NAG Toolbox for MATLAB

g01ez

1 Purpose

g01ez returns the probability associated with the upper tail of the Kolmogorov–Smirnov two sample distribution, via the function name.

2 Syntax

```
[result, ifail] = g01ez(n1, n2, 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 g01ez 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

5.1 Compulsory Input Parameters

1: **n1 – int32 scalar**

The number of observations in the first sample, n_1.

Constraint: **n1** ≥ 1.

2: **n2 – int32 scalar**

The number of observations in the second sample, n_2.

Constraint: **n2** ≥ 1.

3: **d – double scalar**

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: $0.0 \leq \mathbf{d} \leq 1.0$.

5.2 Optional Input Parameters

None.

5.3 Input Parameters Omitted from the MATLAB Interface

None.

5.4 Output Parameters

1: **result – double scalar**

The result of the function.

2: **ifail – int32 scalar**

0 unless the function detects an error (see Section 6).

6 Error Indicators and Warnings

Errors or warnings detected by the function:

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 g01ez.

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 g01ez, 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 function 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

```
n1 = int32(5);
n2 = int32(10);
d = 0.5;
[result, ifail] = g01ez(n1, n2, d)
```

```
result =
    0.3506
ifail =
           0
```