NAG Fortran Library Routine Document

G08CDF

Note: before using this routine, please read the Users' Note for your implementation to check the interpretation of ***bold italicised*** terms and other implementation-dependent details.

1 Purpose

G08CDF performs the two sample Kolmogorov–Smirnov distribution test.

2 Specification

```
SUBROUTINE G08CDF(N1, X, N2, Y, NTYPE, D, Z, P, SX, SY, IFAIL)
INTEGER            N1, N2, NTYPE, IFAIL
real               X(N1), Y(N2), D, Z, P, SX(N1), SY(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 $D^-_{n_1,n_2} = \max\{S_2(x) - S_1(x), 0\}$.

G08CDF 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) *Nonparametric 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

1: N1 – INTEGER *Input*

On entry: the number of observations in the first sample, n_1.

Constraint: N1 ≥ 1.

2: X(N1) – ***real*** array *Input*

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

3: N2 – INTEGER *Input*

On entry: the number of observations in the second sample, n_2.

Constraint: N2 ≥ 1.

4: Y(N2) – ***real*** array *Input*

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

5: NTYPE – INTEGER *Input*

On entry: 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.

6: D – ***real*** *Output*

On exit: 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).

7: Z – ***real*** *Output*

On exit: a standardized value Z of the test statistic, D, without any correction for continuity.

8: P – ***real*** *Output*

On exit: 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).

9: SX(N1) – ***real*** array *Output*

On exit: the observations from the first sample sorted in ascending order.

10: SY(N2) – ***real*** array *Output*

On exit: the observations from the second sample sorted in ascending order.

11: IFAIL – INTEGER *Input/Output*

On entry: IFAIL must be set to 0, -1 or 1. Users who are unfamiliar with this parameter should refer to Chapter P01 for details.

On exit: IFAIL $= 0$ unless the routine detects an error (see Section 6).

For environments where it might be inappropriate to halt program execution when an error is detected, the value -1 or 1 is recommended. If the output of error messages is undesirable, then the value 1 is recommended. Otherwise, for users not familiar with this parameter the recommended value is 0. **When the value -1 or 1 is used it is essential to test the value of IFAIL on exit.**

6 Error Indicators and Warnings

If on entry IFAIL $= 0$ or -1, explanatory error messages are output on the current error message unit (as defined by X04AAF).

Errors or warnings detected by the routine:

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.

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 the routine 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

The following 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

Note: the listing of the example program presented below uses ***bold italicised*** terms to denote precision-dependent details. Please read the Users' Note for your implementation to check the interpretation of these terms. As explained in the Essential Introduction to this manual, the results produced may not be identical for all implementations.

```
*     G08CDF Example Program Text
*     Mark 20 Revised. NAG Copyright 2001.
*     Mark 20 Revised.  To call thread-safe G05 routines.
*     .. Parameters ..
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
      INTEGER          NMAX, MMAX
      PARAMETER        (NMAX=100,MMAX=50)
*     .. Local Scalars ..
      real             D, P, Z
      INTEGER          IFAIL, IGEN, M, N, NTYPE
*     .. Local Arrays ..
      real             SX(NMAX), SY(MMAX), X(NMAX), Y(MMAX)
      INTEGER          ISEED(4)
*     .. External Subroutines ..
      EXTERNAL         G05KBF, G05LGF, G08CDF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G08CDF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) N, M
      WRITE (NOUT,*)
      IF (N.LE.NMAX .AND. M.LE.MMAX) THEN
         IGEN = 0
         ISEED(1) = 0
         CALL G05KBF(IGEN,ISEED)
         IFAIL = 0
         CALL G05LGF(0.0e0,2.0e0,N,X,IGEN,ISEED,IFAIL)
         CALL G05LGF(0.25e0,2.25e0,M,Y,IGEN,ISEED,IFAIL)
         READ (NIN,*) NTYPE
         IFAIL = -1
*
         CALL G08CDF(N,X,M,Y,NTYPE,D,Z,P,SX,SY,IFAIL)
*
         IF (IFAIL.NE.0) WRITE (NOUT,99999) '** IFAIL = ', IFAIL
         WRITE (NOUT,99998) 'Test statistic D = ', D
         WRITE (NOUT,99998) 'Z statistic      = ', Z
         WRITE (NOUT,99998) 'Tail probability = ', P
      ELSE
         WRITE (NOUT,99997) 'N or M is out of range: N = ', N,
     +     ' and M = ', M
      END IF
      STOP
*
99999 FORMAT (1X,A,I2)
99998 FORMAT (1X,A,F8.4)
99997 FORMAT (1X,A,I7,A,I7)
      END
```

9.2 Program Data

```
G08CDF Example Program Data
 100 50
 1
```

9.3 Program Results

```
 G08CDF Example Program Results

 Test statistic D =   0.3600
 Z statistic      =   0.0624
 Tail probability =   0.0003
```