NAG Fortran Library Routine Document

G11BBF

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

G11BBF computes a table from a set of classification factors using a given percentile or quantile, for example the median.

2 Specification

```
 SUBROUTINE G11BBF(TYPE, WEIGHT, N, NFAC, ISF, LFAC, IFAC, LDF, PERCNT,
1                  Y, WT, TABLE, MAXT, NCELLS, NDIM, IDIM, ICOUNT, IWK,
2                  WK, IFAIL)
 INTEGER           N, NFAC, ISF(NFAC), LFAC(NFAC), IFAC(LDF,NFAC), LDF,
1                  MAXT, NCELLS, NDIM, IDIM(NFAC), ICOUNT(MAXT),
2                  IWK(2*NFAC+N), IFAIL
 real              PERCNT, Y(N), WT(*), TABLE(MAXT), WK(2*N)
 CHARACTER*1       TYPE, WEIGHT
```

3 Description

A data set may include both classification variables and general variables. The classification variables, known as factors, take a small number of values known as levels. For example, the factor sex would have the levels male and female. These can be coded as 1 and 2 respectively. Given several factors, a multi-way table can be constructed such that each cell of the table represents one level from each factor. For example, the two factors sex and habitat, habitat having three levels (inner-city, suburban and rural) define the 2×3 contingency table

Sex	**Habitat**		
	Inner-city	Suburban	Rural
Male			
Female			

For each cell statistics can be computed. If a third variable in the data set was age then for each cell the median age could be computed:

Sex	**Habitat**		
	Inner-city	Suburban	Rural
Male	24	31	37
Female	21.5	28.5	33

That is, the median age for all observations for males living in rural areas is 37, the median being the 50% quantile. Other quantiles can also be computed: the p percent quantile or percentile, q_p, is the estimate of the value such that p percent of observations are less than q_p. This is calculated in two different ways depending on whether the tabulated variable is continuous or discrete. Let there be m values in a cell and let $y_{(1)}$, $y_{(2)}, \ldots, y_{(m)}$ be the values for that cell sorted into ascending order. Also, associated with each value there is a weight, $w_{(1)}$, $w_{(2)}, \ldots, w_{(m)}$, which could represent the observed frequency for that value,

with $W_j = \sum_{i=1}^{j} w_{(i)}$ and $W'_j = \sum_{i=1}^{j} w_{(i)} - \frac{1}{2}w_{(j)}$. For the p percentile let $p_w = (p/100)W_m$ and $p'_w = (p/100)W'_m$, then the percentiles for the two cases are as given below.

If the variable is discrete, that is, it takes only a limited number of (usually integer) values, then the percentile is defined as

$$\begin{cases} y_{(j)} & \text{if } W_{j-1} < p_W < W_j \\ \dfrac{y_{(j+1)} + y_{(j)}}{2} & \text{if } p_w = W_j. \end{cases}$$

If the data is continuous then the quantiles are estimated by linear interpolation.

$$\begin{cases} y_{(1)} & \text{if } p'_w \leq W'_1 \\ (1-f)y_{(j-1)} + fy_{(j)} & \text{if } W'_{j-1} < p'_w \leq W'_j \\ y_{(m)} & \text{if } p'_w > W'_m, \end{cases}$$

where $f = (p'_w - W'_{j-1})/(W'_j - W'_{j-1})$.

4 References

John J A and Quenouille M H (1977) *Experiments: Design and Analysis* Griffin

Kendall M G and Stuart A (1969) *The Advanced Theory of Statistics (Volume 1)* (3rd Edition) Griffin

5 Parameters

1: TYPE – CHARACTER*1 *Input*

On entry: indicates if the variable to be tabulated is discrete or continuous.

If TYPE = 'D', the percentiles are computed for a discrete variable.

If TYPE = 'C', the percentiles are computed for a continuous variable using linear interpolation.

Constraint: TYPE = 'D' or 'C'.

2: WEIGHT – CHARACTER*1 *Input*

On entry: indicates if there are weights associated with the variable to be tabulated.

If WEIGHT = 'U', weights are not input and unit weights are assumed.

If WEIGHT = 'W', weights must be supplied in WT.

Constraint: WEIGHT = 'U' or 'W'.

3: N – INTEGER *Input*

On entry: the number of observations.

Constraint: N $\geq$ 2.

4: NFAC – INTEGER *Input*

On entry: the number of classifying factors in IFAC.

Constraint: NFAC $\geq$ 1.

5: ISF(NFAC) – INTEGER array *Input*

On entry: indicates which factors in IFAC are to be used in the tabulation.

If ISF(i) $>$ 0, the ith factor in IFAC is included in the tabulation.

Note that if ISF(i) ≤ 0, for $i = 1, 2, \ldots,$ NFAC then the statistic for the whole sample is calculated and returned in a 1×1 table.

6: LFAC(NFAC) – INTEGER array *Input*

On entry: the number of levels of the classifying factors in IFAC.

Constraint: if ISF(i) > 0, LFAC(i) ≥ 2, for $i = 1, 2, \ldots,$ NFAC.

7: IFAC(LDF,NFAC) – INTEGER array *Input*

On entry: the NFAC coded classification factors for the N observations.

Constraint: $1 \leq$ IFAC(i, j) $\leq$ LFAC(j) for $i = 1, 2, \ldots,$ N; $j = 1, 2, \ldots,$ NFAC.

8: LDF – INTEGER *Input*

On entry: the first dimension of the array IFAC as declared in the (sub)program from which G11BBF is called.

Constraint: LDF $\geq$ N.

9: PERCNT – ***real*** *Input*

On entry: the percentile to be tabulated, p.

Constraint: $0.0 < p < 100.0$.

10: Y(N) – ***real*** array *Input*

On entry: the variable to be tabulated.

11: WT(*) – ***real*** array *Input*

Note: the dimension of the array WT must be at least N if WEIGHT = 'W' and 1 otherwise.

On entry: if WEIGHT = 'W', WT must contain the N weights. Otherwise WT is not referenced.

Constraint: if WEIGHT = 'W', WT(i) ≥ 0.0, for $i = 1, 2, \ldots,$ N.

12: TABLE(MAXT) – ***real*** array *Output*

On exit: the computed table. The NCELLS cells of the table are stored so that for any two factors the index relating to the factor occurring later in LFAC and IFAC changes faster. For further details see Section 8.

13: MAXT – INTEGER *Input*

On entry: the maximum size of the table to be computed.

Constraint: MAXT $\geq$ product of the levels of the factors included in the tabulation.

14: NCELLS – INTEGER *Output*

On exit: the number of cells in the table.

15: NDIM – INTEGER *Output*

On exit: the number of factors defining the table.

16: IDIM(NFAC) – INTEGER array *Output*

On exit: the first NDIM elements contain the number of levels for the factors defining the table.

17: ICOUNT(MAXT) – INTEGER array *Output*

On exit: a table containing the number of observations contributing to each cell of the table, stored identically to TABLE.

18: IWK(2*NFAC+N) – INTEGER array *Workspace*
19: WK(2*N) – ***real*** array *Workspace*

20: 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,
or NFAC < 1,
or LDF $<$ N,
or TYPE $\neq$ 'D' or 'C',
or WEIGHT $\neq$ 'U' or 'W',
or PERCNT ≤ 0.0,
or PERCNT ≥ 100.0.

IFAIL $= 2$

On entry, ISF$(i) > 0$ and LFAC$(i) \leq 1$, for some i,
or IFAC$(i,j) < 1$, for some i,j,
or IFAC$(i,j) >$ LFAC(j), for some i,j,
or MAXT is too small,
or WEIGHT $=$ 'W' and WT$(i) < 0.0$, for some i.

IFAIL $= 3$

At least one cell is empty.

7 Accuracy

Not applicable.

8 Further Comments

The tables created by G11BBF and stored in TABLE and ICOUNT are stored in the following way. Let there be n factors defining the table with factor k having l_k levels, then the cell defined by the levels i_1, i_2,...,i_n of the factors is stored in the mth cell given by:

$$m = 1 + \sum_{k=1}^{n} [(i_k - 1)c_k],$$

where $c_j = \prod_{k=j+1}^{n} l_k$, for $j = 1, 2, \ldots, n-1$ and $c_n = 1$.

9 Example

The data, given by John and Quenouille (1977), is for a 3×6 factorial experiment in 3 blocks of 18 units. The data is input in the order, blocks, factor with 3 levels, factor with 6 levels, yield, and the 3×6 table of treatment medians for yield over blocks is computed and printed.

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.

```
*     G11BBF Example Program Text
*     Mark 17 Release. NAG Copyright 1995.
*     .. Parameters ..
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
      INTEGER          NMAX, MMAX, LTMAX
      PARAMETER        (NMAX=54,MMAX=3,LTMAX=18)
*     .. Local Scalars ..
      real             PERCNT
      INTEGER          I, IFAIL, J, K, LDF, MAXT, N, NCELLS, NCOL, NDIM,
     +                 NFAC, NROW
      CHARACTER        TYPE, WEIGHT
*     .. Local Arrays ..
      real             TABLE(LTMAX), WK(2*NMAX), WT(NMAX), Y(NMAX)
      INTEGER          ICOUNT(LTMAX), IDIM(MMAX), IFAC(NMAX,MMAX),
     +                 ISF(MMAX), IWK(2*MMAX+NMAX), LFAC(MMAX)
*     .. External Subroutines ..
      EXTERNAL         G11BBF
*     .. Executable Statements ..
      WRITE (NOUT,*) 'G11BBF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) TYPE, WEIGHT, N, NFAC, PERCNT
      IF (N.LE.NMAX .AND. NFAC.LE.MMAX) THEN
         IF (WEIGHT.EQ.'W' .OR. WEIGHT.EQ.'w' .OR. WEIGHT.EQ.'V' .OR.
     +       WEIGHT.EQ.'v') THEN
            DO 20 I = 1, N
               READ (NIN,*) (IFAC(I,J),J=1,NFAC), Y(I), WT(I)
   20       CONTINUE
         ELSE
            DO 40 I = 1, N
               READ (NIN,*) (IFAC(I,J),J=1,NFAC), Y(I)
   40       CONTINUE
         END IF
         READ (NIN,*) (LFAC(J),J=1,NFAC)
         READ (NIN,*) (ISF(J),J=1,NFAC)
         LDF = NMAX
         MAXT = LTMAX
         IFAIL = 0
*
         CALL G11BBF(TYPE,WEIGHT,N,NFAC,ISF,LFAC,IFAC,LDF,PERCNT,Y,WT,
     +              TABLE,MAXT,NCELLS,NDIM,IDIM,ICOUNT,IWK,WK,IFAIL)
*
         WRITE (NOUT,*)
         WRITE (NOUT,99999) ' TABLE for ', PERCNT, 'th percentile'
         WRITE (NOUT,*)
         NCOL = IDIM(NDIM)
         NROW = NCELLS/NCOL
         K = 1
         DO 60 I = 1, NROW
            WRITE (NOUT,99998) (TABLE(J),'(',ICOUNT(J),')',J=K,K+NCOL-1)
            K = K + NCOL
   60    CONTINUE
      END IF
      STOP
*
99999 FORMAT (A,F4.0,A)
99998 FORMAT (1X,6(F8.2,A,I2,A))
```

```
      END
```

9.2 Program Data

```
G11BBF Example Program Data

'C' 'U' 54 3 50.0

1 1 1 274
1 2 1 361
1 3 1 253
1 1 2 325
1 2 2 317
1 3 2 339
1 1 3 326
1 2 3 402
1 3 3 336
1 1 4 379
1 2 4 345
1 3 4 361
1 1 5 352
1 2 5 334
1 3 5 318
1 1 6 339
1 2 6 393
1 3 6 358
2 1 1 350
2 2 1 340
2 3 1 203
2 1 2 397
2 2 2 356
2 3 2 298
2 1 3 382
2 2 3 376
2 3 3 355
2 1 4 418
2 2 4 387
2 3 4 379
2 1 5 432
2 2 5 339
2 3 5 293
2 1 6 322
2 2 6 417
2 3 6 342
3 1 1  82
3 2 1 297
3 3 1 133
3 1 2 306
3 2 2 352
3 3 2 361
3 1 3 220
3 2 3 333
3 3 3 270
3 1 4 388
3 2 4 379
3 3 4 274
3 1 5 336
3 2 5 307
3 3 5 266
3 1 6 389
3 2 6 333
3 3 6 353

3 3 6
0 1 1
```

9.3 Program Results

```
 G11BBF Example Program Results

 TABLE for  50.th percentile

   226.00( 3)  320.25( 3)  299.50( 3)  385.75( 3)  348.00( 3)  334.75( 3)
   329.25( 3)  343.25( 3)  365.25( 3)  370.50( 3)  327.25( 3)  378.00( 3)
   185.50( 3)  328.75( 3)  319.50( 3)  339.25( 3)  286.25( 3)  350.25( 3)
```