NAG Fortran Library Routine Document

G02BRF

Note: before using this routine, please read the Users' Note for your implementation to check the interpretation of ***bold italicised*** terms and other implementation-dependent details.

1 Purpose

G02BRF 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 preserved, and the ranks of the observations are not available on exit from the routine.

2 Specification

```
 SUBROUTINE G02BRF(N, M, X, IX, MISS, XMISS, ITYPE, RR, IRR, NCASES,
1                  INCASE, KWORKA, KWORKB, KWORKC, WORK1, WORK2, IFAIL)
 INTEGER           N, M, IX, MISS(M), ITYPE, IRR, NCASES, INCASE(N),
1                  KWORKA(N), KWORKB(N), KWORKC(N), IFAIL
 real              X(IX,M), XMISS(M), RR(IRR,M), WORK1(N), WORK2(N)
```

3 Description

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

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

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 $w_i = 0$ 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 observations are first ranked as follows.

For a given variable, j say, each of the observations x_{ij} for which $w_i = 1$, $(i = 1,2,\ldots,n)$ has associated with it an additional number, the 'rank' of the observation, which indicates the magnitude of that observation relative to the 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 quantities calculated are:

(a) Kendall's tau rank correlation coefficients:

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

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 $T_j = \sum t_j(t_j - 1)$ 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.

(b) Spearman's rank correlation coefficients:

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

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

and $T^*_j = \sum t_j(t_j^2 - 1)$ 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) *Nonparametric Statistics for the Behavioral Sciences* McGraw-Hill

5 Parameters

1: N – INTEGER *Input*

On entry: the number, n, of observations or cases.

Constraint: N $\geq$ 2.

2: M – INTEGER *Input*

On entry: the number, m, of variables.

Constraint: M $\geq$ 2.

3: X(IX,M) – ***real*** array *Input*

On entry: X(i, j) must be set to x_{ij}, the value of the ith observation on the jth variable, where $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$.

4: IX – INTEGER *Input*

On entry: the first dimension of the array X as declared in the (sub)program from which G02BRF is called.

Constraint: IX $\geq$ N.

5: MISS(M) – INTEGER array *Input/Output*

On entry: 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.

On exit: the array MISS is overwritten by the routine, and the information it contained on entry is lost.

6: XMISS(M) – ***real*** array *Input/Output*

On entry: 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).

On exit: the array XMISS is overwritten by the routine, and the information it contained on entry is lost.

7: ITYPE – INTEGER *Input*

On entry: the type of correlation coefficients which are to be calculated. If ITYPE $= -1$, only Kendall's tau coefficients are calculated; if ITYPE $= 0$, both Kendall's tau and Spearman's coefficients are calculated; if ITYPE $= 1$, only Spearman's coefficients are calculated.

8: RR(IRR,M) – ***real*** array *Output*

On exit: the requested correlation coefficients. If only Kendall's tau coefficients are requested (ITYPE $= -1$), then RR(j,k) contains Kendall's tau for the jth and kth variables; if only Spearman's coefficients are requested (ITYPE $= 1$), then 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 $= 0$), then 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.)

9: IRR – INTEGER *Input*

On entry: the first dimension of the array RR as declared in the (sub)program from which G02BRF is called.

Constraint: IRR $\geq$ M.

10: NCASES – INTEGER *Output*

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

11: INCASE(N) – INTEGER array *Output*

On exit: 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, INCASE(i) $= w_i$ (see Section 3), for $i = 1,2,\ldots,n$.

12: KWORKA(N) – INTEGER array *Workspace*
13: KWORKB(N) – INTEGER array *Workspace*
14: KWORKC(N) – INTEGER array *Workspace*
15: WORK1(N) – ***real*** array *Workspace*
16: WORK2(N) – ***real*** array *Workspace*

17: IFAIL – INTEGER *Input/Output*

On entry: IFAIL must be set to 0, -1 or 1. Users who are unfamiliar with this parameter should refer to Chapter P01 for details.

On exit: IFAIL $= 0$ unless the routine detects an error (see Section 6).

For environments where it might be inappropriate to halt program execution when an error is detected, the value -1 or 1 is recommended. If the output of error messages is undesirable, then the value 1 is recommended. Otherwise, for users not familiar with this parameter the recommended value is 0. **When the value -1 or 1 is used it is essential to test the value of IFAIL on exit.**

6 Error Indicators and Warnings

If on entry IFAIL $= 0$ or -1, explanatory error messages are output on the current error message unit (as defined by X04AAF).

Errors or warnings detected by the routine:

IFAIL $= 1$

On entry, N < 2.

IFAIL $= 2$

On entry, M < 2.

IFAIL $= 3$

On entry, IX $<$ N,
or IRR $<$ M.

IFAIL $= 4$

On entry, ITYPE < -1,
or ITYPE > 1.

IFAIL $= 5$

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

7 Accuracy

Users are warned of the need to exercise extreme care in their selection of missing values. The routine treats all values in the inclusive range $(1 \pm \text{ACC}) \times xm_j$, where xm_j is the missing value for variable j specified by the user, and ACC is a machine-dependent constant (see the Users' Note for your implementation) as missing values for variable j.

The user 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 the routine depends on n and m, and the occurrence of missing values.

9 Example

The following program reads in a set of data consisting of nine observations on each of three variables. Missing values of 0.99 and 0.0 are declared for the first and third variables respectively; no missing value is specified for the second variable. The program then calculates and prints both Kendall's tau and Spearman's rank correlation coefficients for all three variables, omitting completely all cases containing missing values; cases 5, 8 and 9 are therefore eliminated, leaving only six cases in the calculations.

9.1 Program Text

Note: the listing of the example program presented below uses ***bold italicised*** terms to denote precision-dependent details. Please read the Users' Note for your implementation to check the interpretation of these terms. As explained in the Essential Introduction to this manual, the results produced may not be identical for all implementations.

```
*     G02BRF Example Program Text
*     Mark 14 Revised.  NAG Copyright 1989.
*     .. Parameters ..
      INTEGER          M, N, IA, ICORR
      PARAMETER        (M=3,N=9,IA=N,ICORR=M)
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
*     .. Local Scalars ..
      INTEGER          I, IFAIL, ITYPE, J, NCASES
*     .. Local Arrays ..
      real             A(IA,M), CORR(ICORR,M), WA(N), WB(N), XMISS(M)
      INTEGER          INOUT(N), IW(N), JW(N), KW(N), MISS(M)
*     .. External Subroutines ..
      EXTERNAL         G02BRF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G02BRF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) ((A(I,J),J=1,M),I=1,N)
      WRITE (NOUT,*)
      WRITE (NOUT,99999) 'Number of variables (columns) =', M
      WRITE (NOUT,99999) 'Number of cases     (rows)    =', N
      WRITE (NOUT,*)
      WRITE (NOUT,*) 'Data matrix is:-'
      WRITE (NOUT,*)
      WRITE (NOUT,99998) (J,J=1,M)
      WRITE (NOUT,99997) (I,(A(I,J),J=1,M),I=1,N)
      WRITE (NOUT,*)
*
*     Set up missing values before calling routine
*
      MISS(1) = 1
      MISS(2) = 0
      MISS(3) = 1
      XMISS(1) = 0.99e0
      XMISS(3) = 0.00e0
      ITYPE = 0
      IFAIL = 1
*
      CALL G02BRF(N,M,A,IA,MISS,XMISS,ITYPE,CORR,ICORR,NCASES,INOUT,IW,
     +            JW,KW,WA,WB,IFAIL)
*
      IF (IFAIL.NE.0) THEN
         WRITE (NOUT,99999) 'Routine fails, IFAIL =', IFAIL
      ELSE
         WRITE (NOUT,*) 'Matrix of rank correlation coefficients:'
         WRITE (NOUT,*) 'Upper triangle -- Spearman''s'
         WRITE (NOUT,*) 'Lower triangle -- Kendall''s tau'
         WRITE (NOUT,*)
         WRITE (NOUT,99998) (I,I=1,M)
         WRITE (NOUT,99997) (I,(CORR(I,J),J=1,M),I=1,M)
         WRITE (NOUT,*)
         WRITE (NOUT,99999) 'Number of cases actually used:', NCASES
      END IF
      STOP
*
99999 FORMAT (1X,A,I3)
99998 FORMAT (1X,3I12)
99997 FORMAT (1X,I3,3F12.4)
      END
```

9.2 Program Data

```
G02BRF Example Program Data
1.70      1.00      0.50
2.80      4.00      3.00
0.60      6.00      2.50
1.80      9.00      6.00
0.99      4.00      2.50
1.40      2.00      5.50
1.80      9.00      7.50
2.50      7.00      0.00
0.99      5.00      3.00
```

9.3 Program Results

```
 G02BRF Example Program Results

 Number of variables (columns) =  3
 Number of cases     (rows)    =  9

 Data matrix is:-

            1           2           3
   1      1.7000      1.0000      0.5000
   2      2.8000      4.0000      3.0000
   3      0.6000      6.0000      2.5000
   4      1.8000      9.0000      6.0000
   5      0.9900      4.0000      2.5000
   6      1.4000      2.0000      5.5000
   7      1.8000      9.0000      7.5000
   8      2.5000      7.0000      0.0000
   9      0.9900      5.0000      3.0000

 Matrix of rank correlation coefficients:
 Upper triangle -- Spearman's
 Lower triangle -- Kendall's tau

            1           2           3
   1      1.0000      0.2941      0.4058
   2      0.1429      1.0000      0.7537
   3      0.2760      0.5521      1.0000

 Number of cases actually used:  6
```