

## G02BCF – 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

G02BCF 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 cases with missing values from only those calculations involving the variables for which the values are missing.

### 2 Specification

```

SUBROUTINE G02BCF(N, M, X, IX, MISS, XMISS, XBAR, STD, SSP, ISSP,
1          R, IR, NCASES, COUNT, IC, IFAIL)
  INTEGER      N, M, IX, MISS(M), ISSP, IR, NCASES, IC, IFAIL
  real        X(IX,M), XMISS(M), XBAR(M), STD(M), SSP(ISSP,M),
1          R(IR,M), COUNT(IC,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  $w_{ij} = 0$  if the  $i$ th observation for the  $j$ th variable is a missing value, i.e., if a missing value,  $xm_j$ , has been declared for the  $j$ th variable, and  $x_{ij} = xm_j$  (see also Section 7); and  $w_{ij} = 1$  otherwise, for  $i = 1, 2, \dots, n$ ;  $j = 1, 2, \dots, m$ .

The quantities calculated are:

(a) Means:

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

(b) Standard deviations:

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

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

$$S_{jk} = \sum_{i=1}^n w_{ij} w_{ik} (x_{ij} - \bar{x}_{j(k)}) (x_{ik} - \bar{x}_{k(j)}), \quad j, k = 1, 2, \dots, m$$

where

$$\bar{x}_{j(k)} = \frac{\sum_{i=1}^n w_{ij} w_{ik} x_{ij}}{\sum_{i=1}^n w_{ij} w_{ik}} \quad \text{and} \quad \bar{x}_{k(j)} = \frac{\sum_{i=1}^n w_{ik} w_{ij} x_{ik}}{\sum_{i=1}^n w_{ik} w_{ij}}$$

(i.e., the means used in the calculation of the sums of squares and cross-products of deviations are based on the same set of observations as are the cross-products.)

- (d) Pearson product-moment correlation coefficients:

$$R_{jk} = \frac{S_{jk}}{\sqrt{S_{jj(k)}S_{kk(j)}}}, \quad j, k, = 1, 2, \dots, m$$

where  $S_{jj(k)} = \sum_{i=1}^n w_{ij}w_{ik}(x_{ij} - \bar{x}_{j(k)})^2$  and  $S_{kk(j)} = \sum_{i=1}^n w_{ik}w_{ij}(x_{ik} - \bar{x}_{k(j)})^2$  and  $\bar{x}_{j(k)}$  and  $\bar{x}_{k(j)}$  are as defined in (c) above.

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

If  $S_{jj(k)}$  or  $S_{kk(j)}$  is zero,  $R_{jk}$  is set to zero.

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

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

(The diagonal terms,  $c_{jj}$ , for  $j = 1, 2, \dots, m$ , 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 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 G02BCF is called.  
*Constraint:*  $IX \geq N$ .
- 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  $j$ 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.
- 6:** XMISS(M) — *real* array *Input*  
*On entry:* XMISS( $j$ ) must be set to the missing value,  $xm_j$ , 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).

- 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:** SSP(ISSP,M) — *real* array Output  
*On exit:* SSP( $j, k$ ) is the cross-product of deviations,  $S_{jk}$ , for  $j, k = 1, 2, \dots, m$ .
- 10:** ISSP — INTEGER Input  
*On entry:* the first dimension of the array SSP as declared in the (sub)program from which G02BCF is called.  
*Constraint:* ISSP  $\geq$  M.
- 11:** R(IR,M) — *real* array Output  
*On exit:* R( $j, k$ ) is the product-moment correlation coefficient,  $R_{jk}$ , between the  $j$ th and  $k$ th variables, for  $j, k = 1, 2, \dots, m$ .
- 12:** IR — INTEGER Output  
*On entry:* the first dimension of the array R as declared in the (sub)program from which G02BCF is called.  
*Constraint:* IR  $\geq$  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 coefficients (when cases involving missing values have been eliminated).
- 14:** COUNT(IC,M) — *real* array Output  
*On exit:* COUNT( $j, k$ ) is the number of cases,  $c_{jk}$ , actually used in the calculation of  $S_{jk}$ , and  $R_{jk}$ , the sum of cross-products and correlation coefficient for the  $j$ th and  $k$ th variables, for  $j, k = 1, 2, \dots, m$ .
- 15:** IC — INTEGER Input  
*On entry:* the first dimension of the array COUNT as declared in the (sub)program from which G02BCF is called.  
*Constraint:* IC  $\geq$  M.
- 16:** 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 or gives a warning (see Section 6).  
**For this routine**, because the values of output parameters may be useful even if IFAIL  $\neq$  0 on exit, users are recommended to set IFAIL to -1 before entry. **It is then essential to test the value of IFAIL on exit.** To suppress the output of an error message when soft failure occurs, set IFAIL to 1.

## 6 Error Indicators and Warnings

Errors or warnings specified by the routine:

IFAIL = 1

On entry,  $N < 2$ .

IFAIL = 2

On entry,  $M < 2$ .

IFAIL = 3

On entry,  $IX < N$ ,  
or  $ISSP < M$ ,  
or  $IR < M$ ,  
or  $IC < M$ .

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

## 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  $(1 \pm ACC) \times xm_j$ , 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,  $-1.0$  and 0.0 are declared for the first, second and third variables respectively. The means, standard deviations, sums of squares and cross-products of deviations from means, and Pearson product-moment correlation 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.

## 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.

```

*      G02BCF Example Program Text
*      Mark 14 Revised.  NAG Copyright 1989.
*      .. Parameters ..
INTEGER          M, N, IA, ISSP, ICORR, IC
PARAMETER       (M=3,N=5,IA=N,ISSP=M,ICORR=M,IC=M)
INTEGER          NIN, NOUT
PARAMETER       (NIN=5,NOUT=6)
*      .. Local Scalars ..
INTEGER          I, IFAIL, J, NCASES
*      .. Local Arrays ..
real           A(IA,M), AMEAN(M), CASES(IC,M), CORR(ICORR,M),
+               SSP(ISSP,M), STD(M), XMISS(M)
INTEGER          MISS(M)
*      .. External Subroutines ..
EXTERNAL        G02BCF
*      .. Executable Statements ..
WRITE (NOUT,*) 'G02BCF 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) = 1
MISS(3) = 1
XMISS(1) = 0.0e0
XMISS(2) = -1.0e0
XMISS(3) = 0.0e0
IFAIL = 1
*
CALL G02BCF(N,M,A,IA,MISS,XMISS,AMEAN,STD,SSP,ISSP,CORR,ICORR,
+          NCASES,CASES,IC,IFAIL)
*
IF (IFAIL.NE.0) THEN
  WRITE (NOUT,99996) 'Routine fails, IFAIL =', IFAIL
ELSE
  WRITE (NOUT,*) 'Variable   Mean   St. dev.'
  WRITE (NOUT,99995) (I,AMEAN(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,(CORR(I,J),J=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,(CASES(I,J),J=1,M),I=1,M)
END IF
STOP

*
99999 FORMAT (1X,A,I2)
99998 FORMAT (1X,6I12)
99997 FORMAT (1X,I3,3F12.4)
99996 FORMAT (1X,A,I2)
99995 FORMAT (1X,I5,2F11.4)
END

```

## 9.2 Program Data

G02BCF 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

G02BCF 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 of deviations

	1	2	3
1	62.7500	21.0000	10.0000
2	21.0000	45.0000	-6.0000
3	10.0000	-6.0000	5.0000

## Correlation coefficients

	1	2	3
1	1.0000	0.9707	0.9449
2	0.9707	1.0000	-0.6547
3	0.9449	-0.6547	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

---