G11AAF – 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

G11AAF computes χ^2 statistics for a two-way contingency table. For a 2×2 table with a small number of observations exact probabilities are computed.

2 Specification

```
 SUBROUTINE G11AAF(NROW, NCOL, NOBST, LDT, EXPT, CHIST, PROB, CHI,
1                  G, DF, IFAIL)
 INTEGER           NROW, NCOL, NOBST(LDT,NCOL), LDT, IFAIL
 real              EXPT(LDT,NCOL), CHIST(LDT,NCOL), PROB, CHI, G, DF
```

3 Description

For a set of n observations classified by two variables, with r and c levels respectively, a two-way table of frequencies with r rows and c columns can be computed.

$$\begin{array}{cccc|c} n_{11} & n_{12} & \dots & n_{1c} & n_{1.} \\ n_{21} & n_{22} & \dots & n_{2c} & n_{2.} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n_{r1} & n_{r2} & \dots & n_{rc} & n_{r.} \\ \hline n_{.1} & n_{.2} & \dots & n_{.c} & n \end{array}$$

To measure the association between the two classification variables two statistics that can be used are:

The Pearson χ^2 statistic $= \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-f_{ij})^2}{f_{ij}}$, and the likelihood ratio test statistic $= 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \times \log(n_{ij}/f_{ij})$.

Where f_{ij} are the fitted values from the model that assumes the effects due to the classification variables are additive, i.e., there is no association. These values are the expected cell frequencies and are given by,

$$f_{ij} = n_{i.}n_{.j}/n.$$

Under the hypothesis of no association between the two classification variables, both these statistics have, approximately, a χ^2 distribution with $(c-1)(r-1)$ degrees of freedom. This distribution is arrived at under the assumption that the expected cell frequencies, f_{ij}, are not too small. For a discussion of this point see Everitt [1]. He concludes by saying, "... in the majority of cases the chi-square criterion may be used for tables with expectations in excess of 0.5 in the smallest cell".

In the case of the 2×2 table, i.e., $c = 2$ and $r = 2$, the χ^2 approximation can be improved by using Yates' continuity correction factor. This decreases the absolute value of $(n_{ij} - f_{ij})$ by $\frac{1}{2}$. For 2×2 tables with a small value of n the exact probabilities from Fisher's test are computed. These are based on the hypergeometric distribution and are computed using G01BLF. A two-tail probability is computed as $\min(1, 2p_u, 2p_l)$, where p_u and p_l are the upper and lower one-tail probabilities from the hypergeometric distribution.

4 References

[1] Everitt B S (1977) *The Analysis of Contingency Tables* Chapman and Hall

[2] Kendall M G and Stuart A (1973) *The Advanced Theory of Statistics (Volume 2)* Griffin (3rd Edition)

5 Parameters

1: NROW — INTEGER *Input*

On entry: the number of rows in the contingency table, r.

Constraint: NROW ≥ 2.

2: NCOL — INTEGER *Input*

On entry: the number of columns in the contingency table, c.

Constraint: NCOL ≥ 2.

3: NOBST(LDT,NCOL) — INTEGER array *Input*

On entry: the contingency table NOBST(i,j) must contain n_{ij} for $i = 1, 2, \ldots, r$; $j = 1, 2, \ldots, c$.

Constraint: NOBST$(i,j) \geq 0$ for $i = 1, 2, \ldots, r$; $j = 1, 2, \ldots, c$.

4: LDT — INTEGER *Input*

On entry: the first dimension of the arrays NOBST, EXPT and CHIST as declared in the (sub)program from which G11AAF is called.

Constraint: LDT $\geq$ NROW.

5: EXPT(LDT,NCOL) — ***real*** array *Output*

On exit: the table of expected values. EXPT(i,j) contains f_{ij} for $i = 1, 2, \ldots, r$; $j = 1, 2, \ldots, c$.

6: CHIST(LDT,NCOL) — ***real*** array *Output*

On exit: the table of χ^2 contributions. CHIST(i,j) contains $\frac{(n_{ij}-f_{ij})^2}{f_{ij}}$ for $i = 1, 2, \ldots, r$; $j = 1, 2, \ldots, c$.

7: PROB — ***real*** *Output*

On exit: if $c = 2$, $r = 2$ and $n \leq 40$ then PROB contains the two-tail significance level for Fisher's exact test, otherwise PROB contains the significance level from the Pearson χ^2 statistic.

8: CHI — ***real*** *Output*

On exit: the Pearson χ^2 statistic.

9: G — ***real*** *Output*

On exit: the likelihood ratio test statistic.

10: DF — ***real*** *Output*

On exit: the degrees of freedom for the statistics.

11: 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**.

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 specified by the routine:

IFAIL = 1

On entry, NROW < 2,
or NCOL < 2,
or LDT < NROW.

IFAIL = 2

On entry, a value in NOBST < 0, or all values in NOBST are zero.

IFAIL = 3

On entry, a 2×2 table has a row or column with both values 0.

IFAIL = 4

At least one cell has expected frequency, f_{ij}, ≤ 0.5. The χ^2 approximation may be poor.

7 Accuracy

For the accuracy of the probabilities for Fisher's exact test see G01BLF.

8 Further Comments

The routine G01AFF allows for the automatic amalgamation of rows and columns. In most circumstances this is not recommended, see Everitt [1].

Multi-dimensional contingency tables can be analysed using log-linear models fitted by G02GBF.

9 Example

The data below, taken from Everitt [1], is from 141 patients with brain tumours. The row classification variable is the site of the tumour: frontal lobes, temporal lobes and other cerebral areas. The column classification variable is the type of tumour: benign, malignant and other cerebral tumours.

23	9	6	38
21	4	3	28
34	24	17	75
78	37	26	141

The data is read in and the statistics 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.

```
*     G11AAF Example Program Text
*     Mark 16 Release. NAG Copyright 1992.
*     .. Parameters ..
      INTEGER          NIN, NOUT
      PARAMETER        (NIN=5,NOUT=6)
      INTEGER          CMAX, RMAX
      PARAMETER        (CMAX=3,RMAX=3)
```

```
*     .. Local Scalars ..
      real             CHI, DF, G, PROB
      INTEGER          I, IFAIL, J, NCOL, NROW
*     .. Local Arrays ..
      real             CHIST(RMAX,CMAX), EXPT(RMAX,CMAX)
      INTEGER          NOBST(RMAX,CMAX)
*     .. External Subroutines ..
      EXTERNAL         G11AAF
*     .. Executable Statements ..
      WRITE (NOUT,*) ' G11AAF Example Program Results'
*     Skip heading in data file
      READ (NIN,*)
      READ (NIN,*) NROW, NCOL
      IF (NROW.LE.RMAX .AND. NCOL.LE.CMAX) THEN
         DO 20 I = 1, NROW
            READ (NIN,*) (NOBST(I,J),J=1,NCOL)
   20    CONTINUE
         IFAIL = -1
*
         CALL G11AAF(NROW,NCOL,NOBST,RMAX,EXPT,CHIST,PROB,CHI,G,DF,
     +               IFAIL)
*
         IF (IFAIL.EQ.0 .OR. IFAIL.EQ.3) THEN
            WRITE (NOUT,*)
            WRITE (NOUT,99999) ' Probability = ', PROB
            WRITE (NOUT,99998) ' Pearson Chi-square statistic = ', CHI
            WRITE (NOUT,99998) ' Likelihood ratio test statistic = ', G
            WRITE (NOUT,99997) ' Degrees of freedom = ', DF
         END IF
      END IF
      STOP
*
99999 FORMAT (A,F6.4)
99998 FORMAT (A,F8.3)
99997 FORMAT (A,F4.0)
      END
```

9.2 Program Data

```
G11AAF Example Program Data
3 3                        : NROW NCOL
23  9  6                   : NOBST
21  4  3
34 24 17
```

9.3 Program Results

```
 G11AAF Example Program Results

Probability = 0.0975
Pearson Chi-square statistic =    7.844
Likelihood ratio test statistic =    8.096
Degrees of freedom =   4.
```