

## NAG Toolbox

### nag\_nonpar\_test\_ks\_1sample\_user (g08cc)

#### 1 Purpose

`nag_nonpar_test_ks_1sample_user (g08cc)` performs the one sample Kolmogorov–Smirnov distribution test, using a user-specified distribution.

#### 2 Syntax

```
[d, z, p, sx, ifail] = nag_nonpar_test_ks_1sample_user(x, cdf, ntype, 'n', n)
[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  $S_n(x_{(i)})$  and  $F_0(x_{(i)})$  represent the sample cumulative distribution function and the theoretical (null) cumulative distribution function respectively at the point  $x_{(i)}$ , where  $x_{(i)}$  is the  $i$ th 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 **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 argument **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  $D_n = \max\{D_n^+, D_n^-\}$ .

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  $D_n^+ = \max\{S_n(x_{(i)}) - F_0(x_{(i)}), 0\}$ .

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  $D_n^- = \max\{F_0(x_{(i)}) - S_n(x_{(i-1)}), 0\}$ . This is only true for continuous distributions. See Section 9 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)** – REAL (KIND=nag\_wp) array

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

2: **cdf** – REAL (KIND=nag\_wp) FUNCTION, supplied by the user.

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

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

#### Input Parameters

1: **x** – REAL (KIND=nag\_wp)

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

#### Output Parameters

1: **result**

The value of the theoretical (null) cumulative distribution function evaluated at **x**.

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

3: **ntype** – INTEGER

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** – INTEGER

*Default:* the dimension of the array **x**.

$n$ , the number of observations in the sample.

*Constraint:* **n**  $\geq$  1.

## 5.3 Output Parameters

1: **d** – REAL (KIND=nag\_wp)

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

2: **z** – REAL (KIND=nag\_wp)

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

3: **p** – REAL (KIND=nag\_wp)

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)** – REAL (KIND=nag\_wp) array

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

5: **ifail** – INTEGER

**ifail** = 0 unless the function detects an error (see Section 5).

## 6 Error Indicators and Warnings

Errors or warnings detected by the function:

**ifail** = 1

On entry, **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  $x < y$ .

**ifail** = -99

An unexpected error has been triggered by this routine. Please contact NAG.

**ifail** = -399

Your licence key may have expired or may not have been installed correctly.

**ifail** = -999

Dynamic memory allocation failed.

## 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 `nag_nonpar_test_ks_1sample_user` (g08cc) increases with  $n$  until  $n > 100$  at which point it drops and then increases slowly.

For a discrete theoretical cumulative distribution function  $F_0(x)$ ,  $D_n^- = \max\{F_0(x_{(i)}) - S_n(x_{(i)}), 0\}$ . 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

The following example performs the one sample Kolmogorov–Smirnov test to test whether a sample of 30 observations arise firstly from a uniform distribution  $U(0, 1)$  or secondly from a Normal distribution with mean 0.75 and standard deviation 0.5. The two-sided test statistic,  $D_n$ , the standardized test statistic,  $Z$ , and the upper tail probability,  $p$ , are computed and then printed for each test.

### 9.1 Program Text

```
function g08cc_example

fprintf('g08cc example results\n\n');

global xmean std;
xmean = 0.75;
std    = 0.5;

x = [0.01; 0.30; 0.20; 0.90; 1.20; 0.09; 1.30; 0.18; 0.90; 0.48;
     1.98; 0.03; 0.50; 0.07; 0.70; 0.60; 0.95; 1.00; 0.31; 1.45;
     1.04; 1.25; 0.15; 0.75; 0.85; 0.22; 1.56; 0.81; 0.57; 0.55];

% Parameters
ntype = nag_int(1);

[d, z, p, sx, ifail] = g08cc( ...
    x, @cdf, ntype);
```

```
fprintf('Test against normal distribution:\n');
fprintf('          mean = %7.2f\n', xmean);
fprintf('          standard deviation = %7.2f\n', std);

fprintf('\n\nTest statistic D = %8.4f\n', d);
fprintf('Z statistic      = %8.4f\n', z);
fprintf('Tail probability = %8.4f\n', p);

function [result] = cdf(x)
% Cumulative distribution function.
% Here: for normal distribution, with mean = 0.75 and s.d. = 0.5
global xmean std;

z = (x-xmean)/std;
[result,ifail] = s15ab(z);
```

## 9.2 Program Results

g08cc example results

```
Test against normal distribution:
          mean =      0.75
          standard deviation =    0.50
```

```
Test statistic D =    0.1439
Z statistic      =    0.7882
Tail probability =    0.5262
```

---