NAG Library Routine Document

G01EZF

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

```
FUNCTION G01EZF (N1, N2, D, IFAIL)
REAL (KIND=nag_wp) G01EZF
INTEGER            N1, N2, IFAIL
REAL (KIND=nag_wp) 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

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

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 (1973). For the case where $\min(n_1, n_2) \leq 10\%$ 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 Kim and Jenrich (1973).

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

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** 80–129 American Mathematical Society

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

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

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 (KIND=nag_wp) *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: $0.0 \leq$ D ≤ 1.0.

4: 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, 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

```
    Program g01ezfe

!     G01EZF Example Program Text

!     Mark 24 Release. NAG Copyright 2012.

!     .. Use Statements ..
      Use nag_library, Only: g01ezf, nag_wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter               :: nin = 5, nout = 6
!     .. Local Scalars ..
      Real (Kind=nag_wp)               :: d, prob
      Integer                          :: ifail, n1, n2
!     .. Executable Statements ..
      Write (nout,*) 'G01EZF Example Program Results'
      Write (nout,*)

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

!     Display titles
      Write (nout,*) '    D      N1    N2    Two-sided probability'
      Write (nout,*)

d_lp: Do
        Read (nin,*,Iostat=ifail) n1, n2, d
        If (ifail/=0) Then
          Exit d_lp
        End If

!       Calculate probaility
        ifail = -1
        prob = g01ezf(n1,n2,d,ifail)
        If (ifail/=0) Then
          If (ifail/=3) Then
            Exit d_lp
          End If
        End If

!       Display results
        Write (nout,99999) d, n1, n2, prob
      End Do d_lp

99999 Format (1X,F7.4,2X,I4,2X,I4,10X,F7.4)
    End Program g01ezfe
```

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
```