NAG C Library Chapter Introduction

g11 – Contingency Table Analysis

Contents

1 Scope of the Chapter

This chapter contains a function for the analysis of two-way contingency tables. Functions in Chapter g02 may be used to fit generalised linear models to discrete data.

2 Background

2.1 Discrete Data

Discrete data can be usefully categorized into three types.

(1) *Binary data*. The variables can take one of two values, for example yes or no. The data may be grouped, for example the number of yes responses in ten questions.

(2) *Categorical data.* The variables can take one of two or more values or levels, but the values are not considered to have any ordering; for example the values may be red, green, blue or brown.

(3) *Ordered categorical data.* This is similar to categorical data but an ordering can be placed on the levels, for example: poor, average or good.

A second important categorization to be made is whether one of the discrete variables can be considered as a response variable or whether it is just the association between the discrete variables that is being considered. If the response variable is binary then a logistic or probit regression model can be used. These are special cases of the generalised linear model with binomial errors. Handling a categorical or ordered categorical response variable is more complex; for discussion of appropriate models see McCullagh and Nelder (1983).

To investigate the association between discrete variables a contingency table can be used.

2.2 Tabulation

The basic summary for multivariate discrete data is the multidimensional table in which each dimension is specified by a discrete variable. If the cells of the table are the number of observations with the corresponding values of the discrete variables then it is a contingency table. The discrete variables that can be used to classify a table are known as factors. For example, the factor sex would have the levels male and female. These can be coded as 1 and 2 respectively. Given several factors a multi-way table can be constructed such that each cell of the table represents one level from each factor. For example, a sample of 120 observations with the two factors sex and habitat, habitat having three levels: inner-city, suburban and rural, would give the 2 by 3 contingency table:

Sex	**Habitat**		
	Inner-city	Suburban	Rural
Male	32	27	15
Female	21	19	6

If the sample also contains continuous variables such as age, the average for the observations in each cell could be computed,

Sex	**Habitat**		
	Inner-city	Suburban	Rural
Male	25.5	30.3	35.6
Female	23.2	29.1	30.4

or other summary statistics.

Given a table, the totals or means for rows, columns etc. may be required. Thus the above contingency table with marginal totals is:

Sex	**Habitat**			
	Inner-city	Suburban	Rural	Total
Male	32	27	15	74
Female	21	19	6	46
Total	53	46	21	120

Note that the marginal totals for columns is itself a 2 by 1 table. Also, other summary statistics could be used to produce the marginal tables such as means or medians. Having computed the marginal tables, the cells of the original table may be expressed in terms of the margins, for example, in the above table the cells could be expressed as percentages of the column totals.

2.3 Contingency Tables

The simplest case is the two-way table formed when considering two discrete variables. For a data set of n observations classified by the two variables with r and c levels respectively, a two-way table of frequencies or counts with r rows and c columns can be computed.

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

If p_{ij} is the probability of an observation in cell ij then the the model which assumes no association between the two variables is the model

$$p_{ij} = p_{i.}p_{.j}$$

where $p_{i.}$ is the marginal probability for the row variable and $p_{.j}$ is the marginal probability for the column variable, the marginal probability being the probabilty of observing a particular value of the variable ignoring all other variables. The appropriateness of this model can be assessed by two commonly used statistics:

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

$$2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \times \log(n_{ij}/f_{ij}).$$

The f_{ij} are the fitted values from the model; these values are the expected cell frequencies and are given by

$$f_{ij} = n\hat{p}_{ij} = n\hat{p}_{i.}\hat{p}_{.j} = n(n_{i.}/n)(n_{.j}/n) = n_{i.}n_{.j}/n.$$

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.

In the case of the 2 by 2 table, i.e., $c = 2$ and $r = 2$, the χ^2-approximation can be improved by using Yates's continuity correction factor. This decreases the absolute value of $(n_{ij} - f_{ij})$ by $\frac{1}{2}$. For 2 by 2 tables with a small values of n the exact probabilities can be computed; this is known as Fisher's exact test.

An alternative approach, which can easily be generalised to more than two variables, is to use log-linear models. A log-linear model for two variables can be written as

$$\log(p_{ij}) = \log(p_{i.}) + \log(p_{.j}).$$

A model like this can be fitted as a generalised linear model with Poisson error with the cell counts, n_{ij}, as the response variable.

3 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)* Griffin (4th Edition)

McCullagh P and Nelder J A (1983) *Generalized Linear Models* Chapman and Hall

4 Available Functions

g11aac χ^2 statistic for two-way contingency table

g11bac Computes multiway table from set of classification factors using selected statistic

g11bbc Computes multiway table from set of classification factors using given percentile/quantile

The following routines may also be used to analyse discrete data:

g02gcc Fit a log-linear model to a contingency table

g02gbc Fit generalized linear models to binary data