NAG Library Routine Document

G08CBF

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

G08CBF performs the one sample Kolmogorov–Smirnov test, using one of the distributions provided.

2 Specification

```
SUBROUTINE G08CBF (N, X, DIST, PAR, ESTIMA, NTYPE, D, Z, P, SX, IFAIL)
INTEGER              N, NTYPE, IFAIL
REAL (KIND=nag_wp) X(N), PAR(2), D, Z, P, SX(N)
CHARACTER(*)         DIST
CHARACTER(1)         ESTIMA
```

3 Description

The data consist of a single sample of n observations denoted by $x_1, x_2, \ldots, x_n$. Let $S_n\big(x_{(i)}\big)$ and $F_0\big(x_{(i)}\big)$ 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 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 null 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 $D_n^+ = \max\big\{S_n\big(x_{(i)}\big) - F_0\big(x_{(i)}\big), 0\big\}$ for both discrete and continuous null distributions.

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 if the null distribution is discrete then $D_n^- = \max\big\{F_0\big(x_{(i)}\big) - S_n\big(x_{(i)}\big), 0\big\}$ and if the null distribution is continuous then $D_n^- = \max\big\{F_0\big(x_{(i)}\big) - S_n\big(x_{(i-1)}\big), 0\big\}$.

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 correction for continuity 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 Parameters

1: N – INTEGER *Input*

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

Constraint: N ≥ 3.

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

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

Constraint: the sample observations supplied must be consistent, in the usual manner, with the null distribution chosen, as specified by the parameters DIST and PAR. For further details see Section 9.

3: DIST – CHARACTER(*) *Input*

On entry: the theoretical (null) distribution from which it is suspected the data may arise.

DIST = 'U'
The uniform distribution over (a, b).

DIST = 'N'
The Normal distribution with mean μ and variance σ^2.

DIST = 'G'
The gamma distribution with shape parameterα and scale parameter β, where the mean $= \alpha\beta$.

DIST = 'BE'
The beta distribution with shape parameters α and β, where the mean $= \alpha/(\alpha + \beta)$.

DIST = 'BI'
The binomial distribution with the number of trials, m, and the probability of a success, p.

DIST = 'E'
The exponential distribution with parameter λ, where the mean $= 1/\lambda$.

DIST = 'P'
The Poisson distribution with parameter μ, where the mean $= \mu$.

DIST = 'NB'
The negative binomial distribution with the number of trials, m, and the probability of success, p.

DIST = 'GP'
The generalized Pareto distribution with shape parameter ξ and scale β.

Any number of characters may be supplied as the actual parameter, however only the characters, maximum 2, required to uniquely identify the distribution are referenced.

Constraint: DIST = 'U', 'N', 'G', 'BE', 'BI', 'E', 'P', 'NB' or 'GP'.

4: PAR(2) – REAL (KIND=nag_wp) array *Input/Output*

On entry: if ESTIMA = 'S', PAR must contain the known values of the parameter(s) of the null distribution as follows.

If a uniform distribution is used, then PAR(1) and PAR(2) must contain the boundaries a and b respectively.

If a Normal distribution is used, then PAR(1) and PAR(2) must contain the mean, μ, and the variance, σ^2, respectively.

If a gamma distribution is used, then PAR(1) and PAR(2) must contain the parameters α and β respectively.

If a beta distribution is used, then PAR(1) and PAR(2) must contain the parameters α and β respectively.

If a binomial distribution is used, then PAR(1) and PAR(2) must contain the parameters m and p respectively.

If an exponential distribution is used, then PAR(1) must contain the parameter λ.

If a Poisson distribution is used, then PAR(1) must contain the parameter μ.

If a negative binomial distribution is used, PAR(1) and PAR(2) must contain the parameters m and p respectively.

If a generalized Pareto distribution is used, PAR(1) and PAR(2) must contain the parameters ξ and β respectively.

If ESTIMA = 'E', PAR need not be set except when the null distribution requested is either the binomial or the negative binomial distribution in which case PAR(1) must contain the parameter m.

On exit: if ESTIMA = 'S', PAR is unchanged; if ESTIMA = 'E', and DIST = 'BI' or DIST = 'NB' then PAR(2) is estimated from the data; otherwise PAR(1) and PAR(2) are estimated from the data.

Constraints:

if DIST = 'U', PAR(1) $<$ PAR(2);
if DIST = 'N', PAR(2) $>$ 0.0;
if DIST = 'G', PAR(1) $>$ 0.0 and PAR(2) $>$ 0.0;
if DIST = 'BE', PAR(1) $>$ 0.0 and PAR(2) $>$ 0.0 and PAR(1) $\leq 10^6$ and PAR(2) $\leq 10^6$;

if DIST = 'BI', $\mathrm{PAR}(1) \geq 1.0$ and $0.0 < \mathrm{PAR}(2) < 1.0$ and $\mathrm{PAR}(1) \times \mathrm{PAR}(2) \times (1.0 - \mathrm{PAR}(2)) \leq 10^6$ and $\mathrm{PAR}(1) < 1/\mathrm{eps}$, where eps = ***machine precision***, see X02AJF;
if DIST = 'E', $\mathrm{PAR}(1) > 0.0$;
if DIST = 'P', $\mathrm{PAR}(1) > 0.0$ and $\mathrm{PAR}(1) \leq 10^6$;
if DIST = 'NB', $\mathrm{PAR}(1) \geq 1.0$ and $0.0 < \mathrm{PAR}(2) < 1.0$ and $\mathrm{PAR}(1) \times (1.0 - \mathrm{PAR}(2))/(\mathrm{PAR}(2) \times \mathrm{PAR}(2)) \leq 10^6$ and $\mathrm{PAR}(1) < 1/\mathrm{eps}$, where eps = ***machine precision***, see X02AJF;
if DIST = 'GP', $\mathrm{PAR}(2) > 0$.

5: ESTIMA – CHARACTER(1) *Input*

On entry: ESTIMA must specify whether values of the parameters of the null distribution are known or are to be estimated from the data.

ESTIMA = 'S'
Values of the parameters will be supplied in the array PAR described above.

ESTIMA = 'E'
Parameters are to be estimated from the data except when the null distribution requested is the binomial distribution or the negative binomial distribution in which case the first parameter, m, must be supplied in PAR(1) and only the second parameter, p, is estimated from the data.

Constraint: ESTIMA = 'S' or 'E'.

6: NTYPE – INTEGER *Input*

On entry: the test 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.

7: 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).

8: Z – REAL (KIND=nag_wp) *Output*

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

9: P – REAL (KIND=nag_wp) *Output*

On exit: the probability, p, associated with the observed value of D where D may be D_n, D_n^+ or D_n^- depending on the value of NTYPE (see Section 3).

10: SX(N) – REAL (KIND=nag_wp) array *Output*

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

11: IFAIL – INTEGER *Input/Output*

On entry: IFAIL must be set to 0, -1 or 1. If you are unfamiliar with this parameter you should refer to Section 3.3 in the Essential Introduction 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 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.**

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, N $= \langle$*value*$\rangle$.
Constraint: N ≥ 3.

IFAIL $= 2$

On entry, DIST $= \langle$*value*$\rangle$ was an illegal value.

IFAIL $= 3$

On entry, NTYPE $= \langle$*value*$\rangle$.
Constraint: NTYPE $= 1$, 2 or 3.

IFAIL $= 4$

On entry, ESTIMA $= \langle$*value*$\rangle$ was an illegal value.

IFAIL $= 5$

On entry, DIST = 'BI' and $m =$ PAR(1) $= \langle$*value*$\rangle$.
Note that m must always be supplied.
Constraint: for the binomial distribution, $1 \leq$ PAR(1) $< 1/$eps, where eps $=$ ***machine precision***, see X02AJF.

On entry, DIST = 'NB' and $m =$ PAR(1) $= \langle$*value*$\rangle$.
Note that m must always be supplied.
Constraint: for the negative binomial distribution, $1 \leq$ PAR(1) $< 1/$eps, where eps $=$ ***machine precision***, see X02AJF.

On entry, ESTIMA = 'S' and PAR(1) $= \langle$*value*$\rangle$; PAR(2) $= \langle$*value*$\rangle$.
Constraint: for the beta distribution, $0 <$ PAR(1) and PAR(2) ≤ 1000000.

On entry, ESTIMA = 'S' and PAR(1) $= \langle$*value*$\rangle$; PAR(2) $= \langle$*value*$\rangle$.
Constraint: for the gamma distribution, PAR(1) and PAR(2) > 0.

On entry, ESTIMA = 'S' and PAR(1) $= \langle$*value*$\rangle$; PAR(2) $= \langle$*value*$\rangle$.
Constraint: for the generalized Pareto distribution with PAR(1) < 0, $0 \leq$ X(i) $\leq -$PAR(2)/PAR(1), for $i = 1, 2, \ldots,$ N.

On entry, ESTIMA = 'S' and PAR(1) $= \langle$*value*$\rangle$; PAR(2) $= \langle$*value*$\rangle$.
Constraint: for the uniform distribution, PAR(1) $<$ PAR(2).

On entry, ESTIMA = 'S' and PAR(1) $= \langle$*value*$\rangle$.
Constraint: for the exponential distribution, PAR(1) > 0.

On entry, ESTIMA = 'S' and PAR(1) $= \langle$*value*$\rangle$.
Constraint: for the Poisson distribution, $0 <$ PAR(1) < 1000000.

On entry, ESTIMA = 'S' and PAR(2) = ⟨*value*⟩.
Constraint: for the binomial distribution, $0 <$ PAR(2) < 1.

On entry, ESTIMA = 'S' and PAR(2) = ⟨*value*⟩.
Constraint: for the generalized Pareto distribution, PAR(2) > 0.

On entry, ESTIMA = 'S' and PAR(2) = ⟨*value*⟩.
Constraint: for the negative binomial distribution, $0 <$ PAR(2) < 1.

On entry, ESTIMA = 'S' and PAR(2) = ⟨*value*⟩.
Constraint: for the Normal distribution, PAR(2) > 0.

IFAIL = 6

On entry, DIST = 'U' and at least one observation is illegal.
Constraint: PAR(1) $\leq$ X(i) $\leq$ PAR(2), for $i = 1, 2, \ldots,$ N.

On entry, DIST = 'G', 'E', 'P', 'NB' or 'GP' and at least one observation is negative.
Constraint: X(i) ≥ 0, for $i = 1, 2, \ldots,$ N.

On entry, DIST = 'BE' and at least one observation is illegal.
Constraint: $0 \leq$ X(i) ≤ 1, for $i = 1, 2, \ldots,$ N.

On entry, DIST = 'BI' and all observations are zero or m.
Constraint: at least one $0.0 <$ X(i) $<$ PAR(1), for $i = 1, 2, \ldots,$ N.

On entry, DIST = 'BI' and at least one observation is illegal.
Constraint: $0 \leq$ X(i) $\leq$ PAR(1), for $i = 1, 2, \ldots,$ N.

On entry, DIST = 'E' or 'P' and all observations are zero.
Constraint: at least one X(i) > 0, for $i = 1, 2, \ldots,$ N.

On entry, DIST = 'GP' and ESTIMA = 'E'.
The parameter estimates are invalid; the data may not be from the generalized Pareto distribution.

IFAIL = 7

On entry, DIST = 'U', 'N', 'G', 'BE' or 'GP', ESTIMA = 'E' and the whole sample is constant. Thus the variance is zero.

IFAIL = 8

On entry, DIST = 'BI', PAR(1) = ⟨*value*⟩, PAR(2) = ⟨*value*⟩.
The variance PAR(1) × PAR(2) × (1 − PAR(2)) exceeds 1000000.

On entry, DIST = 'NB', PAR(1) = ⟨*value*⟩, PAR(2) = ⟨*value*⟩.
The variance PAR(1) × (1 − PAR(2))/(PAR(2) × PAR(2)) exceeds 1000000.

IFAIL = 9

On entry, DIST = 'G' and in the computation of the incomplete gamma function by S14BAF the convergence of the Taylor series or Legendre continued fraction fails within 600 iterations.

IFAIL = −99

An unexpected error has been triggered by this routine. Please contact NAG.

See Section 3.8 in the Essential Introduction for further information.

IFAIL = −399

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

See Section 3.7 in the Essential Introduction for further information.

IFAIL $= -999$

Dynamic memory allocation failed.

See Section 3.6 in the Essential Introduction for further information.

7 Accuracy

The approximation for p, given when $n > 100$, has a relative error of at most 2.5% for most cases. The two-sided probability is approximated by doubling the one-sided probability. This is only good for small p, i.e., $p < 0.10$ but very poor for large p. The error is always on the conservative side, that is the tail probability, p, is over estimated.

8 Parallelism and Performance

G08CBF 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 G08CBF increases with n until $n > 100$ at which point it drops and then increases slowly with n. The time may also depend on the choice of null distribution and on whether or not the parameters are to be estimated.

The data supplied in the parameter X must be consistent with the chosen null distribution as follows:

when DIST = 'U', then PAR$(1) \leq x_i \leq$ PAR(2), for $i = 1, 2, \ldots, n$;

when DIST = 'N', then there are no constraints on the x_i's;

when DIST = 'G', then $x_i \geq 0.0$, for $i = 1, 2, \ldots, n$;

when DIST = 'BE', then $0.0 \leq x_i \leq 1.0$, for $i = 1, 2, \ldots, n$;

when DIST = 'BI', then $0.0 \leq x_i \leq$ PAR(1), for $i = 1, 2, \ldots, n$;

when DIST = 'E', then $x_i \geq 0.0$, for $i = 1, 2, \ldots, n$;

when DIST = 'P', then $x_i \geq 0.0$, for $i = 1, 2, \ldots, n$;

when DIST = 'NB', then $x_i \geq 0.0$, for $i = 1, 2, \ldots, n$;

when DIST = 'GP' and PAR$(1) \geq 0.0$, then $x_i \geq 0.0$, for $i = 1, 2, \ldots, n$;

when DIST = 'GP' and PAR$(1) < 0.0$, then $0.0 \leq x_i \leq -$PAR$(2)/$PAR(1), for $i = 1, 2, \ldots, n$.

10 Example

The following example program reads in a set of data consisting of 30 observations. The Kolmogorov–Smirnov test is then applied twice, firstly to test whether the sample is taken from a uniform distribution, $U(0, 2)$, and secondly to test whether the sample is taken from a Normal distribution where the mean and variance are estimated from the data. In both cases we are testing against H_1; that is, we are doing a two tailed test. The values of D, Z and P are printed for each case.

10.1 Program Text

```
    Program g08cbfe

!     G08CBF Example Program Text

!     Mark 25 Release. NAG Copyright 2014.

!     .. Use Statements ..
      Use nag_library, Only: g08cbf, nag_wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter               :: nin = 5, nout = 6
!     .. Local Scalars ..
      Real (Kind=nag_wp)               :: d, p, z
      Integer                          :: ifail, n, npar, ntype
      Character (2)                    :: dist
      Character (1)                    :: estima
!     .. Local Arrays ..
      Real (Kind=nag_wp)               :: par(2)
      Real (Kind=nag_wp), Allocatable  :: sx(:), x(:)
!     .. Executable Statements ..
      Write (nout,*) 'G08CBF Example Program Results'
      Write (nout,*)

!     Skip heading in data file
      Read (nin,*)

!     Read in problem size and the statistic to calculate
      Read (nin,*) n, ntype

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

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

!     Read in information on the distribution to test against
      Read (nin,*) dist, estima

      Select Case (dist)
      Case ('P','p','E','e')
        npar = 1
      Case Default
        npar = 2
      End Select

!     Read in the distribution parameters if required
!     otherwise they are estimated from the data by G08CBF
!     and PAR need not be set
      If (estima=='S' .Or. estima=='s') Then
        Read (nin,*) par(1:npar)
      Else If (dist=='B' .Or. dist=='b' .Or. dist=='NB' .Or. dist=='nb') Then
!       Read in M for the binomial distribution
        Read (nin,*) par(1)
      End If

!     Perform K-S test
      ifail = 0
      Call g08cbf(n,x,dist,par,estima,ntype,d,z,p,sx,ifail)

!     Display results
      Write (nout,*) 'K-S Test'
      Write (nout,*) 'Distribution: ', dist
      Write (nout,99999) 'Parameters  : ', par(1:npar)
      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,2F8.4)
    End Program g08cbfe
```

10.2 Program Data

```
G08CBF 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
'N' 'E'                                            :: DIST,ESTIMA
```

10.3 Program Results

```
 G08CBF Example Program Results

 K-S Test
 Distribution: N
 Parameters  :   0.6967  0.2564

 Test statistic D =   0.1108
 Z statistic      =   0.6068
 Tail probability =   0.8925
```