NAG Library Routine Document

G02BHF

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

G02BHF computes means and standard deviations, sums of squares and cross-products of deviations from means, and Pearson product-moment correlation coefficients for selected variables omitting completely any cases with a missing observation for any variable (either over all variables in the dataset or over only those variables in the selected subset).

2 Specification

```
SUBROUTINE G02BHF (N, M, X, LDX, MISS, XMISS, MISTYP, NVARS, KVAR, XBAR,    &
                   STD, SSP, LDSSP, R, LDR, NCASES, IFAIL)
INTEGER            N, M, LDX, MISS(M), MISTYP, NVARS, KVAR(NVARS),          &
                   LDSSP, LDR, NCASES, IFAIL
REAL (KIND=nag_wp) X(LDX,M), XMISS(M), XBAR(NVARS), STD(NVARS),             &
                   SSP(LDSSP,NVARS), R(LDR,NVARS)
```

3 Description

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

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

where x_{ij} is the ith observation on the jth variable, together with the subset of these variables, $v_1, v_2, \dots, v_p$, for which information is required.

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. The missing values can be utilized in two slightly different ways; you can indicate which scheme is required.

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

Secondly, let $w_i = 0$ if observation i contains a missing value for any of those variables in the selected subset $v_1, v_2, \dots, v_p$ for which missing values have been declared, i.e., if $x_{ij} = xm_j$ for any j ($j = v_1, v_2, \dots, v_p$) for which an xm_j has been assigned (see also Section 7); and $w_i = 1$ otherwise, for $i = 1, 2, \dots, n$.

The quantities calculated are:

(a) Means:

$$\bar{x}_j = \frac{\sum_{i=1}^{n} w_i x_{ij}}{\sum_{i=1}^{n} w_i}, \quad j = v_1, v_2, \dots, v_p.$$

(b) Standard deviations:

$$s_j = \sqrt{\frac{\sum_{i=1}^{n} w_i (x_{ij} - \bar{x}_j)^2}{\sum_{i=1}^{n} w_i - 1}}, \quad j = v_1, v_2, \ldots, v_p.$$

(c) Sums of squares and cross-products of deviations from means:

$$S_{jk} = \sum_{i=1}^{n} w_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \quad j,k = v_1, v_2, \ldots, v_p.$$

(d) Pearson product-moment correlation coefficients:

$$R_{jk} = \frac{S_{jk}}{\sqrt{S_{jj} S_{kk}}}, \quad j,k = v_1, v_2, \ldots, v_p.$$

If S_{jj} or S_{kk} is zero, R_{jk} is set to zero.

4 References

None.

5 Parameters

1: N – INTEGER *Input*

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

Constraint: $\mathrm{N} \geq 2$.

2: M – INTEGER *Input*

On entry: m, the number of variables.

Constraint: $\mathrm{M} \geq 2$.

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

On entry: $\mathrm{X}(i,j)$ 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 G02BHF is called.

Constraint: $\mathrm{LDX} \geq \mathrm{N}$.

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 (KIND=nag_wp) 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: MISTYP – INTEGER *Input*

On entry: indicates the manner in which missing observations are to be treated.

MISTYP = 1
A case is excluded if it contains a missing value for any of the variables $1, 2, \ldots, m$.

MISTYP = 0
A case is excluded if it contains a missing value for any of the $p(\leq m)$ variables specified in the array KVAR.

8: NVARS – INTEGER *Input*

On entry: p, the number of variables for which information is required.

Constraint: $2 \leq$ NVARS $\leq$ M.

9: KVAR(NVARS) – INTEGER array *Input*

On entry: KVAR(j) must be set to the column number in X of the jth variable for which information is required, for $j = 1, 2, \ldots, p$.

Constraint: $1 \leq$ KVAR(j) $\leq$ M, for $j = 1, 2, \ldots, p$.

10: XBAR(NVARS) – REAL (KIND=nag_wp) array *Output*

On exit: the mean value, of $\bar{x}_j$, of the variable specified in KVAR(j), for $j = 1, 2, \ldots, p$.

11: STD(NVARS) – REAL (KIND=nag_wp) array *Output*

On exit: the standard deviation, s_j, of the variable specified in KVAR(j), for $j = 1, 2, \ldots, p$.

12: SSP(LDSSP, NVARS) – REAL (KIND=nag_wp) array *Output*

On exit: SSP(j, k) is the cross-product of deviations, S_{jk}, for the variables specified in KVAR(j) and KVAR(k), for $j = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, p$.

13: LDSSP – INTEGER *Input*

On entry: the first dimension of the array SSP as declared in the (sub)program from which G02BHF is called.

Constraint: LDSSP $\geq$ NVARS.

14: R(LDR, NVARS) – REAL (KIND=nag_wp) array *Output*

On exit: R(j, k) is the product-moment correlation coefficient, R_{jk}, between the variables specified in KVAR(j) and KVAR(k), for $j = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, p$.

15: LDR – INTEGER *Input*

On entry: the first dimension of the array R as declared in the (sub)program from which G02BHF is called.

Constraint: LDR $\geq$ NVARS.

16: NCASES – INTEGER *Output*

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

17: IFAIL – INTEGER *Input/Output*

On entry: IFAIL must be set to 0, -1 or 1. If you are unfamiliar with this parameter you should refer to Section 3.3 in the Essential Introduction 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, if you are 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.**

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).

Errors or warnings detected by the routine:

IFAIL $= 1$

On entry, N < 2.

IFAIL $= 2$

On entry, NVARS < 2,
or NVARS $>$ M.

IFAIL $= 3$

On entry, LDX $<$ N,
or LDSSP $<$ NVARS,
or LDR $<$ NVARS.

IFAIL $= 4$

On entry, KVAR$(j) < 1$,
or KVAR$(j) >$ M for some $j = 1, 2, \dots,$NVARS.

IFAIL $= 5$

On entry, MISTYP $\neq 1$ or 0

IFAIL $= 6$

After observations with missing values were omitted, no cases remained.

IFAIL $= 7$

After observations with missing values were omitted, only one case remained.

IFAIL $= -99$

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

See Section 3.8 in the Essential Introduction for further information.

IFAIL $= -399$

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

See Section 3.7 in the Essential Introduction for further information.

IFAIL $= -999$

Dynamic memory allocation failed.

See Section 3.6 in the Essential Introduction for further information.

7 Accuracy

G02BHF 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. G02BHF treats all values in the inclusive range $\left(1 \pm 0.1^{(\mathrm{X02BEF}-2)}\right) \times xm_j$, 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

Not applicable.

9 Further Comments

The time taken by G02BHF depends on n and p, 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 four variables. Missing values of 0.0 are declared for the second and fourth variables; no missing values are specified for the first and third variables. The means, standard deviations, sums of squares and cross-products of deviations from means, and Pearson product-moment correlation coefficients for the fourth, first and second variables are then calculated and printed, omitting completely all cases containing missing values for these three selected variables; cases 3 and 4 are therefore eliminated, leaving only three cases in the calculations.

10.1 Program Text

```
    Program g02bhfe

!     G02BHF Example Program Text

!     Mark 25 Release. NAG Copyright 2014.

!     .. Use Statements ..
      Use nag_library, Only: g02bhf, nag_wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter               :: nin = 5, nout = 6
!     .. Local Scalars ..
      Integer                          :: i, ifail, ldr, ldssp, ldx, m,       &
                                          mistyp, n, ncases, nvars
!     .. Local Arrays ..
      Real (Kind=nag_wp), Allocatable  :: r(:,:), ssp(:,:), std(:), x(:,:),   &
                                          xbar(:), xmiss(:)
```

```
      Integer, Allocatable             :: kvar(:), miss(:)
!     .. Executable Statements ..
      Write (nout,*) 'G02BHF Example Program Results'
      Write (nout,*)

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

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

      ldr = nvars
      ldssp = nvars
      ldx = n
      Allocate (r(ldr,nvars),ssp(ldssp,nvars),std(nvars),x(ldx,m),xbar(nvars), &
        xmiss(m),kvar(nvars),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)

!     Read in column IDs
      Read (nin,*) kvar(1:nvars)

!     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,99998)(i,i=1,m)
      Write (nout,99997)(i,x(i,1:m),i=1,n)
      Write (nout,*)

!     Calculate summary statistics
      ifail = 0
      Call g02bhf(n,m,x,ldx,miss,xmiss,mistyp,nvars,kvar,xbar,std,ssp,ldssp,r, &
        ldr,ncases,ifail)

!     Display results
      Write (nout,*) 'Variable   Mean    St. dev.'
      Write (nout,99995)(kvar(i),xbar(i),std(i),i=1,nvars)
      Write (nout,*)
      Write (nout,*) 'Sums of squares and cross-products of deviations'
      Write (nout,99998) kvar(1:nvars)
      Write (nout,99996)(kvar(i),ssp(i,1:nvars),i=1,nvars)
      Write (nout,*)
      Write (nout,*) 'Correlation coefficients'
      Write (nout,99998) kvar(1:nvars)
      Write (nout,99996)(kvar(i),r(i,1:nvars),i=1,nvars)
      Write (nout,*)
      Write (nout,99999) 'Number of cases actually used:', ncases

99999 Format (1X,A,I5)
99998 Format (1X,4I12)
99997 Format (1X,I3,4F12.4)
99996 Format (1X,I3,3F12.4)
99995 Format (1X,I5,2F11.4)
    End Program g02bhfe
```

10.2 Program Data

```
G02BHF Example Program Data
5  4  3  0                 :: N, M, NVARS, MISTYP
 3.0    3.0    1.0    2.0
 6.0    4.0   -1.0    4.0
 9.0    0.0    5.0    9.0
12.0    2.0    0.0    0.0
-1.0    5.0    4.0   12.0  :: End of X
  0      1      0      1   :: MISS
 0.0    0.0    0.0    0.0  :: XMISS
4  1  2                    :: KVAR
```

10.3 Program Results

```
 G02BHF Example Program Results

 Number of variables (columns) =    4
 Number of cases     (rows)    =    5

 Data matrix is:-
             1           2           3           4
   1      3.0000      3.0000      1.0000      2.0000
   2      6.0000      4.0000     -1.0000      4.0000
   3      9.0000      0.0000      5.0000      9.0000
   4     12.0000      2.0000      0.0000      0.0000
   5     -1.0000      5.0000      4.0000     12.0000

 Variable   Mean    St. dev.
     4     6.0000     5.2915
     1     2.6667     3.5119
     2     4.0000     1.0000

 Sums of squares and cross-products of deviations
             4           1           2
   4     56.0000    -30.0000     10.0000
   1    -30.0000     24.6667     -4.0000
   2     10.0000     -4.0000      2.0000

 Correlation coefficients
             4           1           2
   4      1.0000     -0.8072      0.9449
   1     -0.8072      1.0000     -0.5695
   2      0.9449     -0.5695      1.0000

 Number of cases actually used:    3
```