THE ANALYSIS OF CONTINGENCY TABLES

Log-linear Poisson models for weighted numbers

R-76-31
J. de Leeuw, Leyden State University, Dept. Data theory
S. Oppe, Institute for Road Safety Research SWOV
Voorburg, 1976
Institute for Road Safety Research SWOV, The Netherlands

CONTENTS

1. INTRODUCTION

Contingency tables (or cross tables) classify elements of populations or samples (of varying kinds) with reference to one or more characteristics. For instance, classification of fatalities for a year according to age and mode of road usage. If there is only one characteristic, one often speaks of marginal tables. But this is also said of tables originating if one or more variables of a contingency table are added and one or more others are not. Since there is no essential difference between a marginal table and a contingency table (merely a functional difference), our future references will be to contingency tables.

The term "contingency table" is better than "cross table", because it expresses something of the assumptions made in cross-table analysis as regards the contingent factors assumed to play a part in creating the table. This aspect is essential in sampling.

By means of a sample we try on the one hand to describe the population from which the sample is taken, and on the other to verify opinions about this population. The models of analysis described below assume that a sample gives a more or less correct picture of the population, dependent only on random fluctuations.

Assumptions regarding the way chance plays a part form the basis of the model of analysis. Within it, different model specifications are again possible.

Analysis of contingency tables does not usually assume that there are specific relations (such as order relations or even metric relations) between the classes of a characteristic.

Such extra assumptions are possible, however, within specific models, for instance for a variable such as age.

In recent years there has been a new development in the way in which contingency tables are analysed. While it used to be customary (mostly by the Chi-squared test) to verify overall-hypotheses about a table with either one or two characteristics, analysis now increasingly stresses the detailed information the table contains. Furthermore, it is also possible to analyse higer-order tables (subdivided into a numer of characteristics), in order that more complex relationships, i.e. relations between more than two characteristics at once, can be investigated.

2. MODEL

2.1. Fundamental assumption

The fundamental assumption is that the number of accidents in the cells of the contingency table are independent random variables with a Poisson distribution, in which the Poisson distribution parameters may differ. To keep this fairly concrete: if we are concerned with a two-way table with r rows and k columns, then we could write the Poisson assumption for each cell as follows:
there are numbers $\lambda_{ij} \geqslant 0$ (i = 1, .., r; j = 1, .., k) such that:

$$\text{prob}\left[\underset{\sim}{X}_{ij} = x_{ij}\right] = e^{-\lambda_{ij}} \frac{\lambda_{ij}^{x_{ij}}}{x_{ij}!}$$

In this, $\underset{\sim}{X}_{ij}$ is the stochastic variable of cell (i,j) which can assume as values the natural numbers $x_{ij} = 0,1,2,\ldots$ A briefer way of writing this assumption is:

$$\underset{\sim}{X}_{ij} \sim \mathcal{P}(\lambda_{ij}),$$

which we can read as: $\underset{\sim}{X}_{ij}$ has a Poisson distribution with a parameter λ_{ij}.

2.2. Independence assumptions regarding characteristics in models

Although we assume that the $\underset{\sim}{X}_{ij}$ are independent, it is of course possible that there exist relations between parameters λ_{ij}. By investigating the relations between these parameters we can examine whether the characteristics the variables possess are also independent. What do we mean when we say that the rows and columns of an r x k contingency table (with independent Poisson variables $\underset{\sim}{X}_{ij}$) correspond to independent row and column variables? Suppose $\underset{\sim}{X}_{i.}$ and $\underset{\sim}{X}_{.j}$ are the marginal distributions, viz:

$$\underset{\sim}{X}_{i.} = \sum_{j=1}^{k} \underset{\sim}{X}_{ij}$$

and

$$\underset{\sim}{X}_{.j} = \sum_{i=1}^{r} \underset{\sim}{X}_{ij}$$

The requirement that the row and column variables must be independent means that the chances of the r conditional distributions within rows:

$$\text{prob} \left[(\underset{\sim}{X}_{i1} = x_{i1}) \wedge (\underset{\sim}{X}_{i2} = x_{i2}) \wedge \ldots \wedge (\underset{\sim}{X}_{ik} = x_{ik}) \;\middle|\; \underset{\sim}{X}_{i.} = x_{i.} \right]$$

are the same for all i=1, ..., r, and that the k conditional distributions within columns:

$$\text{prob} \left[(\underset{\sim}{X}_{1j} = x_{1j}) \wedge (\underset{\sim}{X}_{2j} = x_{2j}) \wedge \ldots \wedge (\underset{\sim}{X}_{rj} = x_{rj}) \;\middle|\; \underset{\sim}{X}_{.j} = x_{.j} \right]$$

are the same for all j=1, ..., k. Using the independence of the $\underset{\sim}{X}_{ij}$ and the Poisson assumption, we can infer that the conditional distributions within rows are the same as the multinomial distributions:

$$\frac{x_{i.}!}{\prod_{j=1}^{k} x_{ij}!} \cdot \prod_{j=1}^{k} \left(\frac{\lambda_{ij}}{\lambda_{i.}} \right)^{x_{ij}} ,$$

while the conditional distributions within columns are the same as the multinomial distributions:

$$\frac{x_{.j}!}{\prod_{i=1}^{r} x_{ij}!} \cdot \prod_{i=1}^{r} \left(\frac{\lambda_{ij}}{\lambda_{.j}} \right)^{x_{ij}}$$

The row and column variables are thus independent when $\left(\frac{\lambda_{ij}}{\lambda_{i.}}\right)$ is the same for all j, and $\left(\frac{\lambda_{ij}}{\lambda_{.j}}\right)$ is the same for all i.

Necessary and sufficient conditions for this are that there are numbers $a_i \geqslant 0$ (i=1,..,r), $b_j \geqslant 0$ (j=1,..,k), and $m \geqslant 0$, such that $\lambda_{ij} = m\, a_i\, b_j$ for all i,j.

This multiplicative model is mostly converted into a linear model by taking the logarithm:

$$\ln \lambda_{ij} = \mu + \alpha_i + \beta_j \qquad (1)$$

in which $\alpha_i = \ln a_i$ and so on.

Hence, such models are also called log-linear models. The log-linear model (1) is therefore equivalent with the requirement of independence of the row and column variables.

2.3. Saturated and unsaturated models

As stated, in addition to testing hypotheses regarding tables, description of the tables is sometimes also required. If the characteristics are not independent and the above model (1) does not therefore apply, it can be extended with specific parameters for the cells. In that case we have the following model

$$\ln \lambda_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \qquad (2)$$

With regard to this model it is always possible to find such parameters μ, α_i, β_j and γ_{ij} that there is complete agreement between the table one wishes to describe and the model used for description. The significance of the description is that the variation in the numbers of observations of the cells in the table is shown in relation to the structure of the table: it can be seen, for instance, to what extent the variation results from a row-effect, a column-effect or an interaction effect. Although there are now as many parameters as cells and

hence there is no reduction in information, there is an ordering of information. Moreover, note that model (1) is a special case of model (2); it is the same, except for the restriction that $\gamma_{ij} = 0$ for all i, j. Other restrictions are also possible, for instance that the α_i's together form a linear relation or, for example, are equal to zero. In all such cases we speak of unsaturated models. If we are dealing with a sample, these non-saturated models can be regarded as verifiable hypotheses regarding the population from which the sample originates. With a saturated model, such verification is not possible because the model fully describes the data.
As regards the choice of the model of analysis, there is close agreement with linear models as used in analysis of variance. Here, again, we can speak of a breakdown of the table into components: how great is the row-contribution, the column-contribution, the unique cell contribution of each cell? For an incidental table this can be examined by estimating the parameters of the model.
This systematic breakdown therefore provides an efficient review of the information contained in the table. It is also possible to give confidence limits of the estimators for the parameters, so that verification of individual estimators is also possible.
A good description of the relation between analysis of variance models and log-linear models is given bij Nelder & Wedderburn (1972).

2.4. Weighted Poisson models

So far we treated only the numbers of accidents as a function of a number of characteristics. But we are sometimes interested in analysing accident figures normalised for a given exposure factor such as number of inhabitants, road lengths, and so on. If we enter the numbers of accidents in the table with a measure of exposure per cell, which may differ from cell to cell, we can use a more general Poisson model. The fundamental assumption now becomes

$$X_{ij} \sim P(c_{ij}\,\lambda_{ij}),$$

in which the c_{ij} are the given exposure factors, and in which a log-linear model is again assumed for the λ_{ij}.

3. THE DESIGN MATRIX

3.1. General

In matrix notation, the general form of a log-linear model for n Poisson variables $\underset{\sim}{X}_1 \sim P(\lambda_1)$ can be written as

$$\eta = V\theta,$$

in which η is a vector of values $\eta_1 = \ln \lambda_1$. V is a given matrix of the order n x p (known as the design matrix), and θ a vector of p unknown parameters. If the $\underset{\sim}{X}_1$ are arranged in a two-way table and if we replace the index 1 by the row and column indices i and j, then we can rewrite the model

$$\eta_{ij} = \ln \lambda_{ij} = \mu + \alpha_i + \beta_j$$

when r = k = 2, for instance as

$$\begin{bmatrix} \eta_{11} \\ \eta_{21} \\ \eta_{12} \\ \eta_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

Note that in this case the design matrix V is of the order 4 x 5 and rank 3. This become clear if we rewrite the model in the equivalent form:

$$\begin{bmatrix} \eta_{11} \\ \eta_{21} \\ \eta_{12} \\ \eta_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \\ \theta_3 \end{bmatrix}$$

with:

$$\theta_1 = \mu + \bar{\alpha} + \bar{\beta}$$
$$\theta_2 = (\alpha_1 - \bar{\alpha}) = -(\alpha_2 - \bar{\alpha})$$
$$\theta_3 = (\beta_1 - \bar{\beta}) = -(\beta_2 - \bar{\beta})$$

in which $\bar{\alpha}$ and $\bar{\beta}$ respectively are the means of the α's and β's. Generally, it is always possible (and advisable) to choose the design matrix so that its rank is the same as its number of columns. This obviates having to impose extra restrictions on the parameters in order to find a unique solution. If we were to seek a direct solution for the α's and β's, these restrictions would be:

$$\alpha_1 + \alpha_2 = 0 \text{ and } \beta_1 + \beta_2 = 0.$$

The rank matrix of the columns and the number of columns are, for instance, always the same if we choose V so that V'V is diagonal, with V' as the transpose of matrix V (V is then called column-wise orthogonal), or so that V'V is equal to the unit matrix (V is then called column-wise orthonormal).

3.2. Three useful forms of design matrices

3.2.1. Helmert matrices

Let us first consider the case in which we have a single classification. Example: i=1,.....,n corresponds to n age categories, $\underset{\sim}{X}_i$ is the number of accidents in each such category. A first-type design matrix often used is the Helmert matrix.
A complete Helmert matrix for n = 4 is as follows:

1	-1	-1	-1
1	1	-1	-1
1	0	2	-1
1	0	0	3

Note that this V is column-wise orthogonal. The model $\eta = V\theta$ is therefore saturated. A perfect fit is possible if we choose

$\hat{\theta} = (V'V)^{-1}V'\eta$. Unsaturated models are possible by omitting columns of V, which agrees with the hypothesis that some of the elements of θ in the saturated model are equal to zero. The interpretation of Helmert effects becomes clear from the following equivalences:

$$\theta_1 = 0 \Leftrightarrow \sum \eta_i = 0 \Leftrightarrow \sqrt[n]{\prod_{i=1}^{n} \lambda_i} = 1$$

$$\theta_2 = 0 \Leftrightarrow \eta_2 = \eta_1 \Leftrightarrow \lambda_2 = \sqrt[1]{\lambda_1}$$

$$\theta_3 = 0 \Leftrightarrow 2\eta_3 = \eta_1 + \eta_2 \Leftrightarrow \lambda_3 = \sqrt[2]{\lambda_1 \lambda_2}$$

$$\theta_4 = 0 \Leftrightarrow 3\eta_4 = \eta_1 + \eta_2 + \eta_3 \Leftrightarrow \lambda_4 = \sqrt[3]{\lambda_1 \lambda_2 \lambda_3}$$

From this, we can for instance derive:

$$\theta_3 = \theta_4 = 0 \Leftrightarrow \lambda_3 = \lambda_4 = \sqrt[2]{\lambda_1 \lambda_2},$$

and so on. Helmert effects therefore compare every λ_i individually with the geometric mean of the preceding λ_i. In this way, we can discover whether there is a trend in our data, or perhaps a sudden jump.

3.2.2. Orthogonal polynomials

Let us assume that the age categories in our example are intervals of equal length. We might then be interested in the functional relation between age and accident rate. We can describe this functional relation as a polynomial, i.e. as a linear combination of orthogonal polynomials; for n = 3 this gives, for instance, the following (column-wise orthogonal) design matrix:

1 -1 1
1 0 -2
1 1 1

Each constant function on (1,2,3) is of course a multiple of the first column of V, each linear function on (1,2,3) is a linear combination of the first two columns, and each second-degree function is a linear combination of the first three columns. Each function on (1,2,3) can be regarded as a second degree function: this is merely another way of saying that the model defined by V is saturated. Unsaturated models generally have the form $\theta_3 = 0$ or $\theta_2 = \theta_3 = 0$. The hypothesis $\theta_3 = 0$ says that the three points $(1,\eta_1)$, $(2,\eta_2)$, and $(3,\eta_3)$ are on a straight line, the hypothesis $\theta_2 = \theta_3 = 0$ says that $\eta_1 = \eta_2 = \eta_3$. Generally, the hypothesis that $(\eta_1,\ldots,\eta_n)$ is a q^{th} degree polynomial on $(1,2,\ldots,n)$ can be written $\eta_i = \Pi_q(i)$. From our discussion it follows that:

$$\eta_i = \Pi_q(i) \Leftrightarrow \theta_{q+1} = \ldots = \theta_n = 0.$$

The interpretation of polynomial effects in log-linear models is rendered difficult by the use of log-transformation, since:

$$\eta_i = \Pi_q(i) \Leftrightarrow \lambda_i = \exp(\Pi_q(i)) = \exp(\alpha_o + \alpha_1 i + \ldots.$$

$$+\alpha_q i^q) = \left[\exp(i^0)\right]^{\alpha_0} \left[\exp(i^1)\right]^{\alpha_1} \ldots \left[\exp(i^q)\right]^{\alpha_q}.$$

This latter function is rather less simple and unacquaintanced than a polynomial.

3.2.3. Between-Within contrasts

In many cases, categories of our classification break down naturally into different groups. Age can, for instance, be divided into two groups: the under-forties and over-forties. This subdivision can be shown in saturated design-matrix form as

00-20	1	-1	-1	0
20-40	1	-1	+1	0
40-60	1	1	0	-1
60-80	1	1	0	+1

In this case the measurements themselves are therefore in four categories, and we examine so to speak whether subdivision into fewer categories is possible without forfeiting too much information. The first column of V corresponds as usual to the total average, the second column contrasts the two groups (the between-group effect), and the third and fourth columns examine the effects within the groups individually. If there are K groups with n_k elements ($\sum_{k=1}^{K} n_k = n$), then there are generally K-1 between-group effects, and $\sum_{k=1}^{K} (n_k - 1) = n - K$ within-group effects. The most common unsaturated models state that all θ values corresponding to between-group effects are zero. This agrees with the hypothesis that the arithmetic means of the η_i are the same for each group, which is equivalent to the fact that the geometric means of the λ_i are the same for every group.

3.2.4. Combination of design matrices

Let us now examine a two-way classification with, for instance, two classes in the first characteristic (e.g. male against female), and four classes in the second characteristic (e.g. the four age groups in the preceding section). We first choose two design matrices V_1 and V_2 for the separate characteristics.
For example:

$$V_1 = \begin{matrix} +1 & -1 \\ +1 & +1 \end{matrix}$$

$$V_2 = \begin{matrix} +1 & -1 & -1 & 0 \\ +1 & -1 & +1 & 0 \\ +1 & +1 & 0 & -1 \\ +1 & +1 & 0 & +1 \end{matrix}$$

We next form from all 2 x 4 = 8 combinations of columns of V_1 and V_2 the external product (the external product of an n-vector x and an m-vector y is an n x m matrix with the elements $x_i y_j$). This gives the following eight matrices:

V_1	V_2	Product			
1	1	+1	+1	+1	+1
		+1	+1	+1	+1
1	2	-1	-1	+1	+1
		-1	-1	+1	+1
1	3	-1	+1	0	0
		-1	+1	0	0
1	4	0	0	-1	+1
		0	0	-1	+1
2	1	-1	-1	-1	-1
		+1	+1	+1	+1
2	2	+1	+1	-1	-1
		-1	-1	+1	+1
2	3	+1	-1	0	0
		-1	+1	0	0
2	4	0	0	+1	-1
		0	0	-1	+1

We can treat these eight matrices as eight vectors of eight elements, and thus form a design matrix V_{12} with these vectors as columns. Thus:

Design matrix								Belonging to vector
+1	-1	-1	0	-1	+1	+1	0	η_{11}
+1	-1	+1	0	-1	+1	-1	0	η_{12}
+1	+1	0	-1	-1	-1	0	+1	η_{13}
+1	+1	0	+1	-1	-1	0	-1	η_{14}
+1	-1	-1	0	+1	-1	-1	0	η_{21}
+1	-1	+1	0	+1	-1	+1	0	η_{22}
+1	+1	0	-1	+1	+1	0	-1	η_{23}
+1	+1	0	+1	+1	+1	0	+1	η_{24}

The matrix V_{12} so constructed is again column-wise orthogonal, and defines a saturated model. We can say that V_{12} is formed via external products. In using a design matrix built up in this way, we usually wish to investigate a given type of unsaturated models. Let us examine these unsaturated models with respect to our example.
We first choose the column corresponding to the first column of V_1 and the first column of V_2. This is the first column of V_{12}. The hypothesis $\theta_1 = 0$ is equivalent to the hypothesis that the arithmetic mean of the η_{ij} (i=1,2;j=1,2,3,4) is zero, i.e. that the geometric mean of the λ_{ij} equals one.
We next choose the group of columns of V_{12} composed from the first column of V_1 and column two, three or four of V_2. These are columns 2, 3, 4 of V_{12}. The hypothesis $\theta_2 = \theta_3 = \theta_4 = 0$ is equivalent to the hypothesis that the column averages of the η_{ij} are identical, or:

$$\eta_{.1} = \eta_{.2} = \eta_{.3} = \eta_{.4}$$

This is equivalent to:

$$\lambda_{11}\lambda_{12} = \lambda_{12}\lambda_{22} = \lambda_{13}\lambda_{23} = \lambda_{14}\lambda_{24}.$$

In the same way, we can choose the group of columns composed from the first column of V_2 and a non-first column of V_1. This group consists of the fifth column of V_{12}. The hypothesis $\theta_5 = 0$ is:

$$\lambda_{11}\lambda_{12}\lambda_{13}\lambda_{14} = \lambda_{21}\lambda_{22}\lambda_{23}\lambda_{24}.$$

Lastly, there is the group of columns 6, 7, 8 corresponding to a non-first column of V_1 and a non-first column of V_2. The hypothesis $\theta_6 = \theta_7 = \theta_8 = 0$ corresponds to:

$$\eta_{ij} = \frac{\eta_{i.}}{k} + \frac{\eta_{.j}}{r} - \frac{\eta_{..}}{kxr}$$

that is to say with lack of additive interaction in the η_{ij} (cf. model (1) on page 6), which in turn is the same as the lack of multiplicative interaction in the λ_{ij} (for a comparison of these

two forms of interaction see Darroch, 1974; Lancaster, 1973, 1975). It is clear that this form of analysis via external products can be generalised to tables with more than two classifications. We always begin with design matrices for each of the characteristics, form external products, and group the columns of the ultimate design matrix by examing which first columns appear in them. Hence, we form groups of effects corresponding to the additive interactions of the η's (which are known from ordinary analysis of variance, and to multiplicative interactions of the λ's (which can be interpreted in the manner of section 2.2. as independence models). It is important to realise that an interaction hypothesis in the form $\theta_6 = \theta_7 = \theta_8 = 0$ in the above example is either true or not true, regardless of the choice of the original V_1, V_2 ..... The choice of design matrix for a given characteristic is therefore of importance only for better interpretation of the individual θ's, but is of no importance in describing the table according to the contributions of the characteristics or the interactions between them.

4. PARAMETER ESTIMATION AND HYPOTHESIS TESTING

4.1. Introduction

For convenience, we will briefly enumerate the fundamental assumptions for the class of models in which we are interested.

A1: $\underset{\sim}{X}_i \sim \mathcal{P}(\ell_i \lambda_i^o)$

A2: $\underset{\sim}{X}_i$ are independent

A3: $\eta^o = V \theta^o$.

In A1, ℓ is a known vector of weights (or measures of exposure); in A3, $\eta^o = \ln \lambda^o$, and V is a known n x p design matrix, which we shall assume to be column-wise orthonormal. The superscript 'o' with θ, η, and λ is to indicate the 'actual' value of these parameters, and to distinguish them from estimators and variables in specific functions. What interests us in the first place is estimation of the p unknown parameters, and in the second place verifying whether the model A1, A2, A3 is correct. For this, it is important also to formulate in A3 other (equivalent) ways. If V is an n x p column-wise orthonormal matrix, then there is an n x (n - p) column-wise orthonormal matrix V_c such that $V'V_c = 0$. It is clear that A3 can also be written as:

A3: $V_c' \eta^o = 0$.

A third formulation is possible if we define the p-dimensional linear space $\mathcal{V}$ as:

$$\mathcal{V} = \{\eta \mid V_c'\eta = 0\} = \{\eta \mid \eta = V\theta\}.$$

then

A3: $\eta^o \in \mathcal{V}$.

It is generally unfeasible to use estimators and test procedures which are optimal for all conceivable sample sizes. We shall therefore use asymptotic arguments, and derive estimates and tests which have optimum properties if certain factors tend to infinity. For this purpose, we reformulate A1 as:

A1: $\underset{\sim}{X}_i \sim \mathcal{P}(m\, \varrho_i\, \lambda^o_i)$.

The factor m indicates how great our weights ϱ_i and parameters λ^o_i are on average. If we continue our observations, the $\underset{\sim}{X}_i$ will of course tend to infinity. The assumption A1 says, in fact, that all $\underset{\sim}{X}_i$ tend to infinity just as quickly: if m becomes infinitely great, then the values $\underset{\sim}{X}_i/m$ converge (in probability) towards the fixed factors $\varrho_i\, \lambda^o_i$.
For our analyses, it is generally unnecessary to known the value of m; we must merely be prepared to make this assumption. The following facts are known from the general theory of asymptotic statistical analysis. In the first place we shall be interested in estimators that are <u>consistent</u>; viz. if $m \to \infty$ then $\hat{\theta}(m) \xrightarrow{P} \theta^o$. In the second place, we are interested in estimators that are <u>asymptotically normal</u>, which means that their distribution more and more resembles a multinormal distribution if m tends to infinity. For estimators with these two properties, which we can summarise as:

T1: $m^{\frac{1}{2}}\ (\hat{\theta}(m) - \theta^o) \xrightarrow{L} \mathcal{N}(0, \Sigma)$

the asymptotic dispersion matrix Σ satisfies the dissimilarity:

$$\Sigma \geqslant (V'M^oV)^{-1}$$

in which M^o is the diagonal matrix with the values $\varrho_i\, \lambda^o_i$ on the diagonal. Estimators in this class for which the dissimilarity is a similarity, and which in a certain sense are thus as precise as possible, are called <u>efficient</u>. Although nearly all available estimators satisfy T1, they do not necessarily meet the stricter requirement:

$$T2: \; m^{\frac{1}{2}} \; (\hat{\theta}(m) - \theta^{o}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \; (V'M^{o}V)^{-1}).$$

Since efficiency is a necessary property, we shall confine ourselves to efficient estimators (i.e. estimators satisfying T2). Moreover, it is important to add that confidence intervals of estimators and tests of hypotheses via these estimators are generally asymptotically optimal if the estimators are efficient.
It is known that efficient estimators can be found by maximising the likelihood function, that gives the likelihood of the observations as a function of the parameters, and that an asymptotically optimum test of A3 within A1-A2 is possible by calculating the likelihood ratio between the most suitable estimator or estimators and the hypothetical value of the parameter or parameters. The estimation and test theory based on this maximum likelihood is set forth for log-linear Poisson models in Haberman (1974). The theory is modified for weighted Poisson models in De Leeuw (1975). On the whole, calculations based on likelihood are not very simple, and we consider here a different class of estimators and tests (also optimal and efficient), based on Neyman's modified minimum Chi-squared method (1949).

4.2. Modified minimum Chi-squared methods

We begin this section with a known limit theorem for Poisson variables which, applied to A1, says that if $m \longrightarrow \infty$:

$$\frac{\underset{\sim}{X}_i - m \, \rho_i \, \lambda_i^{o}}{(m \, \rho_i \, \lambda_i^{o})^{\frac{1}{2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

If we define

$$\underset{\sim}{Y}_i = \underset{\sim}{X}_i \; / \; (m \; \rho_i)$$

then we can rewrite it in a rather more convenient form:

$$m^{\frac{1}{2}} \; (\underset{\sim}{Y}_i - \lambda_i^{o}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \; \frac{\lambda_i^{o}}{\rho_i}).$$

If, lastly, we define:

$$\underset{\sim}{Z}_i = \ln \underset{\sim}{Y}_i$$

it follows from this that

$$m^{\frac{1}{2}} (\underset{\sim}{Z}_i - \eta_i^o) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \ (\varrho_i \lambda_i^o)^{-1})$$

The modified minimum Chi-squared method we shall discuss below has a simple geometric interpretation. We define the distance measure:

$$\delta (\eta_1, \eta_2) = (\eta_1 - \eta_2)' \underset{\sim}{X} (\eta_1 - \eta_2).$$

The matrix $\underset{\sim}{X}$ is diagonal, and the $\underset{\sim i}{X}$ are on the diagonal. Note that so far we have already demonstrated that:

$$\delta (\underset{\sim}{Z}, \eta^o) \xrightarrow{\mathcal{L}} \chi^2_n$$

if $m \to \infty$ (this follows from the limit distribution of $\underset{\sim}{Z}$, and from $\underset{\sim i}{X}/m \xrightarrow{P} \varrho_i \lambda_i^o$). For estimations, we consider the distance between the vector $\underset{\sim}{Z}$ of observations, and the collection of permitted estimators η. For calculating the modified minimum Chi-squared estimators we must choose $\hat{\eta}$ so that:

$$\delta (\underset{\sim}{Z}, \hat{\eta}) = \min_{\eta \in V} \delta (\underset{\sim}{Z}, \eta).$$

This gives an estimation $\hat{\eta}$ for η^o. The corresponding estimator for θ^o is $V'\hat{\eta}$, and the statistic used for testing A3 is $\delta (\underset{\sim}{Z}, \hat{\eta})$. In the next section we study the distribution of estimators and statistics.

4.3. Calculations and limit distributions

The problem

$$\min_{\eta \in V} \delta (\underset{\sim}{Z}, \eta)$$

can be formulated in two different ways. The first formulation is:

$$\min_{\theta} \delta(Z, V\theta)$$

This gives estimators $\hat{\theta}_I$, and then $\eta_I = V\hat{\theta}_I$. Formulation II uses Lagrange multipliers and can be written as

$$\min_{\eta} \max_{\omega} \delta(Z, \eta) + 2\omega' V_c' \eta .$$

This gives estimators $\hat{\eta}_{II}$ and $\hat{\omega}$, and then $\hat{\theta}_{II} = V' \eta_{II}$. The (n-p)-element vector ω is a vector of indefinite multipliers. The model can now also be written as:

A3 : $\omega = o$.

As the solution of the original problem is unique, obviously

$\hat{\theta}_I = \hat{\theta}_{II} = \hat{\theta}$ and

$\hat{\eta}_I = \hat{\eta}_{II} = \hat{\eta}$.

It follows from formulation I that $\hat{\theta}$ is given by

$$\hat{\theta} = (V'\underset{\sim}{X}V)^{-1}V'\underset{\sim}{X}\underset{\sim}{Z}$$

and hence

$$\hat{\eta} = V(V'\underset{\sim}{X}V)^{-1}V'\underset{\sim}{X}\underset{\sim}{Z}.$$

It further follows that both $\hat{\theta}$ and $\hat{\eta}$ are efficient estimators; in other words:

$$m^{\frac{1}{2}}(\hat{\theta} - \theta^o) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (V'M^oV)^{-1})$$
$$m^{\frac{1}{2}}(\hat{\eta} - \eta^o) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V(V'M^oV)^{-1}V')$$

The asymptotic dispersion matrices can be estimated by:

$$S(\hat{\theta}) = (V'\underset{\sim}{X}V)^{-1}$$

$$S(\hat{\eta}) = V(V'\underset{\sim}{X}V)^{-1}V'$$

It furthermore follows from the given results that:

$$\delta\ (\underset{\sim}{Z}, \hat{\eta}) \xrightarrow{\mathcal{L}} \chi^2_{n-p} \ .$$

Formulation II gives other useful information. We find:

$$\hat{\omega} = (V_c'\underset{\sim}{X}^{-1}V_c)^{-1}V_c'\underset{\sim}{Z}$$

$$\hat{\eta} = \underset{\sim}{Z} - \underset{\sim}{X}^{-1}V_c(V_c'\underset{\sim}{X}^{-1}V_c)^{-1}V_c'\underset{\sim}{Z}.$$

The vectors $\hat{\omega}$ and $\hat{\eta}$ are asymptotically independent, and

$$m^{\frac{1}{2}}\hat{\omega} \xrightarrow{\mathcal{L}} \eta(0, (V_c'M_o^{-1}V_c)^{-1})$$

From equations I and II it also follows that $\delta\ (\underset{\sim}{Z}, \hat{\eta})$ can be written in three different forms:

$$\begin{aligned}\delta\ (\underset{\sim}{Z}, \hat{\eta}) &= \underset{\sim}{Z}' \left[\underset{\sim}{X} - \underset{\sim}{X}V(V'\underset{\sim}{X}V)^{-1}V'\underset{\sim}{X} \right] \underset{\sim}{Z} \\ &= \hat{\omega}'V_c'\underset{\sim}{X}^{-1}V_c\hat{\omega} \\ &= \underset{\sim}{Z}'V_c\ (V_c'\underset{\sim}{X}^{-1}V_c)^{-1}V_c'\underset{\sim}{Z}.\end{aligned}$$

The statistic $\delta\ (\underset{\sim}{Z}, \hat{\eta})$ is therefore also found if we test A3 in the form $V_c'\eta = 0$ or $\omega = 0$ by using the asymptotic distribution of $V_c'\underset{\sim}{Z}$ and $\hat{\omega}$. These tests are known respectively as the Wald test and the Lagrange multiplier test; in this context they are thus equivalent to the Neyman method.
Especially if V_c is a matrix of low-rank, the Wald test will be preferable.

LITERATURE

1. J.N. Darroch. 'Multiplicative and additive interaction in contingency tables'. Biometrika, 1974, p. 207.

2. L.A. Goodman. 'The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications'. J.A.S.A., 1970, p. 226.

3. L.A. Goodman. 'Guided and Unguided Methods for the Selection of Models for a Set of T Multidimensional Contingency Tables'. J.A.S.A., 1973, p. 165.

4. S.J. Haberman. 'The Analysis of Frequency Data'. Univ. of Chicago Press, 1974.

5. H.O. Lancaster. 'The multiplicative definition of interaction'. Austral. J. Statist., 1971, p. 36.

6. J. de Leeuw. 'Maximum Likelihood Estimation for Weighted Poisson Models'. RN005-75. Leyden State University, Dept. Data theory, 1975.

7. J.A. Nelder & R.W.M. Wedderburn. 'Generalized Linear Models'. J.R. Statist. Soc. A, 1972, p. 370.

8. J. Neyman. 'Contributions to the theory of the X^2-test'. Proc. of the Berkely Symp. on Math., Statist. and Probability, 1949, p. 239.

9. R.L. Plackett. 'The Analysis of Categorical Data'. Griffin, London, 1974.

ANNEX 1. CORRECTION FOR BIAS

We can apply a correction for bias if, before calculating the Z_i values, we first add $\frac{1}{2}$ to the X-values, so that Z_i is then defined as:

$$Z_i = \ln \frac{X_i + \frac{1}{2}}{m e_i}$$

Why $\frac{1}{2}$. Suppose we define:

$$\underset{\sim}{Z} = \ln \frac{\underset{\sim}{X} + a}{m e}$$

(For convenience we omit below the i and the small superscript o; we also define $\mu = e\lambda$). If:

$$\underset{\sim}{U} = \frac{(\underset{\sim}{X} - m\mu) + a}{m\mu}$$

then

$$\underset{\sim}{Z} = \eta + \underset{\sim}{U} - \tfrac{1}{2}\underset{\sim}{U}^2 + \tfrac{1}{3}\underset{\sim}{U}^3 - \tfrac{1}{4}\underset{\sim}{U}^4 + \ldots\ldots$$

From which it follows that:

$$E(\underset{\sim}{Z}) = \eta + \frac{1}{m}\left(\frac{2a - 1}{\mu}\right) - \left(\frac{1}{m}\right)^2 \left(\frac{6a^2 - 12a + 5}{12\mu^2}\right) + o(m^{-2}).$$

This correction also has the useful side-effect that Z is now also defined for X = 0.

ANNEX 2. COMPUTER PROGRAMME

In the computer programme it is necessary to read in per variable a design matrix of orthogonal column vectors, the first column vector being generated. The definitive design matrix is constructed in the programme with the aid of the external product method and then converted into an orthonormal matrix. If one is not interested in individual effects, therefore, the simplest method is to introduce Helmert effects. The θ's of the saturated model are calculated with the formula:

$$\theta = (V'\underset{\sim}{X}V)^{-1}V'\underset{\sim\sim}{X}Z$$

In the case of a saturated model for the orthonormal V-matrix this formula reduces to:

$$\theta = V'X^{-1}VV'\underset{\sim\sim}{X}Z = V'\underset{\sim}{Z}$$

The relevant variances, on the basis of which the standard scores are calculated, are on the diagonal of matrix $(V'XV)^{-1}$, which is calculated in the case of saturation as $V'X^{-1}V$, and hence no invertion is needed.
For testing hypotheses in which (always limited) groups of θ's are taken as zero. Formulation II on page 20 is used because in this case only a matrix of limited order need be inverted in order to obtain:

$$\delta\ (\underset{\sim}{Z}, \hat{\underset{\sim}{\eta}}) = \underset{\sim}{Z}'V_c\ (V_c'\ \underset{\sim}{X}^{-1}V_c)^{-1}V_c'\ Z$$

The matrix $V_c'\underset{\sim}{X}^{-1}V_c$ is given as a part matrix of matrix $V'X^{-1}V$ already calculated.

ANNEX 3. EXAMPLE OF AN ANALYSIS

As an illustration, an example is worked out below.
It has been chosen because of the simplicity of the table.

A three-way table was chosen, in which the variables are:
A: The province of Noord-Brabant as against the Rest of the Netherlands.
B: Drinking established as against not established.
C: Location on road (intersection, road section, corner/bend).
The cells of the table show the number of deaths in the years 1971-1973 (Central Statistical Office data), inside built-up areas.

		C_1 (intersection)	C_2 (straight road)	C_3 (corner/bend)
A_1 (N-Br)	B_1 (drinking)	22	48	14
	B_2 (not drinking)	243	272	48
A_2 (Rest of Neth.)	B_1	97	202	68
	B_2	1206	1442	189

These figures are weighted in the analysis for number of inhabitants in Noord-Brabant by factor 18.80 and in the Rest of the Netherlands by factor 115.08.
In analysis, use was made of the following design matrix V , built up from Helmert-effects.

Matrix:

$$V' = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 \\
1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 \\
1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 \\
1 & 1 & -2 & 1 & 1 & -2 & 1 & 1 & -2 & 1 & 1 & -2 \\
1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 \\
1 & -1 & 0 & 1 & -1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 \\
1 & 1 & -2 & 1 & 1 & -2 & -1 & -1 & 2 & -1 & -1 & 2 \\
1 & -1 & 0 & -1 & 1 & 0 & 1 & -1 & 0 & -1 & 1 & 0 \\
1 & 1 & -2 & -1 & -1 & 2 & 1 & 1 & -2 & -1 & -1 & 2 \\
1 & -1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & 1 & -1 & 0 \\
1 & 1 & -2 & -1 & -1 & 2 & -1 & -1 & 2 & 1 & 1 & -2
\end{bmatrix}$$

Effect:

- T: total
- A: Noord-Brabant as against Rest of Netherlands
- B: drinking established as against not established
- C_1: intersection as against road section
- C_2: intersection + road section as against corner/bend
- A x B
- A x C_1
- A x C_2
- B x C_1
- B x C_2
- A x B x C_1
- A x B x C_2

The results of the estimations for the saturated model are given below. The (2x2x3 = 12) estimators agree with a total effect, the main effects, first-order interaction effects and second-order interaction effects.

	Single scores	Chi-squared values	Degrees of freedom
Total effect:	+27.59 * *	761.28	1
Main effects:			
A-effect:	+ 4.02 * *	16.16	1
B-effect:	-24.24 * *	587.35	1
C-effects:	- 5.98 * *	265.27	2
	+14.19 * *		
First-order interaction effects:			
A x B effect:	+ 0.41 N.S.	0.17	1
A x C effects:	+ 0.10 N.S.	0.23	2
	- 0.46 N.S.		
B x C effects:	- 4.04 * *	43.26	2
	- 5.70 * *		
Second-order interaction effects:			
A x B x C effects:	- 0.35 N.S.	1.31	2
	+ 1.03 N.S.		
N.B. 95% limit:	$\pm$ 1.96	3.84	1
		5.99	2

Scores are in standard form. In this way it can already be ascertained which effects are significant, for instance at 5% level. These are marked * *. This test can also be made with an X^2-test. For each effect we then find one X^2-value (see column 2) with the relevant number of degrees of freedom (see column 3). Note that the X^2-values for df=1 are equal to the square of the single scores.
In the Chi-squared tests it is then assumed in each case that all estimators agreeing with the effect are equal to zero, and the X^2-value indicates the extent of the discrepancy between the model so obtained and the data.
Where we are concerned with one degree of freedom per effect, the significance of both tests is by definition identical. In this analysis, the other X^2-values provide the same result as the single-score test. This is not necessarily always the case. The single scores, for instance, may all be (just) not significant, but together yield a significant X^2-value.
It is also possible that only one single score is significant, which does not make the total X^2-value significant.
In such cases the X^2-value and the single scores thus provide additive information.

Interpretation of the data

In general the main effects and the total effect themselves are not very significant in interpreting the data. Here, however, where a correction was made for the number of inhabitants, it can be said as regards the A-effect that per inhabitant there are fewer accidents in the Rest of the Netherlands than in Noord-Brabant (the direction of the effect is shown by the sign!).
In order to interpret this phenomenon, we would have to know something for instance about the degree of urbanisation in Noord-Brabant and in the Rest of the Netherlands and also at least something about the number of traveller and vehicle kilometres. For interpretation of the B x C effect is is important to realise that this relates to cases where drinking was established. It would be

interesting to relate this to built-up/non-built-up areas as well. All this should make it clear that interpretation of the effects is an exercise unconnected with the analysis itself.