NAG Library Routine Document

G02BFF

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

G02BFF computes means and standard deviations of variables, sums of squares and cross-products about zero and correlation-like coefficients for a set of data omitting cases with missing values from only those calculations involving the variables for which the values are missing.

2 Specification

```
SUBROUTINE G02BFF (N, M, X, LDX, MISS, XMISS, XBAR, STD, SSPZ, LDSSPZ,     &
                   RZ, LDRZ, NCASES, CNT, LDCNT, IFAIL)
INTEGER            N, M, LDX, MISS(M), LDSSPZ, LDRZ, NCASES, LDCNT,        &
                   IFAIL
REAL (KIND=nag_wp) X(LDX,M), XMISS(M), XBAR(M), STD(M),                     &
                   SSPZ(LDSSPZ,M), RZ(LDRZ,M), CNT(LDCNT,M)
```

3 Description

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

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

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 $w_{ij} = 0$ 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 $i = 1,2,\ldots,n$ and $j = 1,2,\ldots,m$.

The quantities calculated are:

(a) Means:

$$\bar{x}_j = \frac{\sum_{i=1}^{n} w_{ij}x_{ij}}{\sum_{i=1}^{n} w_{ij}}, \quad j = 1,2,\ldots,m.$$

(b) Standard deviations:

$$s_j = \sqrt{\frac{\sum_{i=1}^{n} w_{ij}\big(x_{ij} - \bar{x}_j\big)^2}{\sum_{i=1}^{n} w_{ij} - 1}}, \quad j = 1,2,\ldots,m.$$

(c) Sums of squares and cross-products about zero:

$$\tilde{S}_{jk} = \sum_{i=1}^{n} w_{ij}w_{ik}x_{ij}x_{ik}, \quad j,k = 1,2,\ldots,m.$$

(d) Correlation-like coefficients:

$$\tilde{R}_{jk} = \frac{\tilde{S}_{jk}}{\sqrt{\tilde{S}_{jj(k)}j(k)\tilde{S}_{kk(j)}}}, \quad j,k = 1,2,\ldots,m,$$

where $\tilde{S}_{jj(k)} = \sum_{i=1}^{n} w_{ij}w_{ik}x_{ij}^2$ and $\tilde{S}_{kk(j)} = \sum_{i=1}^{n} w_{ik}w_{ij}x_{ik}^2$

(i.e., the sums of squares about zero are based on the same set of observations as are used in the calculation of the numerator).

If $\tilde{S}_{jj(k)}$ or $\tilde{S}_{kk(j)}$ is zero, $\tilde{R}_{jk}$ is set to zero.

(e) The number of cases used in the calculation of each of the correlation-like coefficients:

$$c_{jk} = \sum_{i=1}^{n} w_{ij}w_{ik}, \quad j,k = 1,2,\ldots,m.$$

(The diagonal terms, c_{jj}, for $j = 1,2,\ldots,m$, also give the number of cases used in the calculation of the means $\bar{x}_j$ and the standard deviations s_j.)

4 References

None.

5 Arguments

1: N – INTEGER *Input*

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

Constraint: N $\geq$ 2.

2: M – INTEGER *Input*

On entry: m, the number of variables.

Constraint: M $\geq$ 2.

3: X(LDX, M) – REAL (KIND=nag_wp) array *Input*

On entry: 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 $j = 1,2,\ldots,m$.

4: LDX – INTEGER *Input*

On entry: the first dimension of the array X as declared in the (sub)program from which G02BFF is called.

Constraint: LDX $\geq$ N.

5: MISS(M) – INTEGER array *Input*

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.

6: XMISS(M) – REAL (KIND=nag_wp) array *Input*

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).

7: XBAR(M) – REAL (KIND=nag_wp) array *Output*

On exit: the mean value, $\bar{x}_j$, of the jth variable, for $j = 1, 2, \ldots, m$.

8: STD(M) – REAL (KIND=nag_wp) array *Output*

On exit: the standard deviation, s_j, of the jth variable, for $j = 1, 2, \ldots, m$.

9: SSPZ(LDSSPZ, M) – REAL (KIND=nag_wp) array *Output*

On exit: SSPZ(j, k) is the cross-product about zero, $\tilde{S}_{jk}$, for $j = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, m$.

10: LDSSPZ – INTEGER *Input*

On entry: the first dimension of the array SSPZ as declared in the (sub)program from which G02BFF is called.

Constraint: LDSSPZ $\geq$ M.

11: RZ(LDRZ, M) – REAL (KIND=nag_wp) array *Output*

On exit: RZ(j, k) is the correlation-like coefficient, $\tilde{R}_{jk}$, between the jth and kth variables, for $j = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, m$.

12: LDRZ – INTEGER *Input*

On entry: the first dimension of the array RZ as declared in the (sub)program from which G02BFF is called.

Constraint: LDRZ $\geq$ M.

13: NCASES – INTEGER *Output*

On exit: the minimum number of cases used in the calculation of any of the sums of squares and cross-products and correlation-like coefficients (when cases involving missing values have been eliminated).

14: CNT(LDCNT, M) – REAL (KIND=nag_wp) array *Output*

On exit: CNT(j, k) is the number of cases, c_{jk}, actually used in the calculation of $\tilde{S}_{jk}$, and $\tilde{R}_{jk}$, the sum of cross-products and correlation-like coefficient for the jth and kth variables, for $j = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, m$.

15: LDCNT – INTEGER *Input*

On entry: must specify the first dimension of the array CNT as declared in the (sub)program from which G02BFF is called.

Constraint: LDCNT $\geq$ M.

16: IFAIL – INTEGER *Input/Output*

On entry: IFAIL must be set to 0, -1 or 1. If you are unfamiliar with this argument you should refer to Section 3.4 in How to Use the NAG Library and its Documentation for details.

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, because for this routine the values of the output arguments may be useful even if IFAIL $\neq 0$ on exit, the recommended value is -1. **When the value -1 or 1 is used it is essential to test the value of IFAIL on exit.**

On exit: IFAIL $= 0$ unless the routine detects an error or a warning has been flagged (see Section 6).

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).

Note: G02BFF may return useful information for one or more of the following detected errors or warnings.

Errors or warnings detected by the routine:

IFAIL $= 1$

On entry, N < 2.

IFAIL $= 2$

On entry, M < 2.

IFAIL $= 3$

On entry, LDX $<$ N,
or LDSSPZ $<$ M,
or LDRZ $<$ M,
or LDCNT $<$ M.

IFAIL $= 4$

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 means, standard deviations, sums of squares and cross-products, and correlation-like coefficients based on two or more cases are returned by the routine even if IFAIL $= 4$.

IFAIL $= -99$

An unexpected error has been triggered by this routine. Please contact NAG.

See Section 3.9 in How to Use the NAG Library and its Documentation for further information.

IFAIL $= -399$

Your licence key may have expired or may not have been installed correctly.

See Section 3.8 in How to Use the NAG Library and its Documentation for further information.

IFAIL $= -999$

Dynamic memory allocation failed.

See Section 3.7 in How to Use the NAG Library and its Documentation for further information.

7 Accuracy

G02BFF does not use ***additional precision*** arithmetic for the accumulation of scalar products, so there may be a loss of significant figures for large n.

You are warned of the need to exercise extreme care in your selection of missing values. G02BFF treats all values in the inclusive range $\left(1 \pm 0.1^{(\text{X02BEF}-2)}\right) \times xm_j$, where xm_j is the missing value for variable j specified in XMISS.

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 Parallelism and Performance

G02BFF is not threaded in any implementation.

9 Further Comments

The time taken by G02BFF depends on n and m, and the occurrence of missing values.

The routine uses a two-pass algorithm.

10 Example

This example reads in a set of data consisting of five observations on each of three variables. Missing values of 0.0, -1.0 and 0.0 are declared for the first, second and third variables respectively. The means, standard deviations, sums of squares and cross-products about zero, and correlation-like coefficients for all three variables are then calculated and printed, omitting cases with missing values from only those calculations involving the variables for which the values are missing. The program therefore omits cases 4 and 5 in calculating the correlation between the first and second variables, and cases 3 and 4 for the first and third variables, etc.

10.1 Program Text

```
    Program g02bffe

!     G02BFF Example Program Text

!     Mark 26 Release. NAG Copyright 2016.

!     .. Use Statements ..
      Use nag_library, Only: g02bff, nag_wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter               :: nin = 5, nout = 6
!     .. Local Scalars ..
      Integer                          :: i, ifail, ldcnt, ldrz, ldsspz, ldx,  &
                                          m, n, ncases
!     .. Local Arrays ..
      Real (Kind=nag_wp), Allocatable  :: cnt(:,:), rz(:,:), sspz(:,:),        &
                                          std(:), x(:,:), xbar(:), xmiss(:)
      Integer, Allocatable             :: miss(:)
!     .. Executable Statements ..
      Write (nout,*) 'G02BFF Example Program Results'
      Write (nout,*)

!     Skip heading in data file
      Read (nin,*)

!     Read in the problem size
      Read (nin,*) n, m

      ldcnt = m
      ldrz = m
      ldsspz = m
      ldx = n
      Allocate (cnt(ldcnt,m),rz(ldrz,m),sspz(ldsspz,m),std(m),x(ldx,m),        &
        xbar(m),xmiss(m),miss(m))

!     Read in data
      Read (nin,*)(x(i,1:m),i=1,n)

!     Read in missing value flags
      Read (nin,*) miss(1:m)
      Read (nin,*) xmiss(1:m)

!     Display data
```

```
      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)(i,i=1,m)
      Write (nout,99997)(i,x(i,1:m),i=1,n)
      Write (nout,*)

!     Calculate summary statistics
      ifail = 0
      Call g02bff(n,m,x,ldx,miss,xmiss,xbar,std,sspz,ldsspz,rz,ldrz,ncases,    &
        cnt,ldcnt,ifail)

!     Display results
      Write (nout,*) 'Variable   Mean    St. dev.'
      Write (nout,99996)(i,xbar(i),std(i),i=1,m)
      Write (nout,*)
      Write (nout,*) 'Sums of squares and cross-products about' // ' zero'
      Write (nout,99998)(i,i=1,m)
      Write (nout,99997)(i,sspz(i,1:m),i=1,m)
      Write (nout,*)
      Write (nout,*) 'Correlation-like coefficients'
      Write (nout,99998)(i,i=1,m)
      Write (nout,99997)(i,rz(i,1:m),i=1,m)
      Write (nout,*)
      Write (nout,99999)                                                       &
        'Minimum number of cases used for any pair of variables: ', ncases
      Write (nout,*)
      Write (nout,*) 'Numbers used for each pair are:'
      Write (nout,99998)(i,i=1,m)
      Write (nout,99997)(i,cnt(i,1:m),i=1,m)

99999 Format (1X,A,I5)
99998 Format (1X,6I12)
99997 Format (1X,I3,3F12.4)
99996 Format (1X,I5,2F11.4)
    End Program g02bffe
```

10.2 Program Data

```
G02BFF Example Program Data
5  3                        :: N, M
 2.0    3.0    3.0
 4.0    6.0    4.0
 9.0    9.0    0.0
 0.0   12.0    2.0
12.0   -1.0    5.0          :: End of X
  1      1      1           :: MISS
 0.0   -1.0    0.0          :: XMISS
```

10.3 Program Results

```
 G02BFF Example Program Results

 Number of variables (columns) =    3
 Number of cases     (rows)    =    5

 Data matrix is:-

            1           2           3
   1      2.0000      3.0000      3.0000
   2      4.0000      6.0000      4.0000
   3      9.0000      9.0000      0.0000
   4      0.0000     12.0000      2.0000
   5     12.0000     -1.0000      5.0000

 Variable   Mean    St. dev.
     1     6.7500     4.5735
     2     7.5000     3.8730
```

```
     3     3.5000     1.2910

 Sums of squares and cross-products about zero
            1           2           3
   1    245.0000    111.0000     82.0000
   2    111.0000    270.0000     57.0000
   3     82.0000     57.0000     54.0000

 Correlation-like coefficients
            1           2           3
   1      1.0000      0.9840      0.9055
   2      0.9840      1.0000      0.7699
   3      0.9055      0.7699      1.0000

 Minimum number of cases used for any pair of variables:     3

 Numbers used for each pair are:
            1           2           3
   1      4.0000      3.0000      3.0000
   2      3.0000      4.0000      3.0000
   3      3.0000      3.0000      4.0000
```