Chapter

III-12

Statistics

Overview

This chapter describes operations and functions for statistical analysis together with some general guidelines for their use. This is not a statistics tutorial; for that you can consult one of the references at the end of this chapter or the references listed in the documentation of a particular operation or function. The material below assumes that you are familiar with techniques and methods of statistical analysis.

Most statistics operations and functions are named with the prefix “Stats”. Naming exceptions include the random noise functions that have traditionally been named based on the distribution they represent.

There are six natural groups of statistics operations and functions. They include:

- Test operations
- Noise functions
- Probability distribution functions (PDFs)
- Cumulative distribution functions (CDFs)
- Inverse cumulative distribution functions
- General purpose statistics operations and functions

Statistics Test Operations

Test operations analyze the input data to examine the validity of a specific hypothesis. The common test involves a computation of some numeric value (also known as “test statistic”) which is usually compared with a critical value in order to determine if you should accept or reject the test hypothesis (H_0). Most tests compute a critical value for the given significance *alpha* which has the default value 0.05 or a user-provided value via the /ALPH flag. Some tests directly compute the P value which you can compare to the desired significance value.

Critical values have been traditionally published in tables for various significance levels and tails of distributions. They are by far the most difficult technical aspect in implementing statistical tests. The critical values are usually obtained from the inverse of the CDF for the particular distribution, i.e., from solving the equation

$$cdf(criticalValue) = 1 - alpha,$$

where *alpha* is the significance. In some distributions (e.g., Friedman’s) the calculation of the CDF is so computationally intensive that it is impractical (using desktop computers in 2006) to compute for very large parameters. Fortunately, large parameters usually imply that the distributions can be approximated using simpler expressions. Igor’s tests provide whenever possible exact critical values as well as the common relevant approximations.

Comparison of critical values with published table values can sometimes be interesting as there does not appear to be a standard for determining the published critical value when the CDF takes a finite number of discrete values (step-like). In this case the CDF attains the value (1-*alpha*) in a vertical transition so one could use the X value for the vertical transition as a critical value or the X value of the subsequent vertical transition. Some tables reflect a “conservative” approach and print the X value of subsequent transitions.

Statistical test operations can print their results to the history area of the command window and save them in a wave in the current data folder. Result waves have a fixed name associated with the operation. Elements in the wave are designated by dimension labels. You can use the /T flag to display the results of the operation in a table with dimension labels. The argument for this flag determines what happens when you kill the table. You can use/Q in all test operations to prevent printing information in the history area and you can use the /Z flag to make sure that the operations do not report errors except by setting the V_Flag variable to -1.

Statistical test operations tend to include several variations of the named test. You can usually choose to execute one or more variations by specifying the appropriate flags. The following table can be used as a guide for identifying the operation associated with a given test name.

Statistical Test Operations by Name

Test Name	Where to find it
Angular Distance	**StatsAngularDistanceTest**
Bartlett’s test for variances	**StatsVariancesTest**
BootStrap	**StatsResample**
Brown and Forsythe	**StatsANOVA1Test**
Chi-squared test for means	**StatsChiTest**
Cochran’s test	**StatsCochranTest**
Dunn-Holland-Wolfe	**StatsNPMCTest**
Dunnette multicomparison test	**StatsDunnettTest**, **StatsLinearRegression**
Fisher's Exact Test	**StatsContingencyTable**
Fixed Effect Model	**StatsANOVA1Test**
Friedman test on randomized block	**StatsFriedmanTest**
F-test on two distributions	**StatsFTest**
Hodges-Ajne (Batschelet)	**StatsHodgesAjneTest**
Hartigan test for unimodality	**StatsDIPTest**
Hotelling	**StatsCircularTwoSampleTest**, **StatsCircularMeans**
Jackknife	**StatsResample**
Jarque-Bera Test	**StatsJBTest**
Kolmogorov-Smirnov	**StatsKSTest**
Kruskal-Wallis	**StatsKWTest**
Kuiper Test	**StatsCircularMoments**
Levene’s test for variances	**StatsVariancesTest**
Linear Correlation Test	**StatsLinearCorrelationTest**
Linear Order Statistic	**StatsCircularMoments**
Mann-Kendall	**StatsKendallTauTest**
Moore test	**StatsCircularTwoSampleTest**, **StatsCircularMeans**
Nonparametric multiple contrasts	**StatsNPMCTest**
Nonparametric angular-angular correlation	**StatsCircularCorrelationTest**
Nonparametric second order circular analysis	**StatsCircularMeans**
Nonparametric serial randomness (nominal)	**StatsNPNominalSRTest**
Parametric angular-angular correlation	**StatsCircularCorrelationTest**
Parametric angular-Linear correlation	**StatsCircularCorrelationTest**
Parametric second order circular analysis	**StatsCircularMeans**
Parametric serial randomness test	**StatsSRTest**
Rayleigh	**StatsCircularMoments**
Repeated Measures	**StatsANOVA2RMTest**

Test Name	Where to find it
Scheffe equality of means	**StatsScheffeTest**
Shapiro-Wilk test for normality	**StatsShapiroWilkTest**
Spearman	**StatsRankCorrelationTest**
Student-Newman-Keuls	**StatsNPMCTest**
Tukey Test	**StatsTukeyTest** **StatsLinearRegression**, **StatsMultiCorrelationTest**, **StatsNPMCTest**
Two-Factor ANOVA	**StatsANOVA2NRTest**
T-test	**StatsTTest**
Watson's nonparametric two-sample U2	**StatsWatsonUSquaredTest**, **StatsCircularTwoSampleTest**
Watson-Williams	**StatsWatsonWilliamsTest**
Weighted-rank correlation test	**StatsWRCorrelationTest**
Wheeler-Watson nonparametric test	**StatsWheelerWatsonTest**
Wilcoxon-Mann-Whitney two-sample	**StatsWilcoxonRankTest**
Wilcoxon signed rank	**StatsWilcoxonRankTest**

Statistical Test Operations by Data Format

The following tables group statistical operations and functions according to the format of the input data.

Tests for Single Waves

Analysis Method	Comments
StatsChiTest	Compares with known binned values
StatsCircularMoments	WaveStats for circular data
StatsKendallTauTest	Similar to Spearman’s correlation
StatsMedian	Returns the median
StatsNPNominalSRTest	Nonparametric serial randomness test
StatsQuantiles	Computes quantiles and more
StatsResample	Bootstrap analysis
StatsSRTest	Serial randomenss test
StatsTrimmedMean	Returns the trimmed mean
StatsTTest	Compares with known mean
Sort	Reorders the data
WaveStats	Basic statistical description
StatsJBTest	Jarque-Bera test for normality
StatsKSTest	Limited scope test for normality
StatsDIPTest	Hartigan test for unimodality
StatsShapiroWilkTest	Shapiro-Wilk test for normality

Tests for Two Waves

Analysis Method	Comments
StatsChiTest	Chi-squared statistic for comparing two distributions
StatsCochranTest	Randomized block or repeated measures test
StatsCircularTwoSampleTest	Second order analysis of angles
StatsDunnettTest	Compares multiple groups to a control
StatsFTest	Computes ratio of variances
StatsFriedmanTest	Nonparametric ANOVA
StatsKendallTauTest	Similar to Spearman’s correlation
StatsTTest	Compares the means of two distributions
StatsANOVA1Test	One-way analysis of variances
StatsLinearRegression	Linear regression analysis
StatsLinearCorrelationTest	Linear correlation coefficient and its error
StatsRankCorrelationTest	Computes Spearman’s rank correlation
StatsVariancesTest	Compares variances of waves
StatsWilcoxonRankTest	Two-sample or signed rank test
StatsWatsonUSquaredTest	Compares two populations of circular data
StatsWatsonWilliamsTest	Compares mean values of angular distributions
StatsWheelerWatsonTest	Compares two angular distributions

Tests for Multiple or Multidimensional Waves

Analysis Method	Comments
StatsANOVA1Test	One-way analysis of variances
StatsANOVA2Test	Two-factor analysis of variances
StatsANOVA2RMTest	Two-factor repeated measure ANOVA
StatsCochranTest	Randomized block or repeated measures test
StatsContingencyTable	Contingency table analysis
StatsDunnettTest	Comparing multiple groups to a control
StatsFriedmanTest	Nonparametric ANOVA
StatsNPMCTest	Nonparametric multiple comparison tests
StatsScheffeTest	Tests equality of means
StatsTukeyTest	Multiple comparisons based on means
StatsWatsonWilliamsTest	Compares mean values of angular distributions
StatsWheelerWatsonTest	Compares two angular distributions

Statistical Test Operations for Angular/Circular Data

StatsAngularDistanceTest
StatsCircularMoments
StatsCircularMeans
StatsCircularTwoSampleTest
StatsCircularCorrelationTest
StatsHodgesAjneTest
StatsWatsonUSquaredTest
StatsWatsonWilliamsTest
StatsWheelerWatsonTest

Statistical Test Operations: Nonparametric Tests

Operation	Comments
StatsAngularDistanceTest	
StatsFriedmanTest	
StatsCircularTwoSampleTest	Parametric or nonparametric
StatsCircularCorrelationTest	Parametric or nonparameteric
StatsCircularMeans	Parametric or nonparameteric
StatsHodgesAjneTest	
StatsKendallTauTest	
StatsKWTest	
StatsNPMCTest	
StatsNPNominalSRTest	
StatsRankCorrelationTest	
StatsWatsonUSquaredTest	
StatsWheelerWatsonTest	
StatsWilcoxonRankTest	

Noise Functions

The following functions return numbers from a pseudo-random distribution of the specified shapes and parameters. Except for enoise and gnoise where you have an option to select a random number generator, the remaining noise functions use a Mersenne Twister algorithm for the initial uniform pseudo-random distribution. Note that whenever you need repeatable results you should use SetRandomSeed prior to executing any of the noise functions.

The following noise generation functions are available:

binomialNoise

enoise

expnoise

gammaNoise

gnoise

hyperGNoise

logNormalNoise

lorentzianNoise

poissonNoise

StatsPowerNoise

StatsVonMisesNoise

wnoise

Cumulative Distribution Functions

A cumulative distribution function (CDF) is the integral of its respective probability distribution function (PDF). CDFs are usually well behaved functions with values in the range [0,1]. CDFs are important in computing critical values, P values and power of statistical tests.

Many CDFs are computed directly from closed form expressions. Others can be difficult to compute because they involve evaluating a very large number of states, e.g., Friedman or USquared distributions. In these cases you have the following options:

1. Use a built-in table that consists of exact, precomputed values.
2. Compute an approximate CDF based on the prevailing approximation method or using a Monte-Carlo approach.
3. Compute the exact CDF.

Built-in tables are ideal if they cover the range of the parameters that you need. Monte-Carlo methods can be tricky in the sense that repeated application may return small variations in values. Computing the exact CDF may be desirable, but it is often impractical. In most situations the range of parameters that is practical to compute on a desktop machine is already covered in the built-in tables. Larger parameters not have been considered because they take days to compute or because they require 64 bit processors. In addition, most of the approximations tend to improve with increasing size of the parameters.

The functions to calculate values from CDFs are as follows:

StatsBetaCDF

StatsBinomialCDF

StatsCauchyCDF

StatsChiCDF

StatsCMSSDCDF

StatsDExpCDF

StatsErlangCDF

StatsEValueCDF

StatsExpCDF

StatsHyperGCDF

StatsKuiperCDF

StatsLogisticCDF

StatsLogNormalCDF

StatsMaxwellCDF

StatsInvMooreCDF

StatsNBinomialCDF

StatsNCFCDF

StatsNCTCDF

StatsQCDF

StatsRayleighCDF

StatsRectangularCDF

StatsRunsCDF

StatsSpearmanRhoCDF

StatsStudentCDF

StatsTopDownCDF

StatsTriangularCDF

StatsUSquaredCDF

StatsFCDF

StatsFriedmanCDF

StatsGammaCDF

StatsGeometricCDF

StatsNormalCDF

StatsParetoCDF

StatsPoissonCDF

StatsPowerCDF

StatsVonMisesCDF

StatsQCDF

StatsWaldCDF

StatsWeibullCDF

Probability Distribution Functions

Probability distribution functions (PDF) are sometimes known as probability densities. In the case of continuous distributions, the area under the curve of the PDF for each interval equals the probability for the random variable to fall within that interval. The PDFs are useful in calculating event probabilities, characteristic functions and moments of a distribution.

The functions to calculate values from PDFs are as follows:

StatsBetaPDF

StatsBinomialPDF

StatsCauchyPDF

StatsChiPDF

StatsDExpPDF

StatsErlangPDF

StatsErrorPDF

StatsEValuePDF

StatsExpPDF

StatsFPDF

StatsGammaPDF

StatsGeometricPDF

StatsHyperGPDF

StatsLogNormalPDF

StatsMaxwellPDF

StatsBinomialPDF

StatsNCChiPDF

StatsNCFPDF

StatsNCTPDF

StatsNormalPDF

StatsParetoPDF

StatsPoissonPDF

StatsPowerPDF

StatsRayleighPDF

StatsRectangularPDF

StatsStudentPDF

StatsTriangularPDF

StatsVonMisesPDF

StatsWaldPDF

StatsWeibullPDF

Inverse Cumulative Distribution Functions

The inverse cumulative distribution functions return the values at which their respective CDFs attain a given level. This value is typically used as a critical test value. There are very few functions for which the inverse CDF can be written in closed form. In most situations the inverse is computed iteratively from the CDF.

The functions to calculate values from inverse CDFs are as follows:

StatsInvBetaCDF

StatsInvBinomialCDF

StatsInvCauchyCDF

StatsInvChiCDF

StatsInvCMSSDCDF

StatsInvDExpCDF

StatsInvEValueCDF

StatsInvExpCDF

StatsInvFCDF

StatsInvKuiperCDF

StatsInvLogisticCDF

StatsInvLogNormalCDF

StatsInvMaxwellCDF

StatsInvMooreCDF

StatsInvNBinomialCDF

StatsInvNCFCDF

StatsInvNormalCDF

StatsInvParetoCDF

StatsInvQpCDF

StatsInvRayleighCDF

StatsInvRectangularCDF

StatsInvSpearmanCDF

StatsInvStudentCDF

StatsInvTopDownCDF

StatsInvTriangularCDF

StatsInvUSquaredCDF

StatsInvVonMisesCDF

- **StatsInvFriedmanCDF**
- **StatsInvGammaCDF**
- **StatsInvGeometricCDF**
- **StatsInvPoissonCDF**
- **StatsInvPowerCDF**
- **StatsInvQCDF**
- **StatsInvWeibullCDF**

General Purpose Statistics Operations and Functions

This group includes operations and functions that existed before IGOR Pro 6.0 and some general purpose operations and functions that do not belong to the main groups listed.

- **binomial**
- **binomialln**
- **erf**
- **erfc**
- **inverseErf**
- **inverseErfc**
- **Sort**
- **StatsCircularMoments**
- **StatsCorrelation**
- **StatsMedian**
- **StatsQuantiles**
- **StatsResample**
- **StatsTrimmedMean**
- **StudentA**
- **StudentT**
- **WaveStats**
- **StatsPermute**

Hazard and Survival Functions

Igor does not provide built-in functions to calculate the Survival or Hazard functions. They can be calculated easily from the **Probability Distribution Functions** on page III-345 and **Cumulative Distribution Functions** on page III-344.

In the following, the cumulative distribution functions are denoted by $F(x)$ and the probability distribution functions are denoted by $p(x)$.

The Survival Function $S(x)$ is given by

$$S(x) = 1 - F(x).$$

The Hazard function $h(x)$ is given by

$$h(x) = \frac{p(x)}{S(x)} = \frac{p(x)}{1 - F(x)}.$$

The cumulative hazard function $H(x)$ is

$$H(x) = \int_{-\infty}^{x} h(u)\,du,$$

$$H(x) = -\ln\left[1 - F(x)\right].$$

Inverse Survival Function $Z(a)$ is

$$Z(\alpha) = G(1 - \alpha),$$

where $G()$ is the inverse CDF (see **Inverse Cumulative Distribution Functions** on page III-345).

Statistics Procedures

Several procedure files are provided to extend the built-in statistics capability described in this chapter. Some of these procedure files provide user interfaces to the built-in statistics functionality. Others extend the functionality.

In the Analysis menu you will find a Statistics item that brings up a submenu. Selecting any item in the submenu will cause all the statistics-related procedure files to be loaded, making them ready to use. Alternatively, you can load all the statistics procedures by adding the following include statement to the top of your procedure window:

```
#include <AllStatsProcedures>
```

Functionality provided by the statistics procedure files includes the 1D Statistics Report package for automatic analysis of single 1D waves, and the ANOVA Power Calculations Panel, as well as functions to create specialized graphs:

StatsAutoCorrPlot()
StatsBoxPlot()
StatsPlotLag()
StatsProbPlot()
StatsPlotHistogram()

Also included are these convenience functions:

WM_2MeanConfidenceIntervals()
WM_2MeanConfidenceIntervals2()
WM_BernoulliCdf()
WM_BinomialPdf()
WM_CIforPooledMean()
WM_CompareCorrelations()
WM_EstimateMinDetectableDiff()
WM_EstimateReqSampleSize()
WM_EstimateReqSampleSize2()
WM_EstimateSampleSizeForDif()
WM_GetANOVA1Power()
WM_GetGeometricAverage()
WM_GetHarmonicMean()
WM_GetPooledMean()
WM_GetPooledVariance()
WM_MCPointOnRegressionLines()
WM_MeanConfidenceInterval()
WM_OneTailStudentA()
WM_OneTailStudentT()
WM_PlotBiHistogram()
WM_RankForTies()
WM_RankLetterGradesWithTies()
WM_RegressionInversePrediction()
WM_SSEstimatorFunc()
WM_SSEstimatorFunc2()
WM_SSEstimatorFunc3()
WM_VarianceConfidenceInterval()
WM_WilcoxonPairedRanks()
WM_StatsKaplanMeier()

Statistics References

Ajne, B., A simple test for uniformity of a circular distribution, *Biometrica*, *55*, 343-354, 1968.

Bradley, J.V., *Distribution-Free Statistical Tests*, Prentice Hall, Englewood Cliffs, New Jersey, 1968.

Cheung, Y.K., and J.H. Klotz, The Mann Whitney Wilcoxon distribution using linked lists, *Statistica Sinica*, *7*, 805-813, 1997.

Copenhaver, M.D., and B.S. Holland, Multiple comparisons of simple effects in the two-way analysis of variance with fixed effects, *Journal of Statistical Computation and Simulation*, *30*, 1-15, 1988.

Evans, M., N. Hastings, and B. Peacock, *Statistical Distributions*, 3rd ed., Wiley, New York, 2000.

Fisher, N.I., *Statistical Analysis of Circular Data*, 295pp., Cambridge University Press, New York, 1995.

Iman, R.L., and W.J. Conover, A measure of top-down correlation, *Technometrics*, *29*, 351-357, 1987.

Kendall, M.G., *Rank Correlation Methods*, 3rd ed., Griffin, London, 1962.

Klotz, J.H., *Computational Approach to Statistics*, <`http://www.stat.wisc.edu/~klotz/Book.pdf`>.

Moore, B.R., A modification of the Rayleigh test for vector data, *Biometrica*, *67*, 175-180, 1980.

Press, William H., *et al.*, *Numerical Recipes in C*, 2nd ed., 994 pp., Cambridge University Press, New York, 1992.

van de Wiel, M.A., and A. Di Bucchianico, Fast computation of the exact null distribution of Spearman's rho and Page's L statistic for samples with and without ties, *J. of Stat. Plan. and Inference*, *92*, 133-145, 2001.

Wallace, D.L., Simplified Beta-Approximation to the Kruskal-Wallis H Test, *J. Am. Stat. Assoc.*, *54*, 225-230, 1959.

Zar, J.H., *Biostatistical Analysis*, 4th ed., 929 pp., Prentice Hall, Englewood Cliffs, New Jersey, 1999.