NAG Fortran 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 data set or over only those variables in the selected subset).

2 Specification

```
 SUBROUTINE G02BHF(N, M, X, IX, MISS, XMISS, MISTYP, NVARS, KVAR, XBAR,
1                  STD, SSP, ISSP, R, IR, NCASES, IFAIL)
 INTEGER           N, M, IX, MISS(M), MISTYP, NVARS, KVAR(NVARS), ISSP,
1                  IR, NCASES, IFAIL
 real              X(IX,M), XMISS(M), XBAR(NVARS), STD(NVARS),
1                  SSP(ISSP,NVARS), R(IR,NVARS)
```

3 Description

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

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

where x_{ij} is the ith observation on the jth variable, together with the subset of these variables, $v_1, v_2, \ldots, v_p$, 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 utilised in two slightly different ways; the user indicating which scheme is required.

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

Secondly, let $w_i = 0$ if observation i contains a missing value for any of those variables in the selected subset $v_1, v_2, \ldots, v_p$ for which missing values have been declared, i.e., if $x_{ij} = xm_j$ for any j ($j = v_1, v_2, \ldots, v_p$) 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) Means:

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

(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: 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, for $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$.

4: IX – INTEGER *Input*

On entry: the first dimension of the array X as declared in the (sub)program from which G02BHF 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: MISTYP – INTEGER *Input*

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

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

If 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: the number, p, 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*** 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*** array *Output*

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

12: SSP(ISSP,NVARS) – ***real*** 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 $j, k = 1, 2, \ldots, p$.

13: ISSP – INTEGER *Input*

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

Constraint: ISSP $\geq$ NVARS.

14: R(IR,NVARS) – ***real*** 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 $j, k = 1, 2, \ldots, p$.

15: IR – INTEGER *Input*

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

Constraint: IR $\geq$ 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. 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, NVARS < 2,
or NVARS $>$ M.

IFAIL $= 3$

On entry, IX $<$ N,
or ISSP $<$ NVARS,
or IR $<$ NVARS.

IFAIL $= 4$

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

7 Accuracy

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

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 p, and the occurrence of missing values.

The routine uses a two-pass algorithm.

9 Example

The following program 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.

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.

```
*     G02BHF Example Program Text
*     Mark 14 Revised.  NAG Copyright 1989.
*     .. Parameters ..
      INTEGER          M, N, NV, IA, ISSP, ICORR
      PARAMETER        (M=4,N=5,NV=3,IA=N,ISSP=NV,ICORR=NV)
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
*     .. Local Scalars ..
      INTEGER          I, IFAIL, J, MISTYP, NCASES
*     .. Local Arrays ..
      real             A(IA,M), AMEAN(NV), CORR(ICORR,NV), SSP(ISSP,NV),
     +                 STD(NV), XMISS(M)
      INTEGER          KVAR(NV), MISS(M)
*     .. External Subroutines ..
      EXTERNAL         G02BHF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G02BHF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) ((A(I,J),J=1,M),I=1,N)
      KVAR(1) = 4
      KVAR(2) = 1
      KVAR(3) = 2
      MISTYP = 0
      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,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) = 0
      MISS(2) = 1
      MISS(3) = 0
      MISS(4) = 1
      XMISS(2) = 0.0e0
      XMISS(4) = 0.0e0
      IFAIL = 1
*
      CALL G02BHF(N,M,A,IA,MISS,XMISS,MISTYP,NV,KVAR,AMEAN,STD,SSP,ISSP,
     +            CORR,ICORR,NCASES,IFAIL)
*
      IF (IFAIL.NE.0) THEN
         WRITE (NOUT,99999) 'Routine fails, IFAIL =', IFAIL
      ELSE
         WRITE (NOUT,*) 'Variable   Mean    St. dev.'
         WRITE (NOUT,99995) (KVAR(I),AMEAN(I),STD(I),I=1,NV)
         WRITE (NOUT,*)
         WRITE (NOUT,*)
     +     'Sums of squares and cross-products of deviations'
         WRITE (NOUT,99998) (KVAR(I),I=1,NV)
         WRITE (NOUT,99996) (KVAR(I),(SSP(I,J),J=1,NV),I=1,NV)
         WRITE (NOUT,*)
         WRITE (NOUT,*) 'Correlation coefficients'
         WRITE (NOUT,99998) (KVAR(I),I=1,NV)
         WRITE (NOUT,99996) (KVAR(I),(CORR(I,J),J=1,NV),I=1,NV)
```

```
         WRITE (NOUT,*)
         WRITE (NOUT,99999) 'Number of cases actually used:', NCASES
      END IF
      STOP
*
99999 FORMAT (1X,A,I3)
99998 FORMAT (1X,4I12)
99997 FORMAT (1X,I3,4F12.4)
99996 FORMAT (1X,I3,3F12.4)
99995 FORMAT (1X,I5,2F11.4)
      END
```

9.2 Program Data

```
G02BHF Example Program Data
3.00      3.00      1.00      2.00
6.00      4.00      -1.00     4.00
9.00      0.00      5.00      9.00
12.00     2.00      0.00      0.00
-1.00     5.00      4.00      12.00
```

9.3 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
```