NAG Library Routine Document

G01AFF

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

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 (LDNOB, LDPRED, M, N, NOBS, NUM, PRED, CHIS, P, NPOS,     &
                   NDF, M1, N1, IFAIL)
INTEGER            LDNOB, LDPRED, M, N, NOBS(LDNOB,N), NUM, NPOS, NDF,     &
                   M1, N1, IFAIL
REAL (KIND=nag_wp) PRED(LDPRED,N), CHIS, P(21)
```

3 Description

The data consist of the frequencies for the two-way classification, denoted by n_{ij}, for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$ 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 $i = 1, 2, \ldots, m_1$ and $j = 1, 2, \ldots, n_1$, 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

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

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, χ^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: LDNOB – INTEGER *Input*

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

Constraint: LDNOB $\geq$ M.

2: LDPRED – INTEGER *Input*

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

Constraint: LDPRED $\geq$ M.

3: M – INTEGER *Input*

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

Constraint: M > 2.

4: N – INTEGER *Input*

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

Constraint: N > 2.

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

On entry: the elements NOBS(i, j), for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$, must contain the frequencies for the two-way classification. The $(m + 1)$th row and the $(n + 1)$th column of NOBS need not be set.

On exit: contains the following information:

NOBS(i, j), for $i = 1, 2, \ldots, m_1$ and $j = 1, 2, \ldots, n_1$, 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.

Constraint: NOBS$(i,j) \geq 0$, for $i = 1,2,\ldots,\mathrm{M}-1$ and $j = 1,2,\ldots,\mathrm{N}-1$.

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(i).

If NUM $= 1$, shrinkage is required, otherwise shrinkage is not required.

On exit: when Fisher's exact test for a 2×2 classification is used then NUM contains the number of elements used in the array P, otherwise NUM is set to zero.

7: PRED(LDPRED, N) – REAL (KIND=nag_wp) array *Output*

On exit: the elements PRED(i,j), where $i = 1,2,\ldots,\mathrm{M1}$ and $j = 1,2,\ldots,\mathrm{N1}$ 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 utilized.

8: CHIS – REAL (KIND=nag_wp) *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 (KIND=nag_wp) 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: P(NPOS) holds 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. If you are unfamiliar with this parameter you should refer to Section 3.3 in the Essential Introduction for details.

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.**

On exit: IFAIL $= 0$ unless the routine detects an error or a warning has been flagged (see Section 6).

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, LDPRED $<$ M,
or LDNOB $<$ M.

IFAIL $= -99$

An unexpected error has been triggered by this routine. Please contact NAG.

See Section 3.8 in the Essential Introduction for further information.

IFAIL $= -399$

Your licence key may have expired or may not have been installed correctly.

See Section 3.7 in the Essential Introduction for further information.

IFAIL $= -999$

Dynamic memory allocation failed.

See Section 3.6 in the Essential Introduction for further information.

7 Accuracy

The method used is believed to be stable.

8 Parallelism and Performance

Not applicable.

9 Further Comments

The time taken by G01AFF 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$(\mathrm{M}, \mathrm{N}) \leq 40$, 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.

10 Example

In the example program, NPROB determines the number of two-way classifications to be analysed. For each classification the frequencies are read, G01AFF 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.

10.1 Program Text

```
    Program g01affe

!     G01AFF Example Program Text

!     Mark 25 Release. NAG Copyright 2014.

!     .. Use Statements ..
      Use nag_library, Only: g01aff, nag_wp
!     .. Implicit None Statement ..
      Implicit None
!     .. Parameters ..
      Integer, Parameter               :: nin = 5, nout = 6
!     .. Local Scalars ..
      Real (Kind=nag_wp)               :: chis
      Integer                          :: ifail, im, in, j, k, ldnob, ldpred,  &
                                          m, m1, m2, n, n1, n2, ndf, npos, num
!     .. Local Arrays ..
      Real (Kind=nag_wp)               :: p(21)
      Real (Kind=nag_wp), Allocatable  :: pred(:,:)
      Integer, Allocatable             :: nobs(:,:)
!     .. Executable Statements ..
      Write (nout,*) 'G01AFF Example Program Results'
      Write (nout,*)

!     Skip heading in data file
      Read (nin,*)

!     Read in the problem size (where N and M are the number of
!     rows and columns in the two way table NOBS)
      Read (nin,*) im, in, num

!     M and N as supplied to G01AFF must be 1 more than the number
!     of rows and columns of data in NOBS
      m = im + 1
      n = in + 1

      ldnob = m
      ldpred = m
      Allocate (nobs(ldnob,n),pred(ldpred,n))

!     Read in data
      Read (nin,*)(nobs(j,1:in),j=1,im)

      Write (nout,*) 'Data as input -'
      Write (nout,99992) 'Number of rows', im
      Write (nout,99992) 'Number of columns', in
      Write (nout,99992) 'NUM =', num, &
        ' (NUM = 1 means table reduced in size if necessary)'

!     Perform the analysis
      ifail = 0
      Call g01aff(ldnob,ldpred,m,n,nobs,num,pred,chis,p,npos,ndf,m1,n1,ifail)

!     Display results
      If (num==0) Then
        m2 = m - 1
        n2 = n - 1
        If (m1/=m2) Then
```

```
          Write (nout,99992) 'No. of rows reduced from ', m2, ' to ', m1
        End If
        If (n1/=n2) Then
          Write (nout,99992) '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,99991)(k,k=1,n1)
        Write (nout,*) 'Row'
        Do j = 1, m1
          Write (nout,99999) j, nobs(j,1:n1)
          Write (nout,99998) pred(j,1:n1)
          Write (nout,99994) 'Row total = ', nobs(j,n)
        End Do
        Write (nout,*)
        Write (nout,*) 'Column'
        Write (nout,99993) 'totals', nobs(m,1:n1)
        Write (nout,99994) 'Grand total = ', nobs(m,n)
        Write (nout,*)
        Write (nout,99997) 'Chi-squared = ', chis, '   D.F. = ', ndf
      Else
        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 j = 1, 2
          Write (nout,99999) j, nobs(j,1:2)
          Write (nout,99994) 'Row total = ', nobs(j,n)
        End Do
        Write (nout,*)
        Write (nout,*) 'Column'
        Write (nout,99993) 'totals', nobs(m,1:2)
        Write (nout,99994) 'Grand total = ', nobs(m,n)
        Write (nout,*)
        Write (nout,99996) '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,99995)(j,p(j),j=1,num)
      End If

99999 Format (1X,I2,4X,10I6)
99998 Format (8X,12F6.0)
99997 Format (1X,A,F10.3,A,I3)
99996 Format (1X,A,I4,A)
99995 Format (1X,I2,F9.4)
99994 Format (49X,A,I7)
99993 Format (1X,A,10I6)
99992 Format (1X,A,I3,A,I3)
99991 Format (7X,10I6)
    End Program g01affe
```

10.2 Program Data

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

10.3 Program Results

```
 G01AFF Example Program Results

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

 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
```