G08CCF – NAG Fortran Library Routine Document

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

G08CCF performs the one sample Kolmogorov–Smirnov distribution test, using a user-specified distribution.

2 Specification

```
SUBROUTINE G08CCF(N, X, CDF, NTYPE, D, Z, P, SX, IFAIL)
INTEGER           N, NTYPE, IFAIL
real              X(N), CDF, D, Z, P, SX(N)
EXTERNAL          CDF
```

3 Description

The data consist of a single sample of n observations, denoted by $x_1, x_2, \ldots, x_n$.

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 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 the user (in the 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 $D_n = \max\{D_n^+, D_n^-\}$.

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 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 [4], and then tabulated by Smirnov [6]. The asymptotic distributions for the one-sided statistics were obtained by Smirnov [5].

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 [1], 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 [2], Kendall and Stuart [3], Kolmogorov [4], Smirnov [5] and [6].

4 References

[1] Conover W J (1980) *Practical Nonparametric Statistics* Wiley

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

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

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

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

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

[7] Siegel S (1956) *Non-parametric Statistics for the Behavioral Sciences* McGraw–Hill

5 Parameters

1: N — INTEGER *Input*

On entry: the number of observations in the sample, n.

Constraint: $N \geq 1$.

2: X(N) — ***real*** array *Input*

On entry: the sample observations $x_1, x_2, \ldots, x_n$.

3: CDF — ***real*** FUNCTION, supplied by the user. *External Procedure*

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

Its specification is:

```
real FUNCTION CDF(X)
real             X
```

1: X — ***real*** *Input*

On entry: the argument for which CDF must be evaluated.

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

CDF must be declared as EXTERNAL in the (sub)program from which G08CCF is called. Parameters denoted as *Input* must **not** be changed by this procedure.

4: NTYPE — INTEGER *Input*

On entry: 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: D — ***real*** *Output*

On exit: the Kolmogorov-Smirnov test statistic (D_n, D_n^+ or D_n^- according to the value of NTYPE).

6: Z — ***real*** *Output*

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

7: P — ***real*** *Output*

On exit: 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).

8: SX(N) — ***real*** array *Output*

On exit: the sample observations, $x_1, x_2, \ldots, x_n$, sorted in ascending order.

9: IFAIL — INTEGER *Input/Output*

On entry: IFAIL must be set to 0, −1 or 1. For users not familiar with this parameter (described in Chapter P01) the recommended value is 0.

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

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 detected by the routine:

IFAIL = 1

On entry, N < 1.

IFAIL = 2

On entry, NTYPE ≠ 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 non-decreasing 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 the routine 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 the user wishes 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

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.

```
*     G08CCF Example Program Text
*     Mark 14 Release.  NAG Copyright 1989.
*     .. Parameters ..
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
      INTEGER          NMAX
      PARAMETER        (NMAX=30)
*     .. Local Arrays ..
      real             SX(NMAX), X(NMAX)
*     .. Local Scalars ..
      real             D, P, Z
      INTEGER          I, IFAIL, N, NTYPE
*     .. External Functions ..
      real             CDF1, CDF2
      EXTERNAL         CDF1, CDF2
*     .. External Subroutines ..
      EXTERNAL         G08CCF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G08CCF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) N
      WRITE (NOUT,*)
      IF (N.LE.NMAX) THEN
         READ (NIN,*) (X(I),I=1,N)
         READ (NIN,*) NTYPE
         IFAIL = 0
*
         CALL G08CCF(N,X,CDF1,NTYPE,D,Z,P,SX,IFAIL)
*
         WRITE (NOUT,*) 'Test against uniform distribution on (0,2)'
         WRITE (NOUT,*)
         WRITE (NOUT,99999) 'Test statistic D = ', D
         WRITE (NOUT,99999) 'Z statistic      = ', Z
         WRITE (NOUT,99999) 'Tail probability = ', P
```

```
*
         CALL G08CCF(N,X,CDF2,NTYPE,D,Z,P,SX,IFAIL)
*
         WRITE (NOUT,*)
         WRITE (NOUT,*)
     +     'Test against normal distribution with mean = 0.75'
         WRITE (NOUT,*) 'and standard deviation = 0.5.'
         WRITE (NOUT,*)
         WRITE (NOUT,99999) 'Test statistic D = ', D
         WRITE (NOUT,99999) 'Z statistic      = ', Z
         WRITE (NOUT,99999) 'Tail probability = ', P
      ELSE
         WRITE (NOUT,99998) 'N is out of range: N = ', N
      END IF
      STOP
*
99999 FORMAT (1X,A,F8.4)
99998 FORMAT (1X,A,I7)
      END
*
      real FUNCTION CDF1(X)
*     .. Parameters ..
      real              A, B
      PARAMETER         (A=0.0e0,B=2.0e0)
*     .. Scalar Arguments ..
      real              X
*     .. Executable Statements ..
      IF (X.LT.A) THEN
         CDF1 = 0.0e0
      ELSE IF (X.GT.B) THEN
         CDF1 = 1.0e0
      ELSE
         CDF1 = (X-A)/(B-A)
      END IF
      RETURN
      END
*
      real FUNCTION CDF2(X)
*     .. Parameters ..
      real              XMEAN, STD
      PARAMETER         (XMEAN=0.75e0,STD=0.5e0)
*     .. Scalar Arguments ..
      real              X
*     .. Local Scalars ..
      real              Z
      INTEGER           IFAIL
*     .. External Functions ..
      real              S15ABF
      EXTERNAL          S15ABF
*     .. Executable Statements ..
      Z = (X-XMEAN)/STD
      CDF2 = S15ABF(Z,IFAIL)
      RETURN
      END
```

9.2 Program Data

```
G08CCF Example Program Data
 30
 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
 1
```

9.3 Program Results

```
G08CCF Example Program Results

Test against uniform distribution on (0,2)

Test statistic D =   0.2800
Z statistic      =   1.5336
Tail probability =   0.0143

Test against normal distribution with mean = 0.75
and standard deviation = 0.5.

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