NAG Library Routine Document
G08CCF

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 (KIND=nag_wp) X(N), CDF, D, Z, P, SX(N)
EXTERNAL             CDF
```

3 Description

The data consists of a single sample of n observations, denoted by $x_1, x_2, \ldots, x_n$. Let $S_n\left(x_{(i)}\right)$ and $F_0\left(x_{(i)}\right)$ 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 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 $D_n = \max\left\{D_n^+, D_n^-\right\}$.

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

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 $D_n^- = \max\left\{F_0\left(x_{(i)}\right) - S_n\left(x_{(i-1)}\right), 0\right\}$. 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 Arguments

1: N – INTEGER *Input*

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

Constraint: N $\geq$ 1.

2: X(N) – REAL (KIND=nag_wp) array *Input*

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

3: CDF – REAL (KIND=nag_wp) 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.

The specification of CDF is:

```
FUNCTION CDF (X)
REAL (KIND=nag_wp) CDF
REAL (KIND=nag_wp) X
```

1: X – REAL (KIND=nag_wp) *Input*

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

CDF must either be a module subprogram USEd by, or declared as EXTERNAL in, the (sub) program from which G08CCF is called. Arguments denoted as *Input* must **not** be changed by this procedure.

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

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 (KIND=nag_wp) *Output*

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

6: Z – REAL (KIND=nag_wp) *Output*

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

7: P – REAL (KIND=nag_wp) *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 (KIND=nag_wp) 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. If you are unfamiliar with this argument you should refer to Section 3.4 in How to Use the NAG Library and its Documentation for details.

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, if you are not familiar with this argument, the recommended value is 0. **When the value -1 or 1 is used it is essential to test the value of IFAIL on exit.**

On exit: IFAIL $= 0$ unless the routine detects an error or a warning has been flagged (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 or warnings detected by the routine:

IFAIL $= 1$

On entry, $\text{N} < 1$.

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.

See Section 3.9 in How to Use the NAG Library and its Documentation for further information.

IFAIL $= -399$

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

See Section 3.8 in How to Use the NAG Library and its Documentation for further information.

IFAIL $= -999$

Dynamic memory allocation failed.

See Section 3.7 in How to Use the NAG Library and its Documentation for further information.

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 Parallelism and Performance

G08CCF is threaded by NAG for parallel execution in multithreaded implementations of the NAG Library.

Please consult the X06 Chapter Introduction for information on how to control and interrogate the OpenMP environment used within this routine. Please also consult the Users' Note for your implementation for any additional implementation-specific information.

9 Further Comments

The time taken by G08CCF 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.

10 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.

10.1 Program Text

```
!   G08CCF Example Program Text
!   Mark 26 Release. NAG Copyright 2016.

    Module g08ccfe_mod

!     G08CCF Example Program Module:
!            Parameters and User-defined Routines

!     .. Use Statements ..
      Use nag_library, Only: nag_wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Accessibility Statements ..
      Private
      Public                             :: user_cdf
!     .. Parameters ..
      Real (Kind=nag_wp), Parameter      :: std = 0.5_nag_wp
      Real (Kind=nag_wp), Parameter      :: xmean = 0.75_nag_wp
      Integer, Parameter, Public         :: nin = 5, nout = 6
    Contains
      Function user_cdf(x)
!       Cumulative distribution function for the user supplied distribution.
!       In this example, the distribution is the normal distribution, with
!       mean = 0.75 and standard deviation = 0.5

!       .. Use Statements ..
        Use nag_library, Only: s15abf
!       .. Function Return Value ..
        Real (Kind=nag_wp)               :: user_cdf
!       .. Scalar Arguments ..
        Real (Kind=nag_wp), Intent (In) :: x
!       .. Local Scalars ..
        Real (Kind=nag_wp)               :: z
        Integer                          :: ifail
!       .. Executable Statements ..
        z = (x-xmean)/std
        ifail = -1
        user_cdf = s15abf(z,ifail)
        Return
      End Function user_cdf
    End Module g08ccfe_mod
    Program g08ccfe

!     G08CCF Example Main Program

!     .. Use Statements ..
      Use nag_library, Only: g08ccf, nag_wp
      Use g08ccfe_mod, Only: nin, nout, user_cdf
!     .. Implicit None Statement ..
      Implicit None
!     .. Local Scalars ..
      Real (Kind=nag_wp)                 :: d, p, z
      Integer                            :: ifail, n, ntype
!     .. Local Arrays ..
      Real (Kind=nag_wp), Allocatable    :: sx(:), x(:)
!     .. Executable Statements ..
      Write (nout,*) 'G08CCF Example Program Results'
      Write (nout,*)

!     Skip heading in data file
```

```
      Read (nin,*)

!     Read in problem type and required statistic
      Read (nin,*) n, ntype

      Allocate (x(n),sx(n))

!     Read in data
      Read (nin,*) x(1:n)

!     Perform K-S test for user specified distribution
      ifail = 0
      Call g08ccf(n,x,user_cdf,ntype,d,z,p,sx,ifail)

!     Display results
      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

99999 Format (1X,A,F8.4)
    End Program g08ccfe
```

10.2 Program Data

```
G08CCF Example Program Data
 30  1                                             :: N,NTYPE
 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 :: End of X
```

10.3 Program Results

```
 G08CCF Example Program Results

 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
```