visualize a customized DOE. For example, you can estimate input effects and input interactions using ANOVA, linear regression, and response surface modeling, and then visualize results through main effect plots, interaction plots, and multi-vari charts.
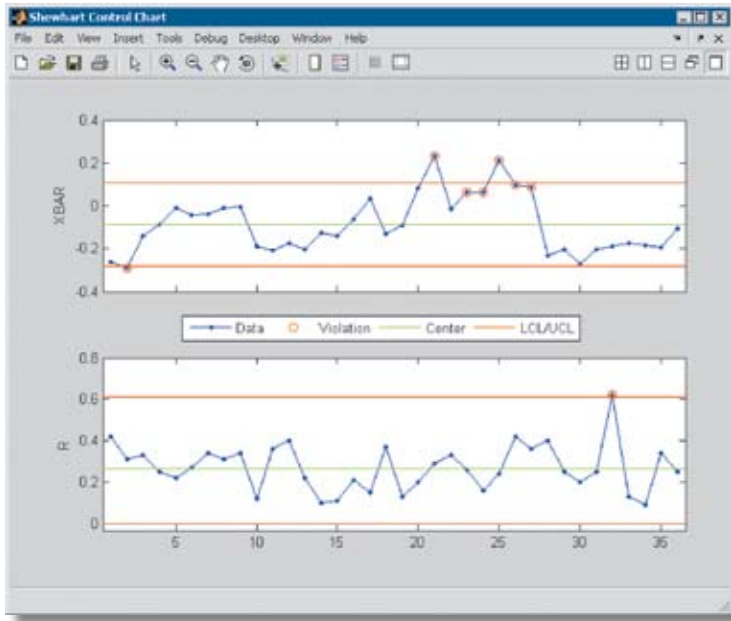
Hypothesis Testing

Random variation often makes it difficult to determine whether samples taken under different conditions really are different. Hypothesis testing is an effective tool for analyzing whether sample-to-sample differences are significant and require further evaluation or are consistent with random and expected data variation.

Statistics Toolbox supports the most widely used parametric and nonparametric hypothesis testing procedures, such as:

- One- and two-sample t-tests

- One-sample z-test

- Nonparametric tests for one sample, paired samples, and two independent samples

- Distribution tests (Chi-square, Jarque-Bera, Lilliefors, and Kolmogorov-Smirnov)

- Comparison of distributions (two-sample Kolmogorov-Smirnov)

- Autocorrelation and randomness tests

- Linear hypotheses tests on regression coefficients
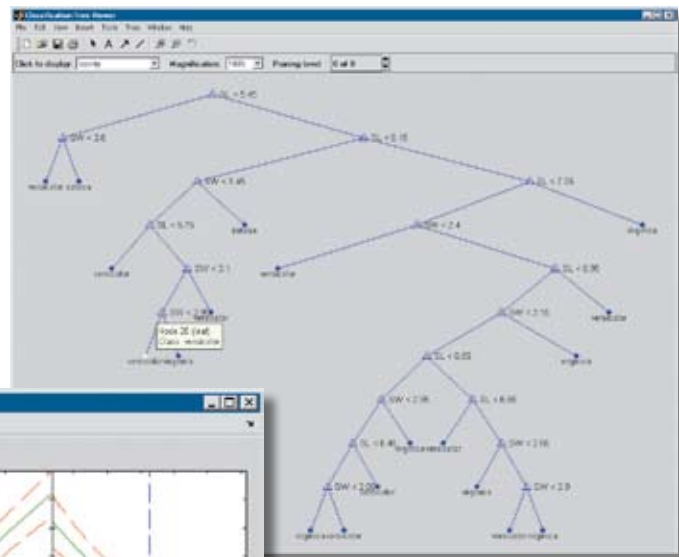
Fitting a decision tree to data. The fitting capabilities in Statistics Toolbox let you visualize a decision tree by drawing a diagram of the decision rule and group assignments.

Model of a chemical reaction of an experiment using Design-of-Experiments (DOE) and surface-fitting capabilities of Statistics Toolbox.

Control charts showing process data and violations of Western Electric control rules. Statistics Toolbox provides a variety of control charts and control rules for monitoring and evaluating products or processes.

Statistical Process Control

Statistics Toolbox provides a set of functions that support Statistical Process Control (SPC). These functions enable you to monitor and improve products or processes by evaluating process variability. SPC functions let you:

- Perform gage repeatability and reproducibility studies

- Estimate process capability

- Create 11 control charts

- Apply Western Electric and Nelson control rules to control chart data

Required Products

MATLAB®

Related Products

Bioinformatics Toolbox™
Read, analyze, and visualize genomic, proteomic, and microarray data

Curve Fitting Toolbox™
Create curve fits and regression models of data

Financial Toolbox™
Analyze financial data and develop financial algorithms

Optimization Toolbox™
Solve standard and large-scale optimization problems

Signal Processing Toolbox™
Perform signal processing, analysis, and algorithm development

Platform and System Requirements

For information on platform and system requirements, visit *www.mathworks.com /products/statistics* ■

Learn More
www.mathworks.com/products/statistics

Resources

VISIT
www.mathworks.com

TECHNICAL SUPPORT
www.mathworks.com/support

ONLINE USER COMMUNITY
www.mathworks.com/matlabcentral

DEMOS
www.mathworks.com/demos

TRAINING SERVICES
www.mathworks.com/training

THIRD-PARTY PRODUCTS AND SERVICES
www.mathworks.com/connections

WORLDWIDE CONTACTS
www.mathworks.com/contact

E-MAIL
info@mathworks.com