NAG Fortran Library Routine Document

G01AFF

1   Purpose

G01AFF performs the analysis of a two-way $r \times c$ contingency table or classification. If $r = c = 2$, and the total number of objects classified is 40 or fewer, then the probabilities for Fisher's exact test are computed. Otherwise, a test statistic is computed (with Yates' correction when $r = c = 2$), which under the assumption of no association between the classifications has approximately a chi-square distribution with $(r - 1) \times (c - 1)$ degrees of freedom.

2   Specification

```
SUBROUTINE G01AFF(INOB, IPRED, M, N, NOBS, NUM, PRED, CHIS, P, NPOS,
1                 NDF, M1, N1, IFAIL)
INTEGER          INOB, IPRED, M, N, NOBS(INOB,N), NUM, NPOS, NDF, M1,
1                N1, IFAIL
real             PRED(IPRED,N), CHIS, P(21)
```

3   Description

The data consist of the frequencies for the two-way classification, denoted by n_{ij}, for $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$ with $m, n > 1$.

A check is made to see whether any row or column of the matrix of frequencies consists entirely of zeros, and if so, the matrix of frequencies is reduced by omitting that row or column. Suppose the final size of the matrix is m_1 by n_1 $(m_1, n_1 > 1)$, and let

$R_i = \sum_{j=1}^{n_1} n_{ij}$, the total frequency for the ith row, for $i = 1, 2, \ldots, m_1$,

$C_j = \sum_{i=1}^{m_1} n_{ij}$, the total frequency for the jth column, for $j = 1, 2, \ldots, n_1$, and

$T = \sum_{i=1}^{m_1} R_i = \sum_{j=1}^{n_1} C_j$, the total frequency.

There are two situations:

(i)   If $m_1 > 2$ and/or $n_1 > 2$, or $m_1 = n_1 = 2$ and $T > 40$, then the matrix of expected frequencies, denoted by r_{ij}, for $i = 1, 2, \ldots, m_1$; $j = 1, 2, \ldots, n_1$, and the test statistic, χ^2, are computed, where

$$r_{ij} = R_i C_j / T, \quad i = 1, 2, \ldots, m_1; \ j = 1, 2, \ldots, n_1$$

and

$$\chi^2 = \sum_{i=1}^{m_1} \sum_{j=1}^{n_1} [|r_{ij} - n_{ij}| - Y]^2 / r_{ij},$$

where

$$Y = \begin{cases} \frac{1}{2} & \text{if } m_1 = n_1 = 2 \\ 0 & \text{otherwise} \end{cases}$$

is Yates' correction for continuity.

Under the assumption that there is no association between the two classifications, X^2 will have approximately a chi-square distribution with $(m_1 - 1) \times (n_1 - 1)$ degrees of freedom.

An option exists which allows for further 'shrinkage' of the matrix of frequencies in the case where $r_{ij} < 1$ for the (i,j)th cell. If this is the case, then row i or column j will be combined with the adjacent row or column with smaller total. Row i is selected for combination if $R_i \times m_1 \leq C_j \times n_1$. This 'shrinking' process is continued until $r_{ij} \geq 1$ for all cells (i,j).

(ii) If $m_1 = n_1 = 2$ and $T \leq 40$, the probabilities to enable Fisher's exact test to be made are computed.

The matrix of frequencies may be rearranged so that R_1 is the smallest marginal (i.e., column and row) total, and $C_2 \geq C_1$. Under the assumption of no association between the classifications, the probability of obtaining r entries in cell (1,1) is computed where

$$P_{r+1} = \frac{R_1! \, R_2! \, C_1! \, C_2!}{T! \, r! \, (R_1 - r)! \, (C_1 - r)! \, (T - C_1 - R_1 + r)!} \;, \quad r = 0, 1, \ldots, R_1.$$

The probability of obtaining the table of given frequencies is returned. A test of the assumption against some alternative may then be made by summing the relevant values of P_r.

4 References

None.

5 Parameters

1: INOB – INTEGER *Input*

On entry: the first dimension of the array NOBS as declared in the (sub)program from which G01AFF is called.

Constraint: INOB $\geq$ M.

2: IPRED – INTEGER *Input*

On entry: the first dimension of the array PRED as declared in the (sub)program from which G01AFF is called.

Constraint: IPRED $\geq$ M.

3: M – INTEGER *Input*

On entry: **one more** than the number of rows of the frequency matrix, NOBS, $m + 1$.

4: N – INTEGER *Input*

On entry: **one more** than the number of columns of the frequency matrix, NOBS, $n + 1$.

5: NOBS(INOB,N) – INTEGER array *Input/Output*

On entry: the elements NOBS(i,j), for $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$, must contain the frequencies for the two-way classification.

On exit: NOBS contains the following information:

NOBS(i,j), for $i = 1, 2, \ldots, m_1$; $j = 1, 2, \ldots, n_1$, contain the frequencies for the two-way classification after 'shrinkage' has taken place (see Section 3).

NOBS$(i, n + 1)$, for $i = 1, 2, \ldots, m_1$, contain the total frequencies in the remaining rows, R_i.

NOBS$(m + 1, j)$, for $j = 1, 2, \ldots, n_1$, contain the total frequencies in the remaining columns, C_j.

NOBS$(m + 1, n + 1)$, contains the total frequency, T.

If any 'shrinkage' has occurred, then all other cells contain no useful information.

6:     NUM – INTEGER   *Input/Output*

On entry: the value assigned to NUM must determine whether automatic 'shrinkage' is required when any $r_{ij} < 1$, as outlined in Section 3(a).

If NUM = 0, shrinkage is not required, and if NUM = 1, shrinkage is required.

On exit: NUM contains the number of elements used in the array P below. When Fisher's exact test for a 2×2 classification is used.

If P is not used, that is if Fisher's exact test is not to be used, then NUM is set to zero.

7:     PRED(IPRED,N) – ***real*** array   *Output*

On exit: the elements $PRED(i, j)$, where $i = 1, 2, \ldots,$M1 and $j = 1, 2, \ldots,$N1 contain the expected frequencies, r_{ij} corresponding to the observed frequencies $NOBS(i, j)$, except in the case when Fisher's exact test for a 2×2 classification is to be used, when PRED is not used. No other elements are utilised.

8:     CHIS – ***real***   *Output*

On exit: the value of the test statistic, χ^2, except when Fisher's exact test for a 2×2 classification is used in which case it is unspecified.

9:     P(21) – ***real*** array   *Output*

P is used only when Fisher's exact test for a 2×2 classification is to be used.

On exit: the first NUM elements contain the probabilities associated with the various possible frequency tables, P_r, for $r = 0, 1, \ldots, R_1$, the remainder are unspecified.

10:    NPOS – INTEGER   *Output*

NPOS is used only when Fisher's exact test for a 2×2 classification is to be used.

On exit: the value of NPOS is the element of P which contains the probability associated with the given table of frequencies.

11:    NDF – INTEGER   *Output*

On exit: the value of NDF gives the number of degrees of freedom for the chi-square distribution, $(m_1 - 1) \times (n_1 - 1)$; when Fisher's exact test is used NDF = 1.

12:    M1 – INTEGER   *Output*

On exit: the number of rows of the two-way classification, after any 'shrinkage', m_1.

13:    N1 – INTEGER   *Output*

On exit: the number of columns of the two-way classification, after any 'shrinkage', n_1.

14:    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

> The number of rows or columns of NOBS is less than 2, possibly after shrinkage.

IFAIL = 2

> At least one frequency is negative, or all frequencies are zero.

IFAIL = 4

> On entry, IPRED < M,
> or   INOB < M.

7   Accuracy

The method used is believed to be stable.

8   Further Comments

The time taken by the routine will increase with M and N, except when Fisher's exact test is to be used, in which case it increases with size of the marginal and total frequencies.

If, on exit, NUM > 0, or alternatively NDF is 1 and $NOBS(M, N) \leq 40$, the probabilities for use in Fisher's exact test for a 2×2 classification will be calculated, and not the test statistic with approximately a chi-square distribution.

9   Example

In the example program, NPROB determines the number of 2-way classifications to be analysed. For each classification the frequencies are read, the routine called, and information given on how much 'shrinkage' has taken place. If Fisher's exact test is to be used, the given frequencies and the array of probabilities associated with the possible frequency tables are printed. Otherwise, if the chi-square test is to be used, the given and expected frequencies, and the test statistic with its degrees of freedom are printed. In the example, there is one 2×3 classification, with shrinkage not requested.

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.

```
*     G01AFF Example Program Text
*     Mark 14 Revised.  NAG Copyright 1989.
*     .. Parameters ..
      INTEGER          NMAX, MMAX, IA, IO
      PARAMETER        (NMAX=8,MMAX=8,IA=MMAX,IO=NMAX)
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
*     .. Local Scalars ..
      real             CHI
      INTEGER          I, IFAIL, J, K, M, M1, M2, N, N1, N2, NDF, NPOS,
     +                 NPROB, NUM
*     .. Local Arrays ..
      real             O(MMAX,NMAX), P(21)
      INTEGER          NA(MMAX,NMAX)
*     .. External Subroutines ..
      EXTERNAL         G01AFF
*     .. Executable Statements ..
```

```
      WRITE (NOUT,*) 'G01AFF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) NPROB
      DO 80 I = 1, NPROB
         READ (NIN,*) M, N, NUM
         IF (M.GE.1 .AND. M.LE.MMAX-1 .AND. N.GE.1 .AND. N.LE.NMAX-1)
    +        THEN
            WRITE (NOUT,*)
            WRITE (NOUT,99999) 'Problem ', I
            DO 20 J = 1, M
               READ (NIN,*) (NA(J,K),K=1,N)
   20       CONTINUE
            WRITE (NOUT,*) 'Data as input -'
            WRITE (NOUT,99991) 'Number of rows', M
            WRITE (NOUT,99991) 'Number of columns', N
            WRITE (NOUT,99991) 'NUM =', NUM,
    +          ' (NUM = 1 means table reduced in size if necessary)'
            IFAIL = 1
            M = M + 1
            N = N + 1
*
            CALL G01AFF(IA,IO,M,N,NA,NUM,O,CHI,P,NPOS,NDF,M1,N1,IFAIL)
*
            WRITE (NOUT,*)
            IF (IFAIL.NE.0) THEN
               WRITE (NOUT,99999)
    +            'Unsuccessful call of G01AFF. IFAIL = ', IFAIL
            ELSE
               IF (NUM.EQ.0) THEN
                  WRITE (NOUT,*) 'Successful call of G01AFF'
                  M2 = M - 1
                  N2 = N - 1
                  IF (M1.NE.M2) THEN
                     WRITE (NOUT,99991) 'No. of rows reduced from ', M2,
    +                   ' to ', M1
                  END IF
                  IF (N1.NE.N2) THEN
                     WRITE (NOUT,99991) 'No. of cols reduced from ', N2,
    +                   ' to ', N1
                  END IF
                  WRITE (NOUT,*)
                  WRITE (NOUT,*)
    +               'Table of observed and expected frequencies'
                  WRITE (NOUT,*)
                  WRITE (NOUT,*) '               Column'
                  WRITE (NOUT,99990) (K,K=1,N1)
                  WRITE (NOUT,*) 'Row'
                  DO 40 J = 1, M1
                     WRITE (NOUT,99998) J, (NA(J,K),K=1,N1)
                     WRITE (NOUT,99997) (O(J,K),K=1,N1)
                     WRITE (NOUT,99993) 'Row total = ', NA(J,N)
   40             CONTINUE
                  WRITE (NOUT,*)
                  WRITE (NOUT,*) 'Column'
                  WRITE (NOUT,99992) 'totals', (NA(M,K),K=1,N1)
                  WRITE (NOUT,99993) 'Grand total = ', NA(M,N)
                  WRITE (NOUT,*)
                  WRITE (NOUT,99996) 'Chi-squared = ', CHI,
    +                '   D.F. = ', NDF
               ELSE
                  WRITE (NOUT,*)
                  WRITE (NOUT,*) 'Fisher''s exact test for 2*2 table'
                  WRITE (NOUT,*)
                  WRITE (NOUT,*) 'Table of observed frequencies'
                  WRITE (NOUT,*)
                  WRITE (NOUT,*) '            Column'
                  WRITE (NOUT,*) '            1    2'
                  WRITE (NOUT,*) 'Row'
                  DO 60 J = 1, 2
                     WRITE (NOUT,99998) J, (NA(J,K),K=1,2)
```

```
                    WRITE (NOUT,99993) 'Row total = ', NA(J,N)
  60                CONTINUE
                    WRITE (NOUT,*)
                    WRITE (NOUT,*) 'Column'
                    WRITE (NOUT,99992) 'totals', (NA(M,K),K=1,2)
                    WRITE (NOUT,99993) 'Grand total = ', NA(M,N)
                    WRITE (NOUT,*)
                    WRITE (NOUT,99995)
     +                 'This table corresponds to element ', NPOS,
     +                 ' in vector P below'
                    WRITE (NOUT,*)
                    WRITE (NOUT,*) 'Vector P'
                    WRITE (NOUT,*)
                    WRITE (NOUT,*) ' I   P(I)'
                    WRITE (NOUT,99994) (J,P(J),J=1,NUM)
                 END IF
              END IF
           ELSE
              STOP
           END IF
  80 CONTINUE
     STOP
*
99999 FORMAT (1X,A,I4)
99998 FORMAT (1X,I2,4X,10I6)
99997 FORMAT (8X,12F6.0)
99996 FORMAT (1X,A,F10.3,A,I3)
99995 FORMAT (1X,A,I4,A)
99994 FORMAT (1X,I2,F9.4)
99993 FORMAT (49X,A,I7)
99992 FORMAT (1X,A,10I6)
99991 FORMAT (1X,A,I3,A,I3)
99990 FORMAT (7X,10I6)
     END
```

9.2   Program Data

```
G01AFF Example Program Data
  1
  2   3   0
   86    51    13
  130   115    41
```

9.3   Program Results

```
G01AFF Example Program Results

Problem    1
Data as input -
Number of rows  2
Number of columns  3
NUM =  0 (NUM = 1 means table reduced in size if necessary)

Successful call of G01AFF

Table of observed and expected frequencies

            Column
           1    2    3
Row
 1        86   51   13
          74.  57.  19.
                                      Row total =     150
 2       130  115   41
         142. 109.  35.
                                      Row total =     286

Column
totals   216  166   54
                                      Grand total =    436

Chi-squared =      6.352   D.F. =   2
```