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


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


5 Parameters

1: n1 – Integer  
   On entry: the number of observations in the first sample, $n_1$.  
   Constraint: $n_1 \geq 1$.

2: n2 – Integer  
   On entry: the number of observations in the second sample, $n_2$.  
   Constraint: $n_2 \geq 1$.
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:* \( 0.0 \leq d \leq 1.0 \).

4:  
   
   fail – NagError * 
   *Input/Output*

   The NAG error parameter (see the Essential Introduction).

6  **Error Indicators and Warnings**

**NE_INT**

On entry, either \( n_1 \) or \( n_2 \) is less than 1: \( n_1 = (\text{value}) \), \( n_2 = (\text{value}) \).

**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 \( d > 1.0 \): \( d = (\text{value}) \).

**NE_BAD_PARAM**

On entry, parameter \( (\text{value}) \) 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", "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\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

<table>
<thead>
<tr>
<th>d</th>
<th>n1</th>
<th>n2</th>
<th>Two-sided probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5000</td>
<td>5</td>
<td>10</td>
<td>0.3506</td>
</tr>
<tr>
<td>Value</td>
<td>x</td>
<td>y</td>
<td>Result</td>
</tr>
<tr>
<td>-------</td>
<td>---</td>
<td>---</td>
<td>--------</td>
</tr>
<tr>
<td>0.5000</td>
<td>10</td>
<td>10</td>
<td>0.1678</td>
</tr>
<tr>
<td>0.5000</td>
<td>20</td>
<td>10</td>
<td>0.0623</td>
</tr>
<tr>
<td>0.4833</td>
<td>20</td>
<td>15</td>
<td>0.0261</td>
</tr>
<tr>
<td>0.1412</td>
<td>400</td>
<td>200</td>
<td>0.0083</td>
</tr>
<tr>
<td>0.2861</td>
<td>200</td>
<td>20</td>
<td>0.0789</td>
</tr>
<tr>
<td>0.2113</td>
<td>1000</td>
<td>20</td>
<td>0.2941</td>
</tr>
<tr>
<td>0.1796</td>
<td>200</td>
<td>50</td>
<td>0.1392</td>
</tr>
<tr>
<td>0.1800</td>
<td>15</td>
<td>200</td>
<td>0.6926</td>
</tr>
<tr>
<td>0.1800</td>
<td>100</td>
<td>100</td>
<td>0.0782</td>
</tr>
</tbody>
</table>