G01EZF – 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

G01EZF returns the probability associated with the upper tail of the Kolmogorov–Smirnov two sample distribution, via the routine name.

2 Specification

```
real FUNCTION G01EZF(N1, N2, D, IFAIL)
INTEGER              N1, N2, IFAIL
real                 D
```

3 Description

Let $F_{n_1}(x)$ and $G_{n_2}(x)$ denote the empirical cumulative distribution functions for the two samples where n_1 and n_2 are the sizes of the first and second samples respectively.

The function G01EZF computes the upper tail probability for the Kolmogorov–Smirnov two sample two-sided test statistic D_{n_1,n_2} where

$$D_{n_1,n_2} = \sup_x |F_{n_1}(x) - G_{n_2}(x)|.$$

The probability is computed exactly if $n_1, n_2 \leq 10000$ and $\max(n_1, n_2) \leq 2500$ using a method given by Kim and Jenrich [6]. For the case where $\min(n_1, n_2) \leq 10$ percent of the $\max(n_1, n_2)$ and $\min(n_1, n_2) \leq 80$ the Smirnov approximation is used. For all other cases the Kolmogorov approximation is used. These two approximations are discussed in [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] Smirnov N (1948) Table for estimating the goodness of fit of empirical distributions *Ann. Math. Statist.* **19** 279–281

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

[6] 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** American Mathematical Society 80–129

5 Parameters

1: N1 — INTEGER *Input*

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

Constraint: N1 ≥ 1.

2: N2 — INTEGER *Input*

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

Constraint: N2 ≥ 1.

3: D — ***real*** *Input*

On entry: the test statistic D_{n_1,n_2}, for the two sample Kolmogorov–Smirnov goodness of fit test, that is the maximum difference between the empirical cumulative distribution functions (CDFs) of the two samples.

Constraint: $0.0 \leq \mathrm{D} \leq 1.0$.

4: 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, N1 < 1,
or N2 < 1.

IFAIL = 2

On entry, D < 0.0,
or D > 1.0.

IFAIL = 3

The approximation solution did not converge in 500 iterations. A tail probability of 1.0 is returned by G01EZF.

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 upper tail probability for the one-sided statistics, $D^{+}_{n_1,n_2}$ or $D^{-}_{n_1,n_2}$, can be approximated by halving the two-sided upper tail probability returned by G01EZF, that is $p/2$. This approximation to the upper tail probability for either $D^{+}_{n_1,n_2}$ or $D^{-}_{n_1,n_2}$ is good for small probabilities, (e.g., $p \leq 0.10$) but becomes poor for larger probabilities.

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 reads in 10 different sample sizes and values for the test statistic D_{n_1,n_2}. The upper tail probability is computed and printed for each case.

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.

```
*     G01EZF Example Program Text
*     Mark 14 Release.  NAG Copyright 1989.
*     .. Parameters ..
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
*     .. Local Scalars ..
      real             D, PROB
      INTEGER          IFAIL, N1, N2
*     .. External Functions ..
      real             G01EZF
      EXTERNAL         G01EZF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G01EZF Example Program Results'
      WRITE (NOUT,*)
      WRITE (NOUT,*) '    D      N1    N2    Two-sided probability'
      WRITE (NOUT,*)
*     Skip heading in data file
      READ (NIN,*)
   20 READ (NIN,*,END=40) N1, N2, D
      IFAIL = -1
*
      PROB = G01EZF(N1,N2,D,IFAIL)
*
      IF (IFAIL.EQ.0) THEN
         WRITE (NOUT,99999) D, N1, N2, PROB
      ELSE
         WRITE (NOUT,99998) D, N1, N2, PROB, ' *IFAIL = ', IFAIL
      END IF
      GO TO 20
   40 STOP
*
99999 FORMAT (1X,F7.4,2X,I4,2X,I4,10X,F7.4)
99998 FORMAT (1X,F7.4,2X,I4,2X,I4,10X,F7.4,A,I2)
      END
```

9.2 Program Data

```
G01EZF Example Program Data.
  5   10   0.5
 10   10   0.5
 20   10   0.5
 20   15   0.4833
400  200   0.1412
200   20   0.2861
1000  20   0.2113
200   50   0.1796
 15  200   0.18
100  100   0.18
```

9.3 Program Results

```
G01EZF Example Program Results

    D       N1    N2    Two-sided probability

 0.5000      5    10          0.3506
 0.5000     10    10          0.1678
 0.5000     20    10          0.0623
 0.4833     20    15          0.0261
 0.1412    400   200          0.0083
 0.2861    200    20          0.0789
 0.2113   1000    20          0.2941
 0.1796    200    50          0.1392
 0.1800     15   200          0.6926
 0.1800    100   100          0.0782
```