

G02BEF – NAG Fortran Library Routine Document

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

G02BEF 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 completely any cases with a missing observation for any variable.

2 Specification

```

SUBROUTINE G02BEF(N, M, X, IX, MISS, XMISS, XBAR, STD, SSPZ,
1             ISSPZ, RZ, IRZ, NCASES, IFAIL)
INTEGER      N, M, IX, MISS(M), ISSPZ, IRZ, NCASES, IFAIL
real        X(IX,M), XMISS(M), XBAR(M), STD(M),
1             SSPZ(ISSPZ,M), RZ(IRZ,M)

```

3 Description

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

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

$$j = 1, 2, \dots, m \quad (m \geq 2),$$

where x_{ij} is the i th observation on the j th 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 j th 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 $i = 1, 2, \dots, n$.

The quantities calculated are:

(a) Means:

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

(b) Standard deviations:

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

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

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

(d) Correlation-like coefficients:

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

If \tilde{S}_{jj} or \tilde{S}_{kk} is zero, \tilde{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 i th observation on the j th variable, for $i = 1, 2, \dots, n$; $j = 1, 2, \dots, m$.
- 4:** IX — INTEGER *Input*
On entry: the first dimension of the array X as declared in the (sub)program from which G02BEF 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, x_{mj} , is to be specified for the j th 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 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, x_{mj} , to be associated with the j th 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) — *real* array *Output*
On exit: the mean value, \bar{x}_j , of the j th variable, for $j = 1, 2, \dots, m$.
- 8:** STD(M) — *real* array *Output*
On exit: the standard deviation, s_j , of the j th variable, for $j = 1, 2, \dots, m$.
- 9:** SSPZ(ISSPZ,M) — *real* array *Output*
On exit: $SSPZ(j, k)$ is the cross-product about zero, \tilde{S}_{jk} , for $j, k = 1, 2, \dots, m$.
- 10:** ISSPZ — INTEGER *Input*
On entry: the first dimension of the array SSPZ as declared in the (sub)program from which G02BEF is called.
Constraint: $ISSPZ \geq M$.

- 11:** RZ(IRZ,M) — *real* array *Output*
On exit: RZ(j, k) is the correlation-like coefficient, \tilde{R}_{jk} , between the j th and k th variables, for $j, k = 1, 2, \dots, m$.
- 12:** IRZ — INTEGER *Input*
On entry: the first dimension of the array RZ as declared in the (sub)program from which G02BEF is called.
Constraint: IRZ \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. For users not familiar with this parameter (described in Chapter P01) the recommended value is 0.
On exit: IFAIL = 0 unless the routine detects an error (see Section 6).

6 Error Indicators and Warnings

Errors detected by the routine:

IFAIL = 1

On entry, N < 2.

IFAIL = 2

On entry, M < 2.

IFAIL = 3

On entry, IX < N,
 or ISSPZ < M,
 or IRZ < 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

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, since the routine treats as missing values for variable j , 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). 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.

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 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 about zero, and correlation-like 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

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.

```

*      G02BEF Example Program Text
*      Mark 14 Revised.  NAG Copyright 1989.
*      .. Parameters ..
      INTEGER          M, N, IA, ISSP, ICORR
      PARAMETER        (M=3,N=5,IA=N,ISSP=M,ICORR=M)
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
*      .. Local Scalars ..
      INTEGER          I, IFAIL, J, NCASES
*      .. Local Arrays ..
      real            A(IA,M), AMEAN(M), CORR(ICORR,M), SSP(ISSP,M),
+                   STD(M), XMISS(M)
      INTEGER          MISS(M)
*      .. External Subroutines ..
      EXTERNAL         G02BEF
*      .. Executable Statements ..
      WRITE (NOUT,*) 'G02BEF 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) = 0.0e0
      XMISS(3) = 0.0e0
      IFAIL = 1

*
      CALL G02BEF(N,M,A,IA,MISS,XMISS,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,99996) (I,AMEAN(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,(SSP(I,J),J=1,M),I=1,M)
  WRITE (NOUT,*)
  WRITE (NOUT,*) 'Correlation-like coefficients'
  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,I2)
99998 FORMAT (1X,6I12)
99997 FORMAT (1X,I3,3F12.4)
99996 FORMAT (1X,I5,2F11.4)
END

```

9.2 Program Data

G02BEF 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

G02BEF 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 about zero

	1	2	3
1	164.0000	18.0000	82.0000
2	18.0000	46.0000	28.0000
3	82.0000	28.0000	50.0000

Correlation-like coefficients

	1	2	3
1	1.0000	0.2072	0.9055
2	0.2072	1.0000	0.5838
3	0.9055	0.5838	1.0000

Number of cases actually used: 3
