NAG Toolbox for MATLAB

g01dh

1 Purpose

g01dh computes the ranks, Normal scores, an approximation to the Normal scores or the exponential scores as requested by you.

2 Syntax

```
[r, ifail] = g01dh(scores, ties, x, 'n', n)
```

3 Description

g01dh computes one of the following scores for a sample of observations, $x_1, x_2, \ldots, x_n$.

1. Rank Scorse

   The ranks are assigned to the data in ascending order, that is the ith observation has score $s_i = k$ if it is the kth smallest observation in the sample.

2. Normal Scores

   The Normal scores are the expected values of the Normal order statistics from a sample of size n. If x_i is the kth smallest observation in the sample, then the score for that observation, s_i, is $E(Z_k)$ where Z_k is the kth order statistic in a sample of size n from a standard Normal distribution and E is the expectation operator.

3. Blom, Tukey and van der Waerden Scores

   These scores are approximations to the Normal scores. The scores are obtained by evaluating the inverse cumulative Normal distribution function, $\Phi^{-1}(\cdot)$, at the values of the ranks scaled into the interval $(0, 1)$ using different scaling transformations.

   The Blom scores use the scaling transformation $(r_i - 3/8)/(n + 1/4)$ for the rank r_i, for $i = 1, 2, \ldots, n$. Thus the Blom score corresponding to the observation x_i is

   $$s_i = \Phi^{-1}\left(\frac{r_i - 3/8}{n + 1/4}\right).$$

   The Tukey scores use the scaling transformation $(r_i - 1/3)/(n + 1/3)$; the Tukey score corresponding to the observation x_i is

   $$s_i = \Phi^{-1}\left(\frac{r_i - 1/3}{n + 1/3}\right).$$

   The van der Waerden scores use the scaling transformation $r_i/(n + 1)$; the van der Waerden score corresponding to the observation x_i is

   $$s_i = \Phi^{-1}\left(\frac{r_i}{n + 1}\right).$$

   The van der Waerden scores may be used to carry out the van der Waerden test for testing for differences between several population distributions, see Conover 1980.

4. Savage Scores

   The Savage scores are the expected values of the exponential order statistics from a sample of size n. They may be used in a test discussed by Savage 1956 and Lehmann 1975. If x_i is the kth smallest observation in the sample, then the score for that observation is

   $$s_i = E(Y_k) = \tfrac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{n-k+1},$$

where Y_k is the kth order statistic in a sample of size n from a standard exponential distribution and E is the expectation operator.

Ties may be handled in one of five ways. Let $x_{t(i)}$, for $i = 1, 2, \ldots, m$, denote m tied observations, that is $x_{t(1)} = x_{t(2)} = \cdots = x_{t(m)}$ with $t(1) < t(2) < \cdots < t(m)$. If the rank of $x_{t(1)}$ is k, then if ties are ignored the rank of $x_{t(j)}$ will be $k + j - 1$. Let the scores ignoring ties be $s^*_{t(1)}, s^*_{t(2)}, \ldots, s^*_{t(m)}$. Then the scores, $s_{t(i)}$, for $i = 1, 2, \ldots, m$, may be calculated as follows:

if averages are used, then $s_{t(i)} = \sum_{j=1}^{m} s^*_{t(j)}/m$;
if the lowest score is used, then $s_{t(i)} = s^*_{t(1)}$;
if the highest score is used, then $s_{t(i)} = s^*_{t(m)}$;
if ties are to be broken randomly, then $s_{t(i)} = s^*_{t(I)}$ where $I \in \{\text{random permutation of } 1, 2, \ldots, m\}$;
if ties are to be ignored, then $s_{t(i)} = s^*_{t(i)}$.

4 References

Blom G 1958 *Statistical Estimates and Transformed Beta-variables* Wiley

Conover W J 1980 *Practical Nonparametric Statistics* Wiley

Lehmann E L 1975 *Nonparametrics: Statistical Methods Based on Ranks* Holden–Day

Savage I R 1956 Contributions to the theory of rank order statistics – the two-sample case *Ann. Math. Statist.* **27** 590–615

Tukey J W 1962 The future of data analysis *Ann. Math. Statist.* **33** 1–67

5 Parameters

5.1 Compulsory Input Parameters

1: **scores – string**

Indicates which of the following scores are required.

scores = 'R'

The ranks.

scores = 'N'

The Normal scores, that is the expected value of the Normal order statistics.

scores = 'B'

The Blom version of the Normal scores.

scores = 'T'

The Tukey version of the Normal scores.

scores = 'V'

The van der Waerden version of the Normal scores.

scores = 'S'

The Savage scores, that is the expected value of the exponential order statistics.

Constraint: **scores** = 'R', 'N', 'B', 'T', 'V' or 'S'.

2: **ties – string**

Indicates which of the following methods is to be used to assign scores to tied observations.

ties = 'A'

The average of the scores for tied observations is used.

ties = 'L'

The lowest score in the group of ties is used.

ties = 'H'

The highest score in the group of ties is used.

ties = 'R'

The random number generator is used to randomly untie any group of tied observations.

ties = 'I'

Any ties are ignored, that is the scores are assigned to tied observations in the order that they appear in the data.

Constraint: **ties** = 'A', 'L', 'H', 'R' or 'I'.

3: **x(n) – double array**

The sample of observations, x_i, for $i = 1, 2, \ldots, n$.

5.2 Optional Input Parameters

1: **n – int32 scalar**

Default: The dimension of the arrays **x**, **r**. (An error is raised if these dimensions are not equal.)

n, the number of observations.

Constraint: $\mathbf{n} \geq 1$.

5.3 Input Parameters Omitted from the MATLAB Interface

iwrk

5.4 Output Parameters

1: **r(n) – double array**

Contains the scores, s_i, for $i = 1, 2, \ldots, n$, as specified by **scores**.

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, **scores** is an invalid character,
or **ties** is an invalid character,
or $\mathbf{n} < 1$.

7 Accuracy

For **scores** = 'R', the results should be accurate to ***machine precision***.

For **scores** = 'S', the results should be accurate to a small multiple of ***machine precision***.

For **scores** = 'N', the results should have a relative accuracy of at least $\max\left(100 \times \epsilon, 10^{-8}\right)$ where ϵ is the ***machine precision***.

For **scores** = 'B', 'T' or 'V', the results should have a relative accuracy of at least $\max\left(10 \times \epsilon, 10^{-12}\right)$.

8 Further Comments

If more accurate Normal scores are required g01da should be used with appropriate settings for the input parameter **etol**.

Note that when ties are resolved randomly the function g05na is used which calls the NAG random number generator g05ka. If you do not initialize the generator then the default seed will be used. If the function is called at different times using the same data and using either the default seed or a fixed seed, by calling g05kb, then the same permutation will arise and ties will thus be resolved in the same way. If you wish ties to be resolved differently then the generator should be initialized to a non-repeatable number using g05kc.

9 Example

```
scores = 'Savage';
ties = 'Average';
x = [2;
     0;
     2;
     2;
     0];
[r, ifail] = g01dh(scores, ties, x)
```

```
r =
    1.4500
    0.3250
    1.4500
    1.4500
    0.3250
ifail =
           0
```