NAG Library Function Document

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,
      const 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.

$$
\begin{array}{ccccc}
n_{11} & n_{12} & \cdots & n_{1c} & n_{1.} \\
n_{21} & n_{22} & \cdots & n_{2c} & n_{2.} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
n_{r1} & n_{r2} & \cdots & n_{rc} & n_{r.} \\
\hline
n_{.1} & n_{.2} & \cdots & n_{.c} & n_{r.}
\end{array}
$$

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_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 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 (3 Volumes)* (4th Edition) Griffin

5 Arguments

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**[**nrow** $\times$ **tdt**] – const Integer *Input*

On entry: the contingency table, **nobst**$[(i-1) \times$ **tdt** $+ j - 1]$ must contain n_{ij}, for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$.

Constraint: **nobst**$[(i-1) \times$ **tdt** $+ j - 1] \geq 0$ for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$.

4: **tdt** – Integer *Input*

On entry: the stride separating matrix column elements in the arrays **nobst**, **expt**, **chist**.

Constraint: **tdt** $\geq$ **ncol**.

5: **expt**[**nrow** $\times$ **tdt**] – double *Output*

On exit: the table of expected values, **expt**$[(i-1) \times$ **tdt** $+ j - 1]$ contains f_{ij}, for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$.

6: **chist**[**nrow** $\times$ **tdt**] – double *Output*

On exit: the table of χ^2 contributions, **chist**$[(i-1) \times$ **tdt** $+ j - 1]$ contains $\frac{\left(n_{ij} - f_{ij}\right)^2}{f_{ij}}$, for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$.

7: **prob** – double * *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** – double * *Output*

On exit: the Pearson χ^2 statistic.

9: **g** – double * *Output*

On exit: the likelihood ratio test statistic.

10: **df** – double * *Output*

On exit: the degrees of freedom for the statistics.

11: **fail** – NagError * *Input/Output*

The NAG error argument (see Section 3.6 in the Essential Introduction).

6 Error Indicators and Warnings

NE_2_INT_ARG_LT

On entry, **tdt** = $\langle value \rangle$ while **ncol** = $\langle value \rangle$. These arguments must satisfy **tdt** $\geq$ **ncol**.

NE_2D_INT_ARR_ELEM

On entry, **nobst**$[(\langle value \rangle) \times$ **tdt** $+ \langle value \rangle] = \langle value \rangle$. 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_INT_ARG_LT

On entry, **ncol** = $\langle value \rangle$.
Constraint: **ncol** ≥ 2.

On entry, **nrow** = $\langle value \rangle$.
Constraint: **nrow** ≥ 2

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 contact NAG for assistance.

NE_LOW_EXPECTED_FREQ

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

NE_TABLE_DEGENERATE

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

7 Accuracy

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

8 Parallelism and Performance

Not applicable.

9 Further Comments

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

10 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.

10.1 Program Text

```
/* nag_chi_sq_2_way_table (g11aac) Example Program.
 *
 * Copyright 2014 Numerical Algorithms Group.
 *
 * Mark 4, 1996.
 * Mark 8 revised, 2004.
 *
 */

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

#define EXPT(I, J)  expt[(I) *tdnobs + J]
#define CHIST(I, J) chist[(I) *tdnobs + J]
#define NOBST(I, J) nobst[(I) *tdnobs + J]
int main(void)
{

  Integer  exit_status = 0, i, j, ncol, *nobst = 0, nrow, tdnobs;
  NagError fail;
  double   chi, *chist = 0, df, *expt = 0, g, prob;

  INIT_FAIL(fail);

  printf("nag_chi_sq_2_way_table (g11aac) Example Program Results\n\n");

  /* Skip heading in data file */
#ifdef _WIN32
  scanf_s("%*[^\n] ");
#else
  scanf("%*[^\n] ");
#endif

#ifdef _WIN32
  scanf_s("%"NAG_IFMT" %"NAG_IFMT" %*[^\n] ", &nrow, &ncol);
#else
  scanf("%"NAG_IFMT" %"NAG_IFMT" %*[^\n] ", &nrow, &ncol);
#endif
  if (nrow >= 2 && ncol >= 2)
    {
      if (!(expt = NAG_ALLOC((nrow)*(ncol), double)) ||
          !(chist = NAG_ALLOC((nrow)*(ncol), double)) ||
          !(nobst = NAG_ALLOC((nrow)*(ncol), Integer)))
        {
          printf("Allocation failure\n");
          exit_status = -1;
          goto END;
        }
      tdnobs = ncol;
    }
  else
    {
      printf("Invalid nrow or ncol.\n");
      exit_status = 1;
      return exit_status;
    }
  for (i = 0; i < nrow; ++i)
    {
      for (j = 0; j < ncol; ++j)
#ifdef _WIN32
        scanf_s("%"NAG_IFMT"", &NOBST(i, j));
#else
        scanf("%"NAG_IFMT"", &NOBST(i, j));
#endif
#ifdef _WIN32
        scanf_s("%*[^\n]");
#else
```

```
        scanf("%*[^\n]");
#endif
    }

  /* nag_chi_sq_2_way_table (g11aac).
   * chi^2 statistics for two-way contingency table
   */
  nag_chi_sq_2_way_table(nrow, ncol, nobst, tdnobs, expt,
                         chist, &prob, &chi, &g, &df, &fail);
  if (fail.code != NE_NOERROR)
    {
      printf("Error from nag_chi_sq_2_way_table (g11aac).\n%s\n",
             fail.message);
      exit_status = 1;
      goto END;
    }

  printf("Probability = %6.4f\n", prob);
  printf("Pearson Chi-square statistic = %8.3f\n", chi);
  printf("Likelihood ratio test statistic = %8.3f\n", g);
  printf("Degrees of freedom = %4.0f\n", df);
 END:
  NAG_FREE(expt);
  NAG_FREE(chist);
  NAG_FREE(nobst);
  return exit_status;
}
```

10.2 Program Data

```
nag_chi_sq_2_way_table (g11aac) Example Program Data
3 3                             : nrow ncol
23  9  6                        : nobst
21  4  3
34 24 17
```

10.3 Program Results

```
nag_chi_sq_2_way_table (g11aac) Example Program Results

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