NAG Toolbox

nag_contab_chisq (g11aa)

1 Purpose

nag_contab_chisq (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] = nag_contab_chisq(nrow, nobs, 'ncol',
ncol)
```

```
[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.

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

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 nag_stat_prob_hypergeom (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** – INTEGER

r, the number of rows in the contingency table.

Constraint: **nrow** ≥ 2.

2: **nobs**($ldnobs$, **ncol**) – INTEGER array

$ldnobs$, the first dimension of the array, must satisfy the constraint $ldnobs \geq$ **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** – INTEGER

Default: the second dimension of the array **nobs**.

c, the number of columns in the contingency table.

Constraint: **ncol** ≥ 2.

5.3 Output Parameters

1: **expt**($ldnobs$, **ncol**) – REAL (KIND=nag_wp) 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**) – REAL (KIND=nag_wp) array

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

3: **prob** – REAL (KIND=nag_wp)

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** – REAL (KIND=nag_wp)

The Pearson χ^2 statistic.

5: **g** – REAL (KIND=nag_wp)

The likelihood ratio test statistic.

6: **df** – REAL (KIND=nag_wp)

The degrees of freedom for the statistics.

7: **ifail** – INTEGER

ifail $= 0$ unless the function detects an error (see Section 5).

6 Error Indicators and Warnings

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

Errors or warnings detected by the function:

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 $= 4$ (*warning*)

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

ifail $= -99$

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

ifail $= -399$

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

ifail $= -999$

Dynamic memory allocation failed.

7 Accuracy

For the accuracy of the probabilities for Fisher's exact test see nag_stat_prob_hypergeom (g01bl).

8 Further Comments

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

Multidimensional contingency tables can be analysed using log-linear models fitted by nag_correg_glm_binomial (g02gb).

9 Example

The data below, taken from Everitt (1977), 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

```
function g11aa_example

fprintf('g11aa example results\n\n');

nrow  = nag_int(3);
nobst = [nag_int(23),  9,  6;
                 21,   4,  3;
                 34,  24, 17];

[expt, chist, prob, chi, g, df, ifail] = ...
  g11aa(nrow, nobst);

% Display results
fprintf('Probability                          = %9.4f\n', prob);
fprintf('Pearson Chi-square statistic    = %8.3f\n', chi);
fprintf('Likelihood ratio test statistic = %8.3f\n', g);
fprintf('Degrees of freedom              = %4.0f\n', df);
```

9.2 Program Results

```
g11aa example results

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