NAG Toolbox for MATLAB

g02bs

1 Purpose

g02bs computes Kendall and/or Spearman non-parametric rank correlation coefficients for a set of data omitting cases with missing values from only those calculations involving the variables for which the values are missing; the data array is preserved, and the ranks of the observations are not available on exit from the function.

2 Syntax

```
[rr, ncases, cnt, ifail] = g02bs(n, x, miss, xmiss, itype, 'm', m)
```

3 Description

The input data consists of n observations for each of m variables, given as an array

$$[x_{ij}], \qquad i = 1,2,\ldots,n(n \geq 2), j = 1,2,\ldots,m(m \geq 2),$$

where x_{ij} is the ith observation on the jth variable. In addition each of the m variables may optionally have associated with it a value which is to be considered as representing a missing observation for that variable; the missing value for the jth variable is denoted by xm_j. Missing values need not be specified for all variables.

Let $w_{ij} = 0$ if the ith observation for the jth variable is a missing value, i.e., if a missing value, xm_j, has been declared for the jth variable, and $x_{ij} = xm_j$ (see also Section 7); and $w_{ij} = 1$ otherwise, for $i = 1,2,\ldots,n$; $j = 1,2,\ldots,m$.

The observations are first ranked, a pair of variables at a time as follows:

For a given pair of variables, j and l say, each of the observations x_{ij} for which the product $w_{ij}w_{il} = 1 (i = 1,2,\ldots,n)$ has associated with it an additional number, the 'rank' of the observation, which indicates the magnitude of that observation relative to the magnitude of the other observations on variable j for which $w_{ij}w_{il} = 1$.

The smallest of these valid observations for variable j is assigned to rank 1, the second smallest valid observation for variable j the rank 2, the third smallest rank 3, and so on until the largest such observation is given the rank n_{jl}, where

$$n_{jl} = \sum_{i=1}^{n} w_{ij}w_{il}.$$

If a number of cases all have the same value for the variable j, then they are each given an 'average' rank, e.g., if in attempting to assign the rank $h + 1$, k observations for which $w_{ij}w_{il} = 1$ were found to have the same value, then instead of giving them the ranks

$$h + 1, h + 2, \ldots, h + k,$$

all k observations would be assigned the rank

$$\frac{2h + k + 1}{2}$$

and the next value in ascending order would be assigned the rank

$$h + k + 1.$$

The variable l is then ranked in a similar way. The process is then repeated for all pairs of variables j and l, for $j = 1,2,\ldots,m$; $l = j,\ldots,m$. Let $y_{ij(l)}$ be the rank assigned to the observation x_{ij} when the jth and lth variables are being ranked, and $y_{il(j)}$ be the rank assigned to the observation x_{il} during the same process, for $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$ and $l = j, j+1, \ldots, m$.

The quantities calculated are:

(a) Kendall's tau rank correlation coefficients:

$$R_{jk} = \frac{\displaystyle\sum_{h=1}^{n}\sum_{i=1}^{n} w_{hj}w_{hk}w_{ij}w_{ik}\,\mathrm{sign}\left(y_{hj(k)} - y_{ij(k)}\right)\,\mathrm{sign}\left(y_{hk(j)} - y_{ik(j)}\right)}{\sqrt{\left[n_{jk}\left(n_{jk}-1\right) - T_{j(k)}\right]\left[n_{jk}\left(n_{jk}-1\right) - T_{k(j)}\right]}}, \qquad j,k = 1,2,\ldots,m,$$

where $n_{jk} = \displaystyle\sum_{i=1}^{n} w_{ij}w_{ik}$

and $\mathrm{sign}\, u = 1$ if $u > 0$

$\mathrm{sign}\, u = 0$ if $u = 0$

$\mathrm{sign}\, u = -1$ if $u < 0$

and $T_{j(k)} = \sum t_j\left(t_j - 1\right)$ where t_j is the number of ties of a particular value of variable j when the jth and kth variables are being ranked, and the summation is over all tied values of variable j.

(b) Spearman's rank correlation coefficients:

$$R^*_{jk} = \frac{n_{jk}\left(n_{jk}^2 - 1\right) - 6\displaystyle\sum_{i=1}^{n} w_{ij}w_{ik}\left(y_{ij(k)} - y_{ik(j)}\right)^2 - \tfrac{1}{2}\left(T^*_{j(k)} + T^*_{k(j)}\right)}{\sqrt{\left[n_{jk}\left(n_{jk}^2 - 1\right) - T^*_{j(k)}\right]\left[n_{jk}\left(n_{jk}^2 - 1\right) - T^*_{k(j)}\right]}}, \qquad j,k = 1,2,\ldots,m,$$

where $n_{jk} = \displaystyle\sum_{i=1}^{n} w_{ij}w_{ik}$

and $T^*_{j(k)} = \sum t_j\left(t_j^2 - 1\right)$, where t_j is the number of ties of a particular value of variable j when the jth and kth variables are being ranked, and the summation is over all tied values of variable j.

4 References

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

5 Parameters

5.1 Compulsory Input Parameters

1: **n – int32 scalar**

n, the number of observations or cases.

Constraint: $\mathbf{n} \geq 2$.

2: **x(ldx,m) – double array**

ldx, the first dimension of the array, must be at least **n**.

$\mathbf{x}(i,j)$ must be set to x_{ij}, the value of the ith observation on the jth variable, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$.

3: **miss(m) – int32 array**

$\mathbf{miss}(j)$ must be set equal to 1 if a missing value, xm_j, is to be specified for the jth variable in the array **x**, or set equal to 0 otherwise. Values of **miss** must be given for all m variables in the array **x**.

4: **xmiss(m) – double array**

xmiss(j) must be set to the missing value, xm_j, to be associated with the jth variable in the array **x**, for those variables for which missing values are specified by means of the array **miss** (see Section 7).

5: **itype – int32 scalar**

The type of correlation coefficients which are to be calculated.

itype $= -1$

Only Kendall's tau coefficients are calculated.

itype $= 0$

Both Kendall's tau and Spearman's coefficients are calculated.

itype $= 1$

Only Spearman's coefficients are calculated.

Constraint: **itype** $= -1$, 0 or 1.

5.2 Optional Input Parameters

1: **m – int32 scalar**

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

m, the number of variables.

Constraint: **m** ≥ 2.

5.3 Input Parameters Omitted from the MATLAB Interface

ldx, ldrr, ldcnt, kworka, kworkb, kworkc, kworkd, work1, work2

5.4 Output Parameters

1: **rr(ldrr,m) – double array**

The requested correlation coefficients.

If only Kendall's tau coefficients are requested (**itype** $= -1$), **rr**(j,k) contains Kendall's tau for the jth and kth variables.

if only Spearman's coefficients are requested (**itype** $= 1$), **rr**(j,k) contains Spearman's rank correlation coefficient for the jth and kth variables.

If both Kendall's tau and Spearman's coefficients are requested (**itype** $= 0$), the upper triangle of **rr** contains the Spearman coefficients and the lower triangle the Kendall coefficients. That is, for the jth and kth variables, where j is less than k, **rr**(j,k) contains the Spearman rank correlation coefficient, and **rr**(k,j) contains Kendall's tau, for $j,k = 1,2,\ldots,m$.

(Diagonal terms, **rr**(j,j), are unity for all three values of **itype**.)

2: **ncases – int32 scalar**

The minimum number of cases used in the calculation of any of the correlation coefficients (when cases involving missing values have been eliminated).

3: **cnt(ldcnt,m) – double array**

The number of cases, n_{jk}, actually used in the calculation of the rank correlation coefficient for the jth and kth variables, for $j,k = 1,2,\ldots,m$.

4: **ifail – int32 scalar**

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

6 Error Indicators and Warnings

Note: g02bs may return useful information for one or more of the following detected errors or warnings.

ifail $= 1$

On entry, $\mathbf{n} < 2$.

ifail $= 2$

On entry, $\mathbf{m} < 2$.

ifail $= 3$

On entry, $\mathbf{ldx} < \mathbf{n}$,
or $\mathbf{ldrr} < \mathbf{m}$,
or $\mathbf{ldcnt} < \mathbf{m}$.

ifail $= 4$

On entry, $\mathbf{itype} < -1$,
or $\mathbf{itype} > 1$.

ifail $= 5$

After observations with missing values were omitted, fewer than two cases remained for at least one pair of variables. (The pairs of variables involved can be determined by examination of the contents of the array **cnt**.) All correlation coefficients based on two or more cases are returned by the function even if **ifail** $= 5$.

7 Accuracy

You are warned of the need to exercise extreme care in your selection of missing values. g02bs treats all values in the inclusive range $(1 \pm \mathrm{ACC}) \times xm_j$, where xm_j is the missing value for variable j specified by you, and ACC is a machine-dependent constant as missing values for variable j.

You must therefore ensure that the missing value chosen for each variable is sufficiently different from all valid values for that variable so that none of the valid values fall within the range indicated above.

8 Further Comments

The time taken by g02bs depends on n and m, and the occurrence of missing values.

9 Example

```
n = int32(9);
x = [1.7, 1, 0.5;
     2.8, 4, 3;
     0.6, 6, 2.5;
     1.8, 9, 6;
     0.99, 4, 2.5;
     1.4, 2, 5.5;
     1.8, 9, 7.5;
     2.5, 7, 0;
     0.99, 5, 3];
miss = [int32(1);
     int32(1);
     int32(1)];
xmiss = [0.99;
     9;
     0];
itype = int32(0);
[rr, ncases, count, ifail] = g02bs(n, x, miss, xmiss, itype)
```

```
rr =
    1.0000    0.1000    0.4058
         0    1.0000    0.0896
    0.2760         0    1.0000
ncases =
           5
count =
     7     5     6
     5     7     6
     6     6     8
ifail =
           0
```