

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 standard distributions provided.

2 Specification

```
SUBROUTINE G08CBF(N, X, DIST, PAR, ESTIMA, NTYPE, D, Z, P, SX, IFAIL)
INTEGER          N, NTYPE, IFAIL
double precision 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 $x_1, x_2, \dots, x_n$. 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 i th 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 $D_n^+ = \max\{S_n(x_{(i)}) - F_0(x_{(i)}), 0\}$ 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 $D_n^- = \max\{F_0(x_{(i)}) - S_n(x_{(i)}), 0\}$ and if the null distribution is continuous then $D_n^- = \max\{F_0(x_{(i)}) - S_n(x_{(i-1)}), 0\}$.

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: $N \geq 3$.
- 2: X(N) – **double precision** array *Input*
On entry: the sample observations $x_1, x_2, \dots, x_n$.
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 8
- 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 $(a, b) - U(a, b)$.
DIST = 'N'
The Normal distribution with mean μ and variance $\sigma^2 - N(\mu, \sigma^2)$.
DIST = 'G'
The gamma distribution with shape parameter α 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 = μ .

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' or 'P'.

4: PAR(2) – **double precision** 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 ESTIMA = 'E', PAR need not be set except when the null distribution requested is the binomial distribution in which case PAR(1) must contain the parameter m .

On exit: if ESTIMA = 'S', PAR is unchanged. If ESTIMA = 'E', then PAR(1) and PAR(2) are set to values as 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) ≤ 10⁶ and PAR(2) ≤ 10⁶;

if DIST = 'BI', PAR(1) ≥ 1.0 and 0.0 < PAR(2) < 1.0 and

PAR(1) × PAR(2) × (1.0 – PAR(2)) ≤ 10⁶ and PAR(1) < 1/eps, where

eps = **machine precision**, see X02AJF;

if DIST = 'E', PAR(1) > 0.0;

if DIST = 'P', PAR(1) > 0.0 and PAR(1) ≤ 10⁶.

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 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 – *double precision*

Output

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

8: Z – *double precision*

Output

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

9: P – *double precision*

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) – *double precision* array

Output

On exit: the sample observations, $x_1, x_2, \dots, x_n$, 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.

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, 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.**

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, $N < 3$.

IFAIL = 2

On entry, an invalid code for DIST has been specified.

IFAIL = 3

On entry, NTYPE $\neq$ 1, 2 or 3.

IFAIL = 4

On entry, ESTIMA $\neq$ 'S' or 'E'.

IFAIL = 5

On entry, the parameters supplied for the specified null distribution are out of range (see Section 5). Apart from a check on the first parameter for the binomial distribution (DIST = 'BI') this error will only occur if ESTIMA = 'S'.

IFAIL = 6

The data supplied in X could not arise from the chosen null distribution, as specified by the parameters DIST and PAR. For further details see Section 8.

IFAIL = 7

The whole sample is constant, i.e., the variance is zero. This error may only occur if (DIST = 'U', 'N', 'G' or 'BE') and ESTIMA = 'E'.

IFAIL = 8

The variance of the binomial distribution (DIST = 'BI') is too large. That is, $mp(1 - p) > 1000000$.

IFAIL = 9

When DIST = 'G', in the computation of the incomplete gamma function by S14BAF the convergence of the Taylor series or Legendre continued fraction fails within 600 iterations. This is an unlikely error exit.

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 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 $\text{PAR}(1) \leq x_i \leq \text{PAR}(2)$, for $i = 1, 2, \dots, n$;

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

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

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

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

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

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

9 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.

9.1 Program Text

```
*      G08CBF Example Program Text
*      Mark 14 Release. NAG Copyright 1989.
*      .. Parameters ..
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
      INTEGER          NMAX, MAXP
      PARAMETER        (NMAX=30,MAXP=2)
*      .. Local Arrays ..
      DOUBLE PRECISION PAR(MAXP), SX(NMAX), X(NMAX)
*      .. Local Scalars ..
      DOUBLE PRECISION D, P, Z
      INTEGER          I, IFAIL, N, NP, NTYPE
*      .. External Subroutines ..
      EXTERNAL         G08CBF
*      .. Executable Statements ..
      WRITE (NOUT,*) 'G08CBF 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,*) NP, (PAR(I),I=1,NP), NTYPE
        IFAIL = 0
*
        CALL G08CBF(N,X,'Uniform',PAR,'Supplied',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
        WRITE (NOUT,*)
*
        READ (NIN,*) NP, (PAR(I),I=1,NP), NTYPE
        IFAIL = 0
*
        CALL G08CBF(N,X,'Normal',PAR,'Estimate',NTYPE,D,Z,P,SX,IFAIL)
*
        WRITE (NOUT,*)
        + 'Test against Normal distribution with parameters estimated from t
        + he data'
        WRITE (NOUT,*)
        WRITE (NOUT,99998) 'Mean = ', PAR(1), ' and variance = ',
        + PAR(2)
        WRITE (NOUT,99999) 'Test statistic D = ', D
        WRITE (NOUT,99999) 'Z statistic      = ', Z
        WRITE (NOUT,99999) 'Tail probability = ', P
      ELSE
        WRITE (NOUT,99997) 'N is out of range: N = ', N
      END IF
*
99999 FORMAT (1X,A,F8.4)
99998 FORMAT (1X,A,F6.4,A,F6.4)
99997 FORMAT (1X,A,I7)
      END
```

9.2 Program Data

G08CBF 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
2 0.0 2.0 1
2 0.0 1.0 1
```

9.3 Program Results

G08CBF 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 parameters estimated from the data

```
Mean = 0.6967 and variance = 0.2564
Test statistic D = 0.1108
Z statistic      = 0.6068
Tail probability = 0.8925
```
