NAG Library Chapter Introduction

G11 – Contingency Table Analysis

Contents

1   Scope of the Chapter

The routines in this chapter are for the analysis of discrete multivariate data. One suite of routines computes tables while other routines are for the analysis of two-way contingency tables, conditional logistic models and one-factor analysis of binary data.

Routines in Chapter G02 may be used to fit generalized linear models to discrete data including binary data and contingency tables.

2   Background to the Problems

2.1   Discrete Data

Discrete variables can be defined as variables which take a limited range of values. Discrete data can be usefully categorized into three types.

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.

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.

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

Data containing discrete variables can be analysed by computing summaries and measures of association and by fitting models.

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 2×3 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 2×1 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   Discrete Response Variables and Logistic Regression

A second important categorization in addition to that given in Section 2.1 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, for example, success or failure, then a logistic or probit regression model can be used. The logistic regression model relates the logarithm of the odds-ratio to a linear model. So if p_i is the probability of success, the model relates $\log(p_i/(1 - p_i))$ to the explanatory variables. If the responses are independent then these models are special cases of the generalized linear model with binomial errors. However, there are cases when the binomial model is not suitable. For example, in a case-control study a number of cases (successes) and number of controls (failures) is chosen for a number of sets of case-controls. In this situation a conditional logistic analysis is required.

Handling a categorical or ordered categorical response variable is more complex, for a discussion on the appropriate models see McCullagh and Nelder (1983). These models generally use a Poisson distribution.

Note that if the response variable is a continuous variable and it is only the explanatory variables that are discrete then the regression models described in Chapter G02 should be used.

2.4   Contingency Tables

If there is no response variable then to investigate the association between discrete variables a contingency table can be computed and a suitable test performed on the table. The simplest case is the two-way table formed when considering two discrete variables. For a dataset 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 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 probability 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

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij}-f_{ij}\right)^2}{f_{ij}},$$

and the likelihood ratio test statistic

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

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

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

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 2×2 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 $1/2$. For 2×2 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 generalized to more than two variables, is to use log-linear models. A log-linear model for two variables can be written as

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

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

2.5 Latent Variable Models

Latent variable models play an important role in the analysis of multivariate data. They have arisen in response to practical needs in many sciences, especially in psychology, educational testing and other social sciences.

Large-scale statistical enquiries, such as social surveys, generate much more information than can be easily absorbed without condensation. Elementary statistical methods help to summarise the data by looking at individual variables or the relationship between a small number of variables. However, with many variables it may still be difficult to see any pattern of inter-relationships. Our ability to visualize relationships is limited to two or three dimensions putting us under strong pressure to reduce the dimensionality of the data and yet preserve as much of the structure as possible. The question is thus one of how to replace the many variables with which we start by a much smaller number, with as little loss of information as possible.

One approach to the problem is to set up a model in which the dependence between the observed variables is accounted for by one or more latent variables. Such a model links the large number of observable variables with a much smaller number of latent variables.

Factor analysis, as described in Chapter G03, is based on a linear model of this kind when the observed variables are continuous. Here we consider the case where the observed variables are binary (e.g., coded 0/1 or true/false) and where there is one latent variable. In educational testing this is known as latent trait analysis, but, more generally, as factor analysis of binary data.

A variety of methods and models have been proposed for this problem. The models used here are derived from the general approach of Bartholomew (1980) and Bartholomew (1984). You are referred to Bartholomew (1980) for further information on the models and to Bartholomew (1987) for details of the method and application.

3 Recommendations on Choice and Use of Available Routines

3.1 Tabulation

The following routines can be used to perform the tabulation of discrete data:

G11BAF computes a multidimensional table from a set of discrete variables or classification factors. The cells of the table may be counts or a summary statistic (total, mean, variance, largest or smallest) computed for an associated continuous variable. Alternatively, G11BAF will update an existing table with further data.

G11BBF computes a multidimensional table from a set of discrete variables or classification factor where the cells are the percentile or quantile for an associated variable. For example, G11BBF can be used to produce a table of medians.

G11BCF computes a marginal table from a table computed by G11BAF or G11BBF using a summary statistic (total, mean, median variance, largest or smallest).

3.2 Analysis of Contingency Tables

G11AAF computes the Pearson and likelihood ratio χ^2 statistics for a two-way contingency table. For 2×2 tables Yates's correction factor is used and for small samples, $n \leq 40$, Fisher's exact test is used.

In addition, G02GCF can be used to fit a log-linear model to a contingency table.

3.3 Binary data

The following routines can be used to analyse binary data:

G11SAF fits a latent variable model to binary data. The frequency distribution of score patterns is required as input data. If your data is in the form of individual score patterns, then the service routine G11SBF may be used to calculate the frequency distribution.

G11CAF estimates the parameters for a conditional logistic model.

In addition, G02GBF fits generalized linear models to binary data.

4 Functionality Index

5 Auxiliary Routines Associated with Library Routine Arguments

None.

6 Routines Withdrawn or Scheduled for Withdrawal

None.

7 References

Bartholomew D J (1980) Factor analysis for categorical data (with Discussion) *J. Roy. Statist. Soc. Ser. B* **42** 293–321

Bartholomew D J (1984) The foundations of factor analysis *Biometrika* **71** 221–232

Bartholomew D J (1987) *Latent Variable Models and Factor Analysis* Griffin

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

Kendall M G and Stuart A (1969) *The Advanced Theory of Statistics (Volume 1)* (3rd Edition) Griffin

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

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