NAG Toolbox for MATLAB

g11aa

1 Purpose

g11aa 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 Syntax

```
[expt, chist, prob, chi, g, df, ifail] = g11aa(nrow, nobs, 'ncol', ncol)
```

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, $\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij}-f_{ij}\right)^2}{f_{ij}}$, and the likelihood ratio test statistic, $2\sum_{i=1}^{r}\sum_{j=1}^{c}n_{ij}\times\log\left(n_{ij}/f_{ij}\right)$, 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 1977. 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 $\left(n_{ij}-f_{ij}\right)$ 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 g01bl. 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

Everitt B S 1977 *The Analysis of Contingency Tables* Chapman and Hall

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

5 Parameters

5.1 Compulsory Input Parameters

1: **nrow – int32 scalar**

r, the number of rows in the contingency table.

Constraint: **nrow** ≥ 2.

2: **nobs**(**ldnobs**,**ncol**) **– int32 array**

ldnobs, the first dimension of the array, must be at least **nrow**.

The contingency table **nobs**(i,j) must contain n_{ij}, for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$.

Constraint: **nobs**$(i,j) \geq 0$, for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$.

5.2 Optional Input Parameters

1: **ncol – int32 scalar**

Default: The dimension of the arrays **nobs**, **expt**, **chist**. (An error is raised if these dimensions are not equal.)

c, the number of columns in the contingency table.

Constraint: **ncol** ≥ 2.

5.3 Input Parameters Omitted from the MATLAB Interface

ldnobs

5.4 Output Parameters

1: **expt**(**ldnobs**,**ncol**) **– double array**

The table of expected values. **expt**(i,j) contains f_{ij}, for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$.

2: **chist**(**ldnobs**,**ncol**) **– double array**

The table of χ^2 contributions. **chist**(i,j) contains $\frac{\left(n_{ij} - f_{ij}\right)^2}{f_{ij}}$, for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$.

3: **prob – double scalar**

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.

4: **chi – double scalar**

The Pearson χ^2 statistic.

5: **g – double scalar**

The likelihood ratio test statistic.

6: **df – double scalar**

The degrees of freedom for the statistics.

7: **ifail – int32 scalar**

0 unless the function detects an error (see Section 6).

6 Error Indicators and Warnings

Note: g11aa may return useful information for one or more of the following detected errors or warnings.

ifail = 1

On entry, **nrow** < 2,
or **ncol** < 2,
or **ldnobs** < **nrow**.

ifail = 2

On entry, a value in **nobs** < 0, or all values in **nobs** 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 g01bl.

8 Further Comments

The function g01af allows for the automatic amalgamation of rows and columns. In most circumstances this is not recommended; see Everitt 1977.

Multi-dimensional contingency tables can be analysed using log-linear models fitted by g02gb.

9 Example

```
nrow = int32(3);
nobst = [int32(23), int32(9), int32(6);
     int32(21), int32(4), int32(3);
     int32(34), int32(24), int32(17)];
[expt, chist, prob, chi, g, df, ifail] = g11aa(nrow, nobst)
```

```
expt =
   21.0213    9.9716    7.0071
   15.4894    7.3475    5.1631
   41.4894   19.6809   13.8298
chist =
    0.1863    0.0947    0.1447
    1.9605    1.5251    0.9063
    1.3519    0.9479    0.7267
prob =
    0.0975
chi =
    7.8441
g =
    8.0958
df =
     4
ifail =
           0
```