NAG Toolbox for MATLAB

g08cd

1 Purpose

g08cd performs the two sample Kolmogorov–Smirnov distribution test.

2 Syntax

```
[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 parameter **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 $D^-_{n_1,n_2} = \max\{S_2(x) - S_1(x), 0\}$.

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) – double array**

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

2: **y(n2) – double array**

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

3: **ntype – int32 scalar**

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 – int32 scalar**

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

the number of observations in the first sample, n_1.

Constraint: **n1** ≥ 1.

2: **n2 – int32 scalar**

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

the number of observations in the second sample, n_2.

Constraint: $\mathbf{n2} \geq 1$.

5.3 Input Parameters Omitted from the MATLAB Interface

None.

5.4 Output Parameters

1: **d – double scalar**

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

2: **z – double scalar**

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

3: **p – double scalar**

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) – double array**

The observations from the first sample sorted in ascending order.

5: **sy(n2) – double array**

The observations from the second sample sorted in ascending order.

6: **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, $\mathbf{n1} < 1$,
or $\mathbf{n2} < 1$.

ifail $= 2$

On entry, $\mathbf{ntype} \neq 1$, 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.

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 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

```
x = [1.5902480498365;
     0.4514344715575777;
     0.7425605404715657;
     0.4500701410943835;
     1.757489613962739;
     0.09494659709066995;
     0.3611426417219349;
     0.8655310991929152;
     0.07874046899432821;
     1.15034115657093;
     1.909225474130498;
     1.248205642566939;
     1.278136611206321;
     1.376242882235565;
     0.4024793749242044;
     1.762011956845783;
     0.6952326353936773;
     0.9876414433367825;
     0.6059391432983038;
     0.2078601007360235;
     1.564305035404043;
     1.746505783532216;
     0.1778870249628505;
     0.1076281399262869;
     0.07704165044712488;
     0.04567587466526592;
     0.8039773560869385;
     1.241984667916896;
     1.797786082814494;
     0.852819977142019;
     1.747622472821148;
     1.912044530456294;
     0.4706479344188908;
     1.243486362938038;
     1.279342351641843;
     0.8903608734478126;
     1.727202408903865;
     0.2581867825828414;
     0.9415224497978657;
     1.74941093213823;
     1.20905525205517;
     0.9257769700555145;
     0.9774905558012059;
     0.8557640721625349;
     0.2648216307181972;
     0.2305151298211215;
     1.214582160870085;
     0.0008087845523660937;
     0.04826634252569287;
     1.422324123840062;
     1.64461234171838;
     1.242902747480194;
     1.859348219383043;
     0.8759548065432361;
     0.9770887783336585;
     1.077045946574884;
     0.07040286404180876;
     0.725208635893021;
```

```
      1.129721212172624;
      0.7794878837807858;
      1.236501744194477;
      0.8327862848080944;
      1.932002357334167;
      0.01096554652191538;
      1.430490703604279;
      0.9004992406397488;
      0.8208346310843287;
      0.7213215653615261;
      1.032756476884525;
      0.3378646599255028;
      1.975473898973496;
      1.307202749307935;
      1.022550018984354;
      0.5703968003457041;
      1.174687331604432;
      1.691927534708742;
      0.09038851332572448;
      0.08328616335992274;
      0.3974243748119088;
      0.4000804015026145;
      1.734779923254609;
      0.072131066247919;
      1.330366567063721;
      0.0890099010719363 3;
      0.6128580701836494;
      1.25978558572682;
      0.3681357787827774;
      1.766612133578424;
      0.5447167419613813;
      0.1468692016200379;
      0.08565369904623682;
      1.73695078139111;
      1.502452715832918;
      1.023663015528883;
      1.548981938796881;
      0.06591087486198742;
      1.089678519368579;
      0.7863284085754246;
      0.4083539838920379;
      0.1336746295053808];
y = [0.3959382482212902;
      2.189059814171972;
      0.5309091506354565;
      1.398213120735397;
      1.226993173102978;
      2.092465206711672;
      1.376152675038049;
      0.2635476162603962;
      0.6826139872981392;
      1.725437958005721;
      1.218220941115947;
      0.2907559334793868;
      0.2888496459940526;
      2.237322438311175;
      1.335206853771786;
      1.482575912633977;
      1.511554656117706;
      1.49444503365132;
      1.647783675961666;
      1.093415807808477;
      1.929516829926758;
      1.771633171736251;
      1.362424985201225;
      0.7600483503138998;
      1.474848032784191;
      1.579250523709051;
      1.100106550316349;
      0.9988674396189126;
```

```
      1.974550189113563;
      0.6728895918936394;
      0.6568563195351251;
      1.038967939288667;
      1.420212415024878;
      2.156198816080763;
      1.879016986416847;
      1.409577351753394;
      1.953857139342458;
      0.3220659541919986;
      2.244466171121614;
      1.738164359951565;
      0.4534069003591212;
      0.8715218615975232;
      1.596006863571959;
      1.501743503682336;
      2.099525595470502;
      0.8538189862774034;
      0.9554670591088673;
      1.668907067970278;
      1.517236797575171;
      1.478170841410264];
ntype = int32(1);
[d, z, p, sx, sy, ifail] = g08cd(x, y, ntype)
```

```
d =
    0.3600
z =
    0.0624
p =
   2.8438e-04
sx =
     array elided
sy =
     array elided
ifail =
           0
```