NAG Library Routine Document

G02BBF

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

G02BBF computes means and standard deviations of variables, sums of squares and cross-products of deviations from means, and Pearson product-moment correlation coefficients for a set of data omitting completely any cases with a missing observation for any variable.

2 Specification

```
 SUBROUTINE G02BBF(N, M, X, LDX, MISS, XMISS, XBAR, STD, SSP, LDSSP, R,
1                  LDR, NCASES, IFAIL)
 INTEGER           N, M, LDX, MISS(M), LDSSP, LDR, NCASES, IFAIL
 double precision  X(LDX,M), XMISS(M), XBAR(M), STD(M), SSP(LDSSP,M),
1                  R(LDR,M)
```

3 Description

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

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

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) Means:

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

(b) Standard deviations:

$$s_j = \sqrt{\frac{\sum_{i=1}^{n} w_i \left(x_{ij} - \bar{x}_j\right)^2}{\sum_{i=1}^{n} w_i - 1}}, \qquad j = 1,2,\ldots,m.$$

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

$$S_{jk} = \sum_{i=1}^{n} w_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \qquad j, k = 1, 2, \ldots, m.$$

(d) Pearson product-moment correlation coefficients:

$$R_{jk} = \frac{S_{jk}}{\sqrt{S_{jj} S_{kk}}}, \qquad j, k = 1, 2, \ldots, m.$$

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: N ≥ 2.

2: M – INTEGER *Input*

On entry: m, the number of variables.

Constraint: M ≥ 2.

3: X(LDX,M) – ***double precision*** 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 G02BBF is called.

Constraint: LDX $\geq$ 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) – ***double precision*** 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: XBAR(M) – ***double precision*** array *Output*

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

8: STD(M) – ***double precision*** array *Output*

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

9: SSP(LDSSP,M) – ***double precision*** array *Output*

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

10: LDSSP – INTEGER *Input*

On entry: the first dimension of the array SSP as declared in the (sub)program from which G02BBF is called.

Constraint: LDSSP $\geq$ M.

11: R(LDR,M) – ***double precision*** array *Output*

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

12: LDR – INTEGER *Input*

On entry: the first dimension of the array R as declared in the (sub)program from which G02BBF is called.

Constraint: LDR $\geq$ M.

13: NCASES – INTEGER *Output*

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

14: 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.

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, 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.**

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, LDX $<$ N,
or LDSSP $<$ M,
or LDR $<$ M.

IFAIL $= 4$

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

IFAIL $= 5$

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

7 Accuracy

G02BBF 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. G02BBF treats all values in the inclusive range $\left(1 \pm 0.1^{(\text{X02BEF}-2)}\right) \times xmj$, where *xmj* is the missing value for variable j specified by you 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 G02BBF depends on n and m, and the occurrence of missing values.

The routine uses a two-pass algorithm.

9 Example

This example reads in a set of data consisting of five observations on each of three variables. Missing values of 0.0 are declared for the first and third variables; no missing value is specified for the second variable. The means, standard deviations, sums of squares and cross-products of deviations from means, and Pearson product-moment correlation coefficients for all three variables are then calculated and printed, omitting completely all cases containing missing values; cases 3 and 4 are therefore eliminated, leaving only three cases in the calculations.

9.1 Program Text

```
*     G02BBF Example Program Text
*     Mark 14 Revised. NAG Copyright 1989.
*     .. Parameters ..
      INTEGER          M, N, LDX, LDSSP, LDR
      PARAMETER        (M=3,N=5,LDX=N,LDSSP=M,LDR=M)
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
*     .. Local Scalars ..
      INTEGER          I, IFAIL, J, NCASES
*     .. Local Arrays ..
      DOUBLE PRECISION R(LDR,M), SSP(LDSSP,M), STD(M), X(LDX,M),
     +                 XBAR(M), XMISS(M)
      INTEGER          MISS(M)
*     .. External Subroutines ..
      EXTERNAL         G02BBF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G02BBF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) ((X(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,(X(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) = 0.0D0
      XMISS(3) = 0.0D0
      IFAIL = 1
*
      CALL G02BBF(N,M,X,LDX,MISS,XMISS,XBAR,STD,SSP,LDSSP,R,LDR,NCASES,
     +            IFAIL)
*
      IF (IFAIL.LT.0) THEN
         WRITE (NOUT,*)
         WRITE (NOUT,99999) ' ** G02BBF returned with IFAIL = ', IFAIL
      ELSE
         IF (IFAIL.NE.0) THEN
            WRITE (NOUT,99999) 'Routine fails, IFAIL =', IFAIL
         ELSE
            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 of deviations'
            WRITE (NOUT,99998) (I,I=1,M)
            WRITE (NOUT,99997) (I,(SSP(I,J),J=1,M),I=1,M)
            WRITE (NOUT,*)
            WRITE (NOUT,*) 'Correlation coefficients'
            WRITE (NOUT,99998) (I,I=1,M)
            WRITE (NOUT,99997) (I,(R(I,J),J=1,M),I=1,M)
            WRITE (NOUT,*)
            WRITE (NOUT,99999) 'Number of cases actually used: ', NCASES
         END IF
      END IF
*
99999 FORMAT (1X,A,I5)
99998 FORMAT (1X,6I12)
99997 FORMAT (1X,I3,3F12.4)
99996 FORMAT (1X,I5,2F11.4)
      END
```

9.2 Program Data

```
G02BBF Example Program Data
2.00      3.00      3.00
4.00      6.00      4.00
9.00      9.00      0.00
0.00      12.00     2.00
12.00     -1.00     5.00
```

9.3 Program Results

```
 G02BBF 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     6.0000     5.2915
```

```
     2     2.6667     3.5119
     3     4.0000     1.0000

 Sums of squares and cross-products of deviations
           1           2           3
   1     56.0000    -30.0000     10.0000
   2    -30.0000     24.6667     -4.0000
   3     10.0000     -4.0000      2.0000

 Correlation coefficients
           1           2           3
   1      1.0000     -0.8072      0.9449
   2     -0.8072      1.0000     -0.5695
   3      0.9449     -0.5695      1.0000

 Number of cases actually used:     3
```