nag_chi_sq_2_way_table (g11aac)

1. Purpose

nag_chi_sq_2_way_table (g11aac) 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

```
#include <nag.h>
#include <nagg11.h>

void nag_chi_sq_2_way_table(Integer nrow, Integer ncol, Integer nobst[],
          Integer tdt, double expt[], double chist[], double *prob,
          double *chi, double *g, double *df, NagError *fail)
```

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:

$$\text{The Pearson } \chi^2 \text{ statistic} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - f_{ij})^2}{f_{ij}},$$

and

$$\text{The likelihood ratio test statistic} = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \times \log(n_{ij}/f_{ij}).$$

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 (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 $(n_{ij} - f_{ij})$ 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 nag_hypergeom_dist (g01blc). 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. Parameters

nrow

Input: the number of rows in the contingency table, r.
Constraint: **nrow** ≥ 2.

ncol
Input: the number of columns in the contingency table, c.
Constraint: **ncol** ≥ 2

nobst[nrow][tdt]
Input: the contingency table, **nobst**[i−1][j−1] must contain n_{ij} for $i = 1, 2, \ldots, r$; $j = 1, 2, \ldots, c$.
Constraint: **nobst**[i−1][j−1] ≥ 0 for $i = 1, 2, \ldots, r$; $j = 1, 2, \ldots, c$.

tdt
Input: the last dimension of the arrays **nobst**, **expt** and **chist** as declared in the function from which nag_chi_sq_2_way_table is called.
Constraint: **tdt** $\geq$ **ncol**.

expt[nrow][tdt]
Output: the table of expected values, **expt**[i−1][j−1] contains f_{ij} for $i = 1, 2, \ldots, r$; $j = 1, 2, \ldots, c$.

chist[nrow][tdt]
Output: the table of χ^2 contributions, **chist**[i−1][j−1] contains $\dfrac{(n_{ij} - f_{ij})^2}{f_{ij}}$ for $i = 1, 2, \ldots, r$; $j = 1, 2, \ldots, c$.

prob
Output: 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.

chi
Output: the Pearson χ^2 statistic.

g
Output: the likelihood ratio test statistic.

df
Output: the degrees of freedom for the statistics.

fail
The NAG error parameter, see the Essential Introduction to the NAG C Library.

5. Error Indications and Warnings

NE_INT_ARG_LT
On entry, **nrow** must not be less than 2: **nrow** = ⟨*value*⟩
On entry, **ncol** must not be less than 2: **ncol** = ⟨*value*⟩

NE_2_INT_ARG_LT
On entry **tdt**=⟨*value*⟩ while **ncol** = ⟨*value*⟩. These parameters must satisfy **tdt** $\geq$ **ncol**

NE_2D_INT_ARR_ELEM
On entry **nobst**[⟨*value*⟩][⟨*value*⟩]=⟨*value*⟩. All elements of this array must be ≥ 0.

NE_2D_INT_ARR_ELEMS
On entry all elements of the array **nobst** are 0. At least one element of this array must be > 0.

NE_TABLE_DEGENERATE
On entry a 2*2 table has a row or column with both elements zero i.e., the table is degenerate.

NE_LOW_EXPECTED_FREQ
At least one cell has expected frequency ≤ 0.5. The chi-square approximation may be poor.

NE_INTERNAL_ERROR
An internal error has occurred in this function. Check the function call and any array sizes. If the call is correct then please consult NAG for assistance.

6. Further Comments

Multi-dimensional contingency tables can be analysed using log-linear models fitted by nag_glm_binomial (g02gbc).

6.1. Accuracy

For the accuracy of the probabilities for Fisher's exact test see nag_hypergeom_dist (g01blc).

6.2. References

Everitt B S (1977) *The Analysis of Contingency Tables.* Chapman and Hall.
Kendall M G and Stuart A (1979) *The Advanced Theory of Statistics (Volume 2).* Griffin (4th Edition).

7. See Also

None

8. 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.

8.1. Program Text

```
/* nag_chi_sq_2_way_table(g11aac) Example Program.
 *
 * Copyright 1996 Numerical Algorithms Group.
 *
 * Mark 4, 1996.
 *
 */

#include <nag.h>
#include <stdio.h>
#include <nag_stdlib.h>
#include <nagg11.h>

#define RMAX 3
#define CMAX 3
#define TDA CMAX

main()
{

  double prob, expt[RMAX][CMAX];
  double g;
  double chist[RMAX][CMAX];
  double df, chi;

  Integer ncol;
  Integer nrow;
  Integer i, j;
  Integer nobst[RMAX][CMAX];
  Integer tda = TDA;

  Vprintf("g11aac Example Program Results\n\n");
```

```
  /*     Skip heading in data file */
  Vscanf("%*[^\n] ");

  Vscanf("%ld %ld %*[^\n] ", &nrow, &ncol);
  if (nrow <= RMAX && ncol <= CMAX)
    {
      for (i = 0; i < nrow; ++i)
        {
          for (j = 0; j < ncol; ++j)
            Vscanf("%ld", &nobst[i][j]);
          Vscanf("%*[^\n]");
        }

      g11aac(nrow, ncol, (Integer *)nobst, tda, (double *)expt,
             (double *)chist, &prob, &chi, &g, &df, NAGERR_DEFAULT);

      Vprintf("Probability = %6.4f\n", prob);
      Vprintf("Pearson Chi-square statistic = %8.3f\n", chi);
      Vprintf("Likelihood ratio test statistic = %8.3f\n", g);
      Vprintf("Degrees of freedom = %4.0f\n", df);
    }
  exit(EXIT_SUCCESS);
}
```

8.2. Program Data

```
g11aac Example Program Data
3 3                          : nrow ncol
23  9  6                     : nobst
21  4  3
34 24 17
```

8.3. Program Results

```
g11aac Example Program Results

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