

# NAG C Library Function Document

## nag\_prob\_2\_sample\_ks (g01ezc)

### 1 Purpose

nag\_prob\_2\_sample\_ks (g01ezc) returns the probability associated with the upper tail of the Kolmogorov–Smirnov two sample distribution.

### 2 Specification

double nag\_prob\_2\_sample\_ks (Integer **n1**, Integer **n2**, double **d**, NagError \*fail)

### 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 nag\_prob\_2\_sample\_ks (g01ezc) 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 (1973). 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 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**  $\geq 1$ .
- 2: **n2** – Integer *Input*  
*On entry:* the number of observations in the second sample,  $n_2$ .  
*Constraint:* **n2**  $\geq 1$ .

- 3: **d** – double *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:*  $0.0 \leq \mathbf{d} \leq 1.0$ .
- 4: **fail** – NagError \* *Input/Output*  
The NAG error parameter (see the Essential Introduction).

## 6 Error Indicators and Warnings

### NE\_INT

On entry, either **n1** or **n2** is less than 1: **n1** =  $\langle value \rangle$ , **n2** =  $\langle value \rangle$ .

### NE\_CONVERGENCE

The Smirnov approximation used for large samples did not converge in 200 iterations. The probability is set to 1.0.

### NE\_REAL

On entry, **d** < 0.0 or **d** > 1.0: **d** =  $\langle value \rangle$ .

### NE\_BAD\_PARAM

On entry, parameter  $\langle value \rangle$  had an illegal value.

### NE\_INTERNAL\_ERROR

An internal error has occurred in this function. Check the function call and any array sizes. If the call is correct then please consult NAG for assistance.

## 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 nag\_prob\_2\_sample\_ks (g01ezc), 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 nag\_prob\_2\_sample\_ks (g01ezc) 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

```

/* nag_prob_2_sample_ks (g01ezc) Example Program.
 *
 * Copyright 2001 Numerical Algorithms Group.
 *
 * Mark 7, 2001.
 */

#include <stdio.h>
#include <nag.h>
#include <nag_stdlib.h>
#include <nagg01.h>

int main(void)
{
    /* Scalars */
    double d__, prob;
    Integer exit_status, n1, n2;
    NagError fail;

    INIT_FAIL(fail);
    exit_status = 0;
    Vprintf("%s\n\n", "g01ezc Example Program Results");
    Vprintf("%s\n\n", "      d      n1      n2      Two-sided probability");

    /* Skip heading in data file */
    Vscanf("%*[^\\n] ");

    while ( scanf("%ld%ld%lf%*[^\\n] ", &n1, &n2, &d__) != EOF)
    {
        prob = g01ezc(n1, n2, d__, &fail);
        if (fail.code != NE_NOERROR)
        {
            Vprintf("Error from g01ezc.\\n%s\\n", fail.message);
            exit_status = 1;
            goto END;
        }

        Vprintf("%7.4f%2s%4ld%2s%4ld%10s%7.4f\\n", d__, "", n1, "",
                n2, "", prob);
    }
    END:
    return exit_status;
}

```

## 9.2 Program Data

g01ezc 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

g01ezc 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

---