NAG Toolbox for MATLAB

g02bp

1 Purpose

g02bp computes Kendall and/or Spearman non-parametric rank correlation coefficients for a set of data omitting completely any cases with a missing observation for any variable; the data array is overwritten with the ranks of the observations.

2 Syntax

```
[x, miss, xmiss, rr, ncases, incase, ifail] = g02bp(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

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

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 $w_i = 0$ if observation i contains a missing value for any of those variables for which missing values have been declared; i.e., if $x_{ij} = xm_j$ for any j for which an xm_j has been assigned (see also Section 7); and $w_i = 1$ otherwise, for $i = 1,2,\ldots,n$.

The quantities calculated are:

(a) Ranks

For a given variable, j say, each of the observations x_{ij} for which $w_i = 1$, for $i = 1,2,\ldots,n$, has associated with it an additional number, the 'rank' of the observation, which indicates the magnitude of that observation relative to the magnitudes of the other observations on that same variable for which $w_i = 1$.

The smallest of these valid observations for variable j is assigned the rank 1, the second smallest observation for variable j the rank 2, the third smallest the rank 3, and so on until the largest such observation is given the rank n_c, where $n_c = \sum_{i=1}^{n} w_i$.

If a number of cases all have the same value for the given 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 $w_i = 1$ 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 process is repeated for each of the m variables.

Let y_{ij} be the rank assigned to the observation x_{ij} when the jth variable is being ranked. For those observations, i, for which $w_i = 0$, $y_{ij} = 0$, for $j = 1,2,\ldots,m$.

The actual observations x_{ij} are replaced by the ranks y_{ij}, for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$.

(b) Non-parametric rank correlation coefficients

(i) Kendall's tau:

$$R_{jk} = \frac{\sum_{h=1}^{n}\sum_{i=1}^{n} w_h w_i \operatorname{sign}\left(y_{hj} - y_{ij}\right) \operatorname{sign}(y_{hk} - y_{ik})}{\sqrt{\left[n_c(n_c-1) - T_j\right]\left[n_c(n_c-1) - T_k\right]}}, \qquad j,k = 1,2,\ldots,m,$$

where $n_c = \sum_{i=1}^{n} w_i$

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

$\operatorname{sign} u = 0$ if $u = 0$

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

and $T_j = \sum t_j\left(t_j - 1\right)$ where t_j is the number of ties of a particular value of variable j, and the summation is over all tied values of variable j.

(ii) Spearman's:

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

where $n_c = \sum_{i=1}^{n} w_i$

and $T^*_j = \sum t_j\left(t_j^2 - 1\right)$ where t_j is the number of ties of a particular value of variable j, 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 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**, **xbar**, **std**, **ssp**, **r**. (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, kworka, kworkb, kworkc, work1, work2

5.4 Output Parameters

1: **x(ldx,m) – double array**

x(i,j) contains the rank y_{ij} of the observation x_{ij}, for $i = 1,2,\ldots,n$ and $j = 1,2,\ldots,m$. (For those observations containing missing values, and therefore excluded from the calculation, $y_{ij} = 0$, for $j = 1,2,\ldots,m$.)

2: **miss(m) – int32 array**

The array **miss** contains the function, and the information it contained on entry is lost.

3: **xmiss(m) – double array**

The array **xmiss** contains the function, and the information it contained on entry is lost.

4: **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, $\mathbf{rr}(j,k)$ contains the Spearman rank correlation coefficient, and $\mathbf{rr}(k,j)$ contains Kendall's tau, for $j,k = 1,2,\ldots,m$.

(Diagonal terms, $\mathbf{rr}(j,j)$, are unity for all three values of **itype**.)

5: **ncases – int32 scalar**

The number of cases, n_c, actually used in the calculations (when cases involving missing values have been eliminated).

6: **incase(n) – int32 array**

$\mathbf{incase}(i)$ holds the value 1 if the ith case was included in the calculations, and the value 0 if the ith case contained a missing value for at least one variable. That is, $\mathbf{incase}(i) = w_i$ (see Section 3), for $i = 1,2,\ldots,n$.

7: **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, $\mathbf{n} < 2$.

ifail $= 2$

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

ifail $= 3$

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

ifail $= 4$

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

ifail $= 5$

After observations with missing values were omitted, fewer than 2 cases remained.

7 Accuracy

You are warned of the need to exercise extreme care in your selection of missing values. g02bp 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 g02bp 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(0);
     int32(1)];
xmiss = [0.99;
     4.566269392561278e+257;
     0];
itype = int32(0);
[xOut, missOut, xmissOut, rr, ncases, incase, ifail] = g02bp(n, x, miss,
xmiss, itype)
```

```
xOut =
    3.0000    1.0000    1.0000
    6.0000    3.0000    3.0000
    1.0000    4.0000    2.0000
    4.5000    5.5000    5.0000
         0         0         0
    2.0000    2.0000    4.0000
    4.5000    5.5000    6.0000
         0         0         0
         0         0         0
missOut =
           1
           3
           1
xmissOut =
    0.9900
         0
         0
rr =
    1.0000    0.2941    0.4058
    0.1429    1.0000    0.7537
    0.2760    0.5521    1.0000
ncases =
           6
incase =
           1
           1
           1
           1
           0
           1
           1
           0
           0
ifail =
           0
```