NAG Toolbox for MATLAB

g08cc

1 Purpose

g08cc performs the one sample Kolmogorov–Smirnov distribution test, using a user-specified distribution.

2 Syntax

```
[d, z, p, sx, ifail] = g08cc(x, cdf, ntype, 'n', n)
```

3 Description

The data consists of a single sample of n observations, denoted by $x_1, x_2, \dots, x_n$. Let $S_n\left(x_{(i)}\right)$ and $F_0\left(x_{(i)}\right)$ represent the sample cumulative distribution function and the theoretical (null) cumulative distribution function respectively at the point $x_{(i)}$, where $x_{(i)}$ is the ith smallest sample observation.

The Kolmogorov–Smirnov test provides a test of the null hypothesis H_0: the data are a random sample of observations from a theoretical distribution specified by you (in user-supplied real function **cdf**) against one of the following alternative hypotheses.

(i) H_1 : the data cannot be considered to be a random sample from the specified null distribution.

(ii) H_2 : the data arise from a distribution which dominates the specified null distribution. In practical terms, this would be demonstrated if the values of the sample cumulative distribution function $S_n(x)$ tended to exceed the corresponding values of the theoretical cumulative distribution function $F_{0(x)}$.

(iii) H_3 : the data arise from a distribution which is dominated by the specified null distribution. In practical terms, this would be demonstrated if the values of the theoretical cumulative distribution function $F_0(x)$ tended to exceed the corresponding values of the sample cumulative distribution function $S_n(x)$.

One of the following test statistics is computed depending on the particular alternative hypothesis specified (see the description of the parameter **ntype** in Section 5).

For the alternative hypothesis H_1 :

> D_n – the largest absolute deviation between the sample cumulative distribution function and the theoretical cumulative distribution function. Formally $D_n = \max\left\{D_n^+, D_n^-\right\}$.

For the alternative hypothesis H_2 :

> D_n^+ – the largest positive deviation between the sample cumulative distribution function and the theoretical cumulative distribution function. Formally $D_n^+ = \max\left\{S_n\left(x_{(i)}\right) - F_0\left(x_{(i)}\right), 0\right\}$.

For the alternative hypothesis H_3 :

> D_n^- – the largest positive deviation between the theoretical cumulative distribution function and the sample cumulative distribution function. Formally $D_n^- = \max\left\{F_0\left(x_{(i)}\right) - S_n\left(x_{(i-1)}\right), 0\right\}$. This is only true for continuous distributions. See Section 8 for comments on discrete distributions.

The standardized statistic, $Z = D \times \sqrt{n}$, is also computed, where D may be D_n, D_n^+ or D_n^- depending on the choice of the alternative hypothesis. This is the standardized value of D with no continuity correction applied and the distribution of Z converges asymptotically to a limiting distribution, first derived by Kolmogorov 1933, and then tabulated by Smirnov 1948. The asymptotic distributions for the one-sided statistics were obtained by Smirnov 1933.

The probability, under the null hypothesis, of obtaining a value of the test statistic as extreme as that observed, is computed. If $n \leq 100$, an exact method given by Conover 1980 is used. Note that the method used is only exact for continuous theoretical distributions and does not include Conover's modification for discrete distributions. This method computes the one-sided probabilities. The two-sided probabilities are estimated by doubling the one-sided probability. This is a good estimate for small p, that is $p \leq 0.10$, but it becomes very poor for larger p. If $n > 100$ then p is computed using the Kolmogorov–Smirnov limiting distributions; see Feller 1948, Kendall and Stuart 1973, Kolmogorov 1933, Smirnov 1933 and Smirnov 1948.

4 References

Conover W J 1980 *Practical Nonparametric Statistics* Wiley

Feller W 1948 On the Kolmogorov–Smirnov limit theorems for empirical distributions *Ann. Math. Statist.* **19** 179–181

Kendall M G and Stuart A 1973 *The Advanced Theory of Statistics (Volume 2)* (3rd Edition) Griffin

Kolmogorov A N 1933 Sulla determinazione empirica di una legge di distribuzione *Giornale dell' Istituto Italiano degli Attuari* **4** 83–91

Siegel S 1956 *Non-parametric Statistics for the Behavioral Sciences* McGraw–Hill

Smirnov N 1933 Estimate of deviation between empirical distribution functions in two independent samples *Bull. Moscow Univ.* **2 (2)** 3–16

Smirnov N 1948 Table for estimating the goodness of fit of empirical distributions *Ann. Math. Statist.* **19** 279–281

5 Parameters

5.1 Compulsory Input Parameters

1: **x(n) – double array**

The sample observations $x_1, x_2, \ldots, x_n$.

2: **cdf – string containing name of m-file**

cdf must return the value of the theoretical (null) cumulative distribution function for a given value of its argument.

Its specification is:

```
[result] = cdf(x)
```

Input Parameters

1: **x – double scalar**

The argument for which **cdf** must be evaluated.

Output Parameters

1: **result – double scalar**

The result of the function.

Constraint: **cdf** must always return a value in the range $[0.0, 1.0]$ and **cdf** must always satify the condition that $\mathbf{cdf}(x_1) \leq \mathbf{cdf}(x_2)$ for any $x_1 \leq x_2$

3: **ntype – int32 scalar**

The statistic to be calculated, i.e., the choice of alternative hypothesis.

ntype $= 1$

Computes D_n, to test H_0 against H_1.

ntype $= 2$

Computes D_n^+, to test H_0 against H_2.

ntype $= 3$

Computes D_n^-, to test H_0 against H_3.

Constraint: **ntype** $= 1$, 2 or 3.

5.2 Optional Input Parameters

1: **n – int32 scalar**

Default: The dimension of the arrays **x**, **sx**. (An error is raised if these dimensions are not equal.)

n, the number of observations in the sample.

Constraint: $\mathbf{n} \geq 1$.

5.3 Input Parameters Omitted from the MATLAB Interface

None.

5.4 Output Parameters

1: **d – double scalar**

The Kolmogorov–Smirnov test statistic (D_n, D_n^+ or D_n^- according to the value of **ntype**).

2: **z – double scalar**

A standardized value, Z, of the test statistic, D, without the continuity correction applied.

3: **p – double scalar**

The probability, p, associated with the observed value of D, where D may D_n, D_n^+ or D_n^- depending on the value of **ntype** (see Section 3).

4: **sx(n) – double array**

The sample observations, $x_1, x_2, \ldots, x_n$, sorted in ascending order.

5: **ifail – int32 scalar**

0 unless the function detects an error (see Section 6).

6 Error Indicators and Warnings

Errors or warnings detected by the function:

ifail $= 1$

On entry, $\mathbf{n} < 1$.

ifail $= 2$

On entry, **ntype** $\neq 1$, 2 or 3.

ifail $= 3$

The supplied theoretical cumulative distribution function returns a value less than 0.0 or greater than 1.0, thereby violating the definition of the cumulative distribution function.

ifail $= 4$

The supplied theoretical cumulative distribution function is not a nondecreasing function thereby violating the definition of a cumulative distribution function, that is $F_0(x) > F_0(y)$ for some $x < y$.

7 Accuracy

For most cases the approximation for p given when $n > 100$ has a relative error of less than 0.01. The two-sided probability is approximated by doubling the one-sided probability. This is only good for small p, that is $p < 0.10$, but very poor for large p. The error is always on the conservative side.

8 Further Comments

The time taken by g08cc increases with n until $n > 100$ at which point it drops and then increases slowly.

For a discrete theoretical cumulative distribution function $F_0(x)$, $D_n^- = \max\left\{F_0\left(x_{(i)}\right) - S_n\left(x_{(i)}\right), 0\right\}$. Thus if you wish to provide a discrete distribution function the following adjustment needs to be made,

for D_n^+, return $F(x)$ as x as usual;

for D_n^-, return $F(x - d)$ at x where d is the discrete jump in the distribution. For example $d = 1$ for the Poisson or Binomial distributions.

9 Example

```
g08cc_cdf.m
```

```
function [result] = cdf(x)

  if x < 0
    result = 0;
  elseif x > 2
    result = 1;
  else
    result = x/2;
  end
```

```
x = [0.01;
     0.3;
     0.2;
     0.9;
     1.2;
     0.09;
     1.3;
     0.18;
     0.9;
     0.48;
     1.98;
     0.03;
     0.5;
     0.07000000000000001;
     0.7;
     0.6;
     0.95;
     1;
     0.31;
     1.45;
     1.04;
     1.25;
     0.15;
     0.75;
     0.85;
     0.22;
     1.56;
     0.8100000000000001;
     0.57;
     0.55];
ntype = int32(1);
[d, z, p, sx, ifail] = g08cc(x, 'g08cc_cdf', ntype)
```

```
d =
    0.2800
z =
    1.5336
p =
    0.0143
sx =
    0.0100
    0.0300
    0.0700
    0.0900
    0.1500
    0.1800
    0.2000
    0.2200
    0.3000
    0.3100
    0.4800
    0.5000
    0.5500
    0.5700
    0.6000
    0.7000
    0.7500
    0.8100
    0.8500
    0.9000
    0.9000
    0.9500
    1.0000
    1.0400
    1.2000
    1.2500
    1.3000
    1.4500
    1.5600
    1.9800
ifail =
           0
```