NAG Toolbox

nag_nonpar_test_ks_2sample (g08cd)

1 Purpose

nag_nonpar_test_ks_2sample (g08cd) performs the two sample Kolmogorov–Smirnov distribution test.

2 Syntax

```
[d, z, p, sx, sy, ifail] = nag_nonpar_test_ks_2sample(x, y, ntype, 'n1', n1, 'n2', n2)
[d, z, p, sx, sy, ifail] = g08cd(x, y, ntype, 'n1', n1, 'n2', n2)
```

3 Description

The data consists of two independent samples, one of size n_1, denoted by $x_1, x_2, \ldots, x_{n_1}$, and the other of size n_2 denoted by $y_1, y_2, \ldots, y_{n_2}$. Let $F(x)$ and $G(x)$ represent their respective, unknown, distribution functions. Also let $S_1(x)$ and $S_2(x)$ denote the values of the sample cumulative distribution functions at the point x for the two samples respectively.

The Kolmogorov–Smirnov test provides a test of the null hypothesis H_0: $F(x) = G(x)$ against one of the following alternative hypotheses:

(i) H_1: $F(x) \neq G(x)$.

(ii) H_2: $F(x) > G(x)$. This alternative hypothesis is sometimes stated as, 'The x's tend to be smaller than the y's', i.e., it would be demonstrated in practical terms if the values of $S_1(x)$ tended to exceed the corresponding values of $S_2(x)$.

(iii) H_3: $F(x) < G(x)$. This alternative hypothesis is sometimes stated as, 'The x's tend to be larger than the y's', i.e., it would be demonstrated in practical terms if the values of $S_2(x)$ tended to exceed the corresponding values of $S_1(x)$.

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

For the alternative hypothesis H_1.

D_{n_1,n_2} – the largest absolute deviation between the two sample cumulative distribution functions.

For the alternative hypothesis H_2.

$D^+_{n_1,n_2}$ – the largest positive deviation between the sample cumulative distribution function of the first sample, $S_1(x)$, and the sample cumulative distribution function of the second sample, $S_2(x)$. Formally $D^+_{n_1,n_2} = \max\{S_1(x) - S_2(x), 0\}$.

For the alternative hypothesis H_3.

$D^-_{n_1,n_2}$ – the largest positive deviation between the sample cumulative distribution function of the second sample, $S_2(x)$, and the sample cumulative distribution function of the first sample, $S_1(x)$. Formally $D^-_{n_1,n_2} = \max\{S_2(x) - S_1(x), 0\}$.

nag_nonpar_test_ks_2sample (g08cd) also returns the standardized statistic $Z = \sqrt{\frac{n_1+n_2}{n_1 n_2}} \times D$, where D may be D_{n_1,n_2}, $D^+_{n_1,n_2}$ or $D^-_{n_1,n_2}$ depending on the choice of the alternative hypothesis. The distribution of this statistic converges asymptotically to a distribution given by Smirnov as n_1 and n_2 increase; see Feller (1948), Kendall and Stuart (1973), Kim and Jenrich (1973), Smirnov (1933) or Smirnov (1948).

The probability, under the null hypothesis, of obtaining a value of the test statistic as extreme as that observed, is computed. If $\max(n_1, n_2) \leq 2500$ and $n_1 n_2 \leq 10000$ then an exact method given by Kim and Jenrich (see Kim and Jenrich (1973)) is used. Otherwise p is computed using the approximations

suggested by Kim and Jenrich (1973). Note that the method used is only exact for continuous theoretical distributions. This method computes the two-sided probability. The one-sided probabilities are estimated by halving the two-sided probability. This is a good estimate for small p, that is $p \leq 0.10$, but it becomes very poor for larger p.

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

Kim P J and Jenrich R I (1973) Tables of exact sampling distribution of the two sample Kolmogorov–Smirnov criterion $D_{mn}(m < n)$ *Selected Tables in Mathematical Statistics* **1** 80–129 American Mathematical Society

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**(**n1**) – REAL (KIND=nag_wp) array

The observations from the first sample, $x_1, x_2, \ldots, x_{n_1}$.

2: **y**(**n2**) – REAL (KIND=nag_wp) array

The observations from the second sample, $y_1, y_2, \ldots, y_{n_2}$.

3: **ntype** – INTEGER

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

ntype $= 1$
Computes $D_{n_1 n_2}$, to test against H_1.

ntype $= 2$
Computes $D^+_{n_1 n_2}$, to test against H_2.

ntype $= 3$
Computes $D^-_{n_1 n_2}$, to test against H_3.

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

5.2 Optional Input Parameters

1: **n1** – INTEGER

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

The number of observations in the first sample, n_1.

Constraint: **n1** ≥ 1.

2: **n2** – INTEGER

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

The number of observations in the second sample, n_2.

Constraint: **n2** ≥ 1.

5.3 Output Parameters

1: **d** – REAL (KIND=nag_wp)

The Kolmogorov–Smirnov test statistic ($D_{n_1 n_2}$, $D^+_{n_1 n_2}$ or $D^-_{n_1 n_2}$ according to the value of **ntype**).

2: **z** – REAL (KIND=nag_wp)

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

3: **p** – REAL (KIND=nag_wp)

The tail probability associated with the observed value of D, where D may be $D_{n_1,n_2}, D^+_{n_1,n_2}$ or $D^-_{n_1,n_2}$ depending on the value of **ntype** (see Section 3).

4: **sx**(**n1**) – REAL (KIND=nag_wp) array

The observations from the first sample sorted in ascending order.

5: **sy**(**n2**) – REAL (KIND=nag_wp) array

The observations from the second sample sorted in ascending order.

6: **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, **n1** < 1,
or **n2** < 1.

ifail $= 2$

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

ifail $= 3$

The iterative procedure used in the approximation of the probability for large n_1 and n_2 did not converge. For the two-sided test, $p = 1$ is returned. For the one-sided test, $p = 0.5$ is returned.

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

The large sample distributions used as approximations to the exact distribution should have a relative error of less than 5% for most cases.

8 Further Comments

The time taken by nag_nonpar_test_ks_2sample (g08cd) increases with n_1 and n_2, until $n_1 n_2 > 10000$ or $\max(n_1, n_2) \geq 2500$. At this point one of the approximations is used and the time decreases significantly. The time then increases again modestly with n_1 and n_2.

9 Example

This example computes the two-sided Kolmogorov–Smirnov test statistic for two independent samples of size 100 and 50 respectively. The first sample is from a uniform distribution $U(0, 2)$. The second sample is from a uniform distribution $U(0.25, 2.25)$. The test statistic, D_{n_1,n_2}, the standardized test statistic, Z, and the tail probability, p, are computed and printed.

9.1 Program Text

```
     function g08cd_example

fprintf('g08cd example results\n\n');

x = [ 1.160 1.785 0.322 1.437 1.695 1.770 1.209 0.479 1.122 0.974 ...
      0.290 1.155 0.218 1.595 1.053 1.058 1.282 1.278 1.066 0.725 ...
      0.113 1.516 1.329 1.907 0.101 0.387 1.392 0.613 0.692 1.397 ...
      1.627 0.417 1.079 0.607 0.899 0.493 0.381 1.660 0.233 0.718 ...
      1.376 1.395 1.557 1.610 1.632 0.851 1.824 0.921 0.139 0.618 ...
      0.050 0.956 0.669 1.109 1.882 1.462 1.465 0.201 1.036 1.127 ...
      0.907 0.876 1.199 1.667 1.141 0.820 0.488 0.732 0.725 0.753 ...
      0.760 1.833 0.074 1.101 0.620 1.858 0.681 0.705 0.876 1.096 ...
      1.870 1.597 0.990 0.430 0.410 0.399 1.693 0.492 1.318 0.883 ...
      1.291 1.051 1.934 1.314 1.496 0.391 1.079 0.881 0.983 1.306];

y = [ 1.695 1.452 0.997 1.771 1.114 1.624 2.005 0.782 1.870 0.954 ...
      1.606 2.059 0.774 0.741 1.040 0.521 2.163 0.818 1.781 1.420 ...
      0.558 1.437 2.004 1.325 0.398 0.582 2.047 0.332 1.186 0.890 ...
      1.825 1.324 1.334 0.261 0.299 1.733 1.172 1.000 1.663 1.093 ...
      1.045 2.022 1.174 0.670 1.143 1.189 0.494 1.275 1.122 1.823];

ntype = nag_int(1);
[d, z, p, sx, sy, ifail] = g08cd(...
                                 x, y, ntype);

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

9.2 Program Results

```
     g08cd example results

Test statistic D =   0.1800
Z statistic      =   0.0312
Tail probability =   0.2222
```