AD 747969

TRANSFORMATIONS FOR MULTIVARIATE BINARY DATA

by

P. Bloomfield

D D C
SEP 11 1972
C

Technical Report 18, Series 2

Department of Statistics

PRINCETON UNIVERSITY

August 1972

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
U S Department of Commerce
Springfield VA 22151

25

TRANSFORMATIONS FOR MULTIVARIATE BINARY DATA

by

P. Bloomfield

Summary

The interpretation of statistical data may often be simplified by a preliminary transformation. In the context of contingency tables, one way of achieving this would be to relabel the possible outcomes, or in other words to permute the cells of the table. For a 2^d table, certain permutations have the property that a loglinear model for the cell probabilities transforms in a simple way. These are, in a sense, linear transformations of the original variables.

The aim of making such a transformation is to fit the transformed data by a simple model, such as a low order hierarchical model or one in which certain variables are independent of others. A 2^4 table has been analysed with this end in view. All the models were fitted to the original data, and to do this a computer program has been developed which will fit nonhierarchical models by iterative scaling.

1. Introduction

It is often found that the analysis of statistical data may be simplified by using a suitably chosen transformation. The most common examples involve linear transformations of continuous variables. However, Goodman (1971) gives an example of a contingency table with an apparently rather complex structure, which may be simplified by describing the responses in terms of different variables. In other words, the variables used to index the data need to be transformed. Professor D. R. Cox has also mentioned in lectures the need for study of such transformations.

The simplest type of contingency table is the 2^d table, indexed by d variables, each taking just two values. The possible transformations of such variables are discussed in Section 2, and the subset of linear transformations is defined. Linear transformations have the advantage that a factorial-type model for the probability distribution, such as those discussed by Bahadur (1961) and Birch (1963), is also transformed in a simple way.

The aim of making a transformation is to simplify the structure of the data. For example, one might look for transformed variables to which a simple hierarchical model (Birch, 1963; Bishop, 1969) could be fitted. In Section 3 we examine a 2^4 table extracted from the data of Ries and Smith (1963), which has also been analysed by Cox and Lauh (1967) and Goodman (1971). Two transformations are used to show the type of simplification which might be achieved. In Section 4 we examine a problem which arises in the use of the Deming-Stephan (Deming and Stephan, 1940) algorithm to fit nonhierarchical models.

2. Transformations

Suppose that $\underset{\sim}{X} = \{X_1,\ldots,X_d\}$ is a d-variate random variable such that each X_α takes the values 0 or 1, $\alpha = 1,\ldots, d$. The set of possible values of $\underset{\sim}{X}$ is thus I_d, the set of vertices of the unit d-dimensional hyper-cube. A typical vertex will be denoted by $\underset{\sim}{i} = (i_1,\ldots,i_d)$, where each $i_\alpha = 0$ or 1, $\alpha = 1,\ldots,d$. If we draw a random sample of n $\underset{\sim}{X}$'s from some distribution over I_d, the collection of counts $n_{\underset{\sim}{i}}$ = no. of $\underset{\sim}{X}$'s taking the value $\underset{\sim}{i}$, $\underset{\sim}{i} \in I_d$, is a 2^d contingency table.

One type of transformation which has been used on such data is applied to the cell counts $n_{\underset{\sim}{i}}$, $\underset{\sim}{i} \in I_d$. Thus one might make a variance-stabilising transformation, or some transformation designed to reveal additivity of structure. However, we are concerned in the present paper with transformations not of cell counts but of the original random variable $\underset{\sim}{X}$.

In order to preserve the information present in $\underset{\sim}{X}$, we ask that the transformation should be invertible, and hence the range of the tranformed variable must contain exactly 2^d points. It is simplest to assume that this range is in fact I_d; thus a transformation of $\underset{\sim}{X}$ is simply a permutation of I_d.

Some of these permutations are of course trivial. A re-ordering of the components of $\underset{\sim}{X}$ will rarely be useful, and similarly a re-coding of any component, that is replacing X_α by $1-X_\alpha$. Thus there are $d!2^d$ trivially distinct versions of any transformation. However, this still leaves

$$n_d = \frac{2^d!}{d!2^d} = \frac{(2^d-1)!}{d!} \qquad (2.1)$$

non-trivial transformations (including the identity), a number which increases alarmingly for modest values of d. At $d = 4$, for instance, its value is around 5×10^{10}.

Clearly these transformations differ in the extent to which they change the original variables. In the simplest case of a 2×2 table, however, there is essentially only one type of transformation. We introduce this with an example due to D. R. Cox. Consider an experiment in which a couple are asked their voting intentions. Suppose that we code the responses as

$$X_1 = \begin{cases} 1 & \text{husband votes for Party D} \\ 0 & \text{husband votes for Party R ,} \end{cases} \tag{2.2}$$

with X_2 carrying similar meaning for the wife. If political considerations carried no weight in the choice of a marital partner, and if there were no subsequent interaction, then X_1 and X_2 would be independent, and would thus represent a simple and useful way of coding the responses. However, we could also use the coding

$$\begin{aligned} X_1' &= X_1 \, , \\ X_2' &= \begin{cases} 0 & X_1 = X_2 \\ 1 & X_1 \neq X_2 \, ; \end{cases} \end{aligned} \tag{2.3}$$

here X_2' records whether the voting intentions were the same or different. If it emerged that X_1' and X_2' were independent, then these would be the natural variables with which to record the responses.

For this 2×2 case, there is only one other transformation, to variables

$$X_1'' = X_2 \, , \quad X_2'' = X_2' \, . \tag{2.4}$$

Since $(2^2 - 1)!/2! = 3$, all other transformations may be obtained trivially from these three sets of variables.

It is interesting to note that the variables X_1' , X_2' , X_1'' and X_2'' may be written as linear transformations in residue arithmetic modulo 2 . For $X_2' = X_1 + X_2$ in this arithmetic, and this is the only new variable used. When $d > 2$, certain transformations may still be written in this way; for example, we could have had some third variable $X_3 = X_3' = X_3''$. However, not all transformations are linear when $d > 3$. The easiest way to see this is by counting. There are $2^d - 1$ linear functions of $X_1, \ldots, X_d$, corresponding to the inclusion or exclusion of each variable, and omitting the null function in which all variables are excluded. Thus the first transformed variable may be chosen in $2^d - 1$ possible ways. The second must be distinct from the first and may thus be chosen in $2^d - 2$ ways. The linear space generated by these two contains 3 non-null functions; hence the third variable must be chosen from the remaining $2^d - 4$. Continuing the argument, the total number of invertible linear transformations is

$$(2^d-1)(2^d-2)(2^d-4)\ldots(2^d-2^{d-1}) \quad .$$

Each of these occurs in $d!$ trivially distinct forms; the possibility of recoding any variable, that is interchanging 1 and 0 has been eliminated. This leaves a total of

$$(2^d-1)(2^d-2)\ldots(2^d-2^{d-1})/d! \quad ,$$

essentially distinct transformations, including as before the identity. This may be rewritten as

$$n_d = \frac{2^{d(d-1)2}}{d!} \prod_{\alpha=1}^{d} (2^\alpha - 1) \quad , \qquad (2.5)$$

a number which still increases rapidly as a function of d. However, since

$$n_d = \frac{(2^d-1)2^{d-1}}{d} \; n_{d-1} \; , \qquad (2.6)$$

which may be compared with

$$m_d = \frac{(2^d-1)!}{d(2^{d-1}-1)!} \; m_{d-1} \; , \qquad (2.7)$$

it is clear that n_d increases far more slowly than m_d. Thus the set of linear transformations is an increasingly small subset of the set of all transformations.

A comparison with normal theory suggests that in a first discussion of transformations, we should restrict our attention to linear transformations.

There is, however, a more compelling reason for doing this, to discuss which we need to consider the log-linear model (Birch, 1963) for the probabilities in a 2^k table. We begin by going back to the 2^2 example. Let $p_{\underset{\sim}{i}} = \mathrm{pr}(\underset{\sim}{X} = \underset{\sim}{i}) = \mathrm{pr}(X_1 = i_1, X_2 = i_2)$; $i_1, i_2 = 0,1$. Assuming that $p_{\underset{\sim}{i}} > 0$ for each $\underset{\sim}{i}$, we let $\xi \underset{\sim}{i} = \log P\underset{\sim}{i}$. Then the table of ξ's may be decomposed as in a factorial experiment, as the sum of different components. For our purposes, this is most conveniently written

$$\xi_{\underset{\sim}{i}} = \sum_{\underset{\sim}{j}\in I_2} \lambda_{\underset{\sim}{j}} (-1)^{\underset{\sim}{i}^{\tau}\underset{\sim}{j}} , \quad \underset{\sim}{i}\in I_2 , \tag{2.8}$$

or in full,

$$\left.\begin{aligned} \xi_{00} &= \lambda_{00} + \lambda_{01} + \lambda_{10} + \lambda_{11} , \\ \xi_{01} &= \lambda_{00} - \lambda_{01} + \lambda_{10} - \lambda_{11} , \\ \xi_{10} &= \lambda_{00} + \lambda_{01} - \lambda_{10} - \lambda_{11} , \\ \xi_{11} &= \lambda_{00} - \lambda_{01} - \lambda_{10} + \lambda_{11} . \end{aligned}\right\} \tag{2.9}$$

The superscript τ on a vector or matrix denotes transposition. Here λ_{11} is the "interaction" between X_1 and X_2 . When it vanishes, X_1 and X_2 are independent.

Now $\underset{\sim}{X}' = \underset{\sim}{T}\underset{\sim}{X}$, where

$$\underset{\sim}{T} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} ;$$

note that $T^{-1} = T$ in this residue arithmetic, and also that $(T^{\tau})^{-1} = T^{\tau}$. Thus

$$\left.\begin{aligned}
p'_{\underset{\sim}{i}} &= \mathrm{pr}(\underset{\sim}{X}' = \underset{\sim}{i}) \\
&= \mathrm{pr}(\underset{\sim}{T}\underset{\sim}{X} = \underset{\sim}{i}) \\
&= \mathrm{pr}(\underset{\sim}{X} = \underset{\sim}{T}^{-1}\underset{\sim}{i}) \\
&= p_{\underset{\sim}{T}^{-1}\underset{\sim}{i}} \\
&= \exp \sum_{\underset{\sim}{j}\in I_2} \lambda_{\underset{\sim}{j}} (-1)^{(\underset{\sim}{T}^{-1}\underset{\sim}{i})^{\tau}\underset{\sim}{j}} \\
&= \exp \sum_{\underset{\sim}{j}\in I_2} \lambda_{\underset{\sim}{j}} (-1)^{\underset{\sim}{i}^{\tau}\underset{\sim}{T}^{\tau-1}\underset{\sim}{j}} \\
&= \exp \sum_{\underset{\sim}{k}\in I_2} \lambda_{\underset{\sim}{T}^{\tau}\underset{\sim}{k}} (-1)^{\underset{\sim}{i}^{\tau}\underset{\sim}{k}}
\end{aligned}\right\} \qquad (2.10)$$

Thus the loglinear model for the probabilities $p'_{\underset{\sim}{i}}$, $\underset{\sim}{i}\in I_2$, has the same coefficients as that for $p_{\underset{\sim}{i}}$, $\underset{\sim}{i}\in I_2$, except that they have been permuted according to the transposed linear operator $\underset{\sim}{T}^{\tau}$ in the sense that $\lambda'_{\underset{\sim}{k}} = \lambda_{\underset{\sim}{T}^{\tau}\underset{\sim}{k}}$. Now the condition for independence of X'_1 and X'_2 is $\lambda'_{11} = 0 = \lambda_{01}$, and similarly X''_1 and X''_2 are independent if $\lambda_{10} = 0$. Thus one may tell from the coefficients in the model for the original variables whether either transformation will give rise to independent variables.

The same argument extends readily to $d > 2$. The expansion (2.8) for the log probabilities is still valid provided I_2 is replaced by I_d . A linear transformation can be written as $\underset{\sim}{X}' = \underset{\sim}{T}\underset{\sim}{X}$, where T is a matrix of zeroes and ones, having an inverse in residue arithmetic modulo 2 ; it is not in general true that $\underset{\sim}{T}^{-1} = \underset{\sim}{T}$. The sequence of manipulations (2.10) does not depend on d , and hence it is true for $d \geq 2$ that the coefficients are permuted according to $\underset{\sim}{T}^{\tau}$, that is $\lambda'_{\underset{\sim}{k}} = \lambda_{\underset{\sim}{T}^{\tau}\underset{\sim}{k}}$, $\underset{\sim}{k}\in I_d$. The possible gains from making such a transformation are discussed in the next section.

A different generalisation is to tables in which each variable may take more than two values. The most obvious generalisation is to tables in which each variable takes r values, $0,\ldots,r-1$. The natural arithmetic here is residue arithmetic modulo r , and the natural decomposition is the finite Fourier transform. Specifically, if $w = \exp(2\pi i/r)$, then

$$\xi_{\underset{\sim}{i}} = \log p_{\underset{\sim}{i}} = \log\{\mathrm{pr}(X=\underset{\sim}{i})\}$$

may be written as

$$\xi_{\underset{\sim}{i}} = \sum_{\underset{\sim}{j}\in I_{r,d}} \lambda_{\underset{\sim}{j}}\, w^{\underset{\sim}{i}^{\tau}\underset{\sim}{j}} , \quad \underset{\sim}{i}\in I_{r,d} \quad , \tag{2.11}$$

where $I_{r,d}$ is the set of all d-tuples $(i_1,\ldots, i_d)$ where $0 \leq i_\alpha < r$, $\alpha = 1,\ldots, d$. The inverse formula, defining the λ's , is

$$\lambda_{\underset{\sim}{j}} = (rd)^{-1} \sum_{\underset{\sim}{i}\in I_{r,d}} \xi_{\underset{\sim}{i}}\, w^{-\underset{\sim}{j}^{\tau}\underset{\sim}{i}} \tag{2.12}$$

Now suppose that $\underset{\sim}{T}$ is a (dxd) matrix whose entries t_{ij} are integers satisfying $0 \leq t_{ij} < r$, $i, j = 1,\ldots, d$, and that there exists an inverse $\underset{\sim}{T}^{-1}$ in residue arithmetic modulo r . Let $\underset{\sim}{X}' = \underset{\sim}{T}\underset{\sim}{X}$ in this arithmetic. Then

$$\mathrm{pr}(\underset{\sim}{X}' = \underset{\sim}{i}) = \mathrm{pr}(\underset{\sim}{X} = \underset{\sim}{T}^{-1}\underset{\sim}{i})$$

$$= \exp \sum_{\underset{\sim}{j}\in I_{r,d}} \lambda_{\underset{\sim}{j}}\, w^{(\underset{\sim}{T}^{-1}\underset{\sim}{i})^{\tau}\underset{\sim}{j}}$$

which simplifies to

$$\exp\Bigl(\sum_{\underset{\sim}{k}\in I_{r,d}} \lambda_{\underset{\sim}{T}^{\tau}\underset{\sim}{k}}\Bigr)$$

Thus as in the 2^d case, the effect of this transformation is simply to permute the coefficients according to $\underset{\sim}{T}^{\tau}$, that is $\lambda'_{\underset{\sim}{j}} = \lambda_{\underset{\sim}{T}^{\tau}\underset{\sim}{j}}$.

This generalisation is, however, rather restrictive in its structure. The Fourier decomposition is most suitable when the categories are ordered, but is essentially invariant under cyclic permutations of these categories. The type of data for which this seems natural would be where the categories could meaningfully be arranged in a circle, an unusual situation. Henceforth we shall only consider binary, that is dichotomous, variables.

It should be noted that the property that the model is transformed in a simple way when the variables are transformed linearly (in the present sense) is not restricted to the loglinear model described above. Clearly the same argument applies if any function of the cell probabilities is expanded in a factorial form, such as in the representation proposed by Bahadur (1961).

3. Motivation, and an example

The only reason mentioned as yet for transforming the variables by which a 2^d contingency table is classified has been to attain independence. For $d > 2$, one will rarely be able to transform to d mutually independent variables, but one might hope to find variables which could be partitioned into $\delta < d$ mutually independent blocks. Another possibility is that of conditional independence. Each of these distributions may be described in terms of restricted loglinear models.

The general version of (2.8) is

$$\log\{\mathrm{pr}(\underset{\sim}{X} = \underset{\sim}{i})\} = \sum_{\underset{\sim}{j} \in I_d} \lambda_{\underset{\sim}{j}} (-1)^{\underset{\sim}{i}^T \underset{\sim}{j}}, \quad \underset{\sim}{i} \in I_d \tag{3.1}$$

In a restricted loglinear model, the summation is restricted to some subset $J \subset I_d$. Thus effectively we constrain those λ's whose subscripts do not lie in J to be zero. Since $\exp(\lambda_{\underset{\sim}{0}})$ is merely a normalising constant, we shall always suppose that $\underset{\sim}{0} \in J$.

The models of block-wise independence and of conditional independence correspond to choices of J having certain specific structures; see Goodman (1970). These structures all possess the property of being hierarchical, which may be defined as follows. For $\underset{\sim}{i}, \underset{\sim}{j} \in I_d$, we write

$\underset{\sim}{i} \leq \underset{\sim}{j}$ if $i_\alpha \leq j_\alpha$, $\alpha = 1,\ldots,d$. Then the set J is hierarchical if for every $\underset{\sim}{j} \in J$, J contains every $\underset{\sim}{i} \leq \underset{\sim}{j}$. It may be seen from (3.1) that the parameter $\lambda_{\underset{\sim}{i}}$ describes the degree of interaction of the variables $(X_\alpha : i_\alpha = 1)$. Thus a model is hierarchical if whenever the interaction of a particular set of variables is included in the model, the interactions of all subsets of those variables are also included. The class of hierarchical models is evidently a natural one to consider. We observe here that this notation is in conflict with that of other authors, who have in general indexed interactions by the list of α's for which $j_\alpha = 1$, that is, by a subset of $(1,\ldots, d)$. A model then consists of a family of such subsets, and is hierarchical if the family of subsets is hierarchical in the usual sense. However, our present notation seems to be the most natural one to use in the present context.

One's aim in making a transformation of the kind discussed here would be to find a hierarchical model which fits the transformed data, and preferably a model which displays extra structure of the types mentioned above.

As an example, we consider some data extracted from those of Ries and Smith (1963), and shown in Table 1. In the original data, the quality of water had a third possible value (medium); we have omitted this in order to leave a 2^4 table. As a first step we fitted the unrestricted model, $J = I_4$; the values of the parameters are given in Table 2. This analysis amounts merely to a factorial analysis of the log counts. These data are rather unusual in that some two-variable interactions are smaller in magnitude than some three-variable interactions. This is similar to a feature detected in the complete data by Goodman (1971).

However, if we define new variables by $X'_1 = X_1 + X_2$ modulo 2 , $X'_\alpha = X_\alpha$, $\alpha = 2,3,4$, then the parameters are permuted as has been described above. The permuted parameters are given in Table 3. In the transformed table, the highest ranked three-variable interaction ranks 10th (down from 11th), and the highest ranked two-variable interaction ranks 13th (down from 14th). These changes, while not dramatic in themselves, are accompanied by others which also tend to reduce the weight of the higher order interaction terms. One simple way to measure this is by the sums of squares of the parameters, shown in Table 6. The larger parameter values have been permuted into 'interactions' involving only one variable each, that is into 'main effects'. In fact, the model in which these new variables $\underset{\sim}{X}'$ are independent fits the data well ($X^2 \simeq 13$ on 11 degrees of freedom).

However we shall examine the data a little further, to see what else may be accomplished. For example, one might wish to have the highest ranked three-variable interaction rather smaller. If we define a second transformation by $X_2'' = X_1 + X_2$, $X_\alpha'' = X_\alpha$, $\alpha = 1,3,4$, then the parameters are permuted as in Table 4. The model of complete independence is now less tenable ($X^2 \simeq 18$ on 11 degrees of freedom, between the upper 5% and 10% points). However, the model in which all main effects and all two-variable interactions are included, but no higher order terms, fits this transformation rather better. The X^2 values, each with five degrees of freedom, are 6.0 for variables $\underset{\sim}{X}$, 3.6 for variables $\underset{\sim}{X}'$ and 1.9 for variables $\underset{\sim}{X}''$. These values show that this model in fact fits the original variables adequately. However, they also illustrate the possibility of improving the fit by transformation; a similar three-fold reduction of X^2 in other data could be quite dramatic.

Goodman (1971), examining the complete table, suggested that the transformation $X_1^* = X_1 + X_2$, $X_2^* = X_2 + X_3$, $X_3^* = X_3$, $X_4^* = X_4$ would simplify the data. The resulting permuted paramters are given in

Table 5. The corresponding row of Table 6 suggests that the model with no three- or four-variable interactions should fit slightly better than for variables $\underset{\sim}{X}''$. This is borne out by a X^2 value of around 1.5 (on 5 degrees of freedom).

4. Estimation

The accepted procedure for estimating the parameters in a restricted loglinear model is that of maximum likelihood. Furthermore, it has been shown that when the model is hierarchical, this fitting may be performed by an algorithm known variously as *iterative scaling* and the *Deming-Stephan algorithm*; see for example Ireland and Kullback (1968). Thus one way to search for a suitable transformation to be applied to the data would be to perform various transformations, and then to see whether the transformed data are adequately fitted by some hierarchical model, as fitted by iterative scaling.

Fortunately there is a simpler procedure available. For as we have shown above, the loglinear model transforms in a simple way when the data are transformed linearly. Thus any model for the transformed data corresponds to a model for the original data, and hence may be fitted without performing the transformation. An example of this is given in this Section.

However, a hierarchical model for the transformed data will not in general correspond to a hierarchical model for the original data. Thus we must consider the maximum likelihood fitting of non-hierarchical models. Fortunately again, this may be achieved by using iterative scaling, although not in the usual form.

The logarithm of the likelihood function for the parameters $(\lambda_{\underset{\sim}{j}} , \underset{\sim}{j}\in J)$ of a restricted model, given observed data $(n_{\underset{\sim}{i}} , \underset{\sim}{i}\in I_d)$ is

$$\log(n!) - \sum_{\underset{\sim}{i}\in I_d} \log(n_{\underset{\sim}{i}}!) + \sum_{\underset{\sim}{i}\in I_d} n_{\underset{\sim}{i}} \log(p_{\underset{\sim}{i}}) \qquad (4.1)$$

where $n = \Sigma n_{\underset{\sim}{i}}$, and

$$\log(p_{\underset{\sim}{i}}) = \sum_{\underset{\sim}{j}\in J} \lambda_{\underset{\sim}{j}} (-1)^{\underset{\sim}{i}^\tau \underset{\sim}{j}} \ . \qquad (4.2)$$

The part of this which depends on the parameters simplifies to

$$\sum_{\underset{\sim}{j}\in J} \lambda_{\underset{\sim}{j}} \sum_{\underset{\sim}{i}\in I_d} n_{\underset{\sim}{i}} (-1)^{\underset{\sim}{i}^\tau \underset{\sim}{j}} = \sum_{\underset{\sim}{j}\in J} \lambda_{\underset{\sim}{j}} a_{\underset{\sim}{j}} \ , \qquad (4.3)$$

say. The parameters are of course subject to the constraint $\Sigma p_{\underset{\sim}{i}} = 1$.

Thus (4.3) may be maximised by the method of the undetermined multiplier. Let

$$S(\lambda_{\underset{\sim}{j}} , \underset{\sim}{j}\in J) = S(\underset{\sim}{\lambda}) = \sum_{\underset{\sim}{j}\in J} \lambda_{\underset{\sim}{j}} a_{\underset{\sim}{j}}$$

$$- \theta \sum_{\underset{\sim}{i}\in I_d} \exp\{\sum_{\underset{\sim}{j}\in J} \lambda_{\underset{\sim}{j}} (-1)^{\underset{\sim}{i}^\tau \underset{\sim}{j}}\} \ .$$

Then

$$\frac{\partial S}{\partial \lambda_{\underset{\sim}{k}}} = a_{\underset{\sim}{k}} - \theta \sum_{\underset{\sim}{i}\in I_d} (-1)^{\underset{\sim}{i}^\tau \underset{\sim}{k}} \exp\{\sum_{\underset{\sim}{j}\in J} \lambda_{\underset{\sim}{j}} (-1)^{\underset{\sim}{i}^\tau \underset{\sim}{j}}\} \ , \ \underset{\sim}{k}\in J$$

are the equations to be solved. Since $\underset{\sim}{0}\in J$, the first of these is

$$a_{\underset{\sim}{0}} = \theta \sum_{\underset{\sim}{i}\in I_d} \exp\{\sum_{\underset{\sim}{j}\in J} \lambda_{\underset{\sim}{j}} (-1)^{\underset{\sim}{i}^\tau \underset{\sim}{j}}\} \ .$$

But $a_{\underset{\sim}{0}} = n$, the total number of observations, and the sum on the right hand side is constrained to equal 1. Thus $\theta = n$, and the remaining equations to be solved are

$$\sum_{\underset{\sim}{i}\in I_d} (-1)^{\underset{\sim}{i}^\tau \underset{\sim}{k}} \exp\{\sum_{\underset{\sim}{j}\in J} \lambda_{\underset{\sim}{j}} (-1)^{\underset{\sim}{i}^\tau \underset{\sim}{j}}\} = a_{\underset{\sim}{k}}/n \ , \quad \underset{\sim}{k}\in J \ .$$

or

$$\sum_{\underset{\sim}{i}\in I_d} (-1)^{\underset{\sim}{i}^\tau \underset{\sim}{k}} p_{\underset{\sim}{i}} = a_k/n \ , \quad \underset{\sim}{k}\in J \ . \qquad (4.4)$$

A complementary set of equations, implied by $\lambda_{\underset{\sim}{j}} = 0$, $\underset{\sim}{j} \notin J$, are

$$\sum_{\underset{\sim}{i} \in I_d} (-1)^{\underset{\sim}{i}^{\tau} \underset{\sim}{j}} \log p_{\underset{\sim}{i}} = 0 , \quad \underset{\sim}{j} \notin J .$$

Let $P_{\underset{\sim}{k}}$ be the set of $\underset{\sim}{i}$ for which $(-1)^{\underset{\sim}{i}^{\tau} \underset{\sim}{k}} = 1$, and $N_{\underset{\sim}{k}}$ be the complement in I_d of $P_{\underset{\sim}{k}}$. Then (4.4) may be written

$$\sum_{\underset{\sim}{i} \in P_{\underset{\sim}{k}}} p_{\underset{\sim}{i}} - \sum_{\underset{\sim}{i} \in N_{\underset{\sim}{k}}} p_{\underset{\sim}{i}} = \frac{1}{n} \sum_{\underset{\sim}{i} \in P_{\underset{\sim}{k}}} n_{\underset{\sim}{i}} - \frac{1}{n} \sum_{\underset{\sim}{i} \in N_{\underset{\sim}{k}}} n_{\underset{\sim}{i}} ,$$

where we have rewritten $a_{\underset{\sim}{k}}$ on the right hand side of the equation in the same way as the left hand side has been rewritten. But since the sum of the two terms on each side of the equation equals one, it implies that

$$\left.\begin{aligned} \sum_{\underset{\sim}{i} \in P_{\underset{\sim}{k}}} p_{\underset{\sim}{i}} &= \frac{1}{n} \sum_{\underset{\sim}{i} \in P_{\underset{\sim}{k}}} n_{\underset{\sim}{i}} , \\ \sum_{\underset{\sim}{i} \in N_{\underset{\sim}{k}}} p_{\underset{\sim}{i}} &= \frac{1}{n} \sum_{\underset{\sim}{i} \in N_{\underset{\sim}{k}}} n_{\underset{\sim}{i}} . \end{aligned}\right\} \tag{4.5}$$

These equations have an especially simple interpretation when $\underset{\sim}{k}$ has only one entry equal to 1 . Suppose $k_{\alpha} = 1$, $k_{\beta} = 0$, $\beta \neq \alpha$. Then $P_{\underset{\sim}{k}} = (\underset{\sim}{i} \in I_d : i_{\alpha} = 0)$, and hence

$$\sum_{\underset{\sim}{i} \in P_{\underset{\sim}{k}}} p_{\underset{\sim}{i}} \tag{4.6}$$

is the probability that $X_{\alpha} = 0$. Thus the equations then state that in the fitted distribution, the probability that $X_{\alpha} = 0$ should equal the observed proportion of times that this event occurred. If the two terms on the left hand side of (4.5) are called a fitted one-variable marginal subtable, then the fitted marginal subtable for $X_{\underset{\sim}{\alpha}}$ has to coincide with the observed marginal subtable.

When $\underset{\sim}{k}$ is not of this form, the interpretation is not so clear. However, if the model is hierarchical, then it has been shown that the equations may be grouped, usually not into disjoint sets, in such a way that each group implies that a fitted subtable should coincide with an observed subtable. However, these subtables are not in general one-variable subtables, but describe the joint distribution of a number of variables. Furthermore, there exists a minimal set of such marginal subtables.

In the present context, a different interpretation is more suitable. Define a transformed variable by $Y = \underset{\sim}{X}^T \underset{\sim}{k}$ $\underline{\text{modulo}}$ 2 . Then (4.6) is the probability that $Y = 0$, and equations (4.5) state that the fitted marginal distribution of Y should equal its observed marginal distribution. When only the α'th entry of $\underset{\sim}{k}$ takes the value 1 , $Y = X_\alpha$, and hence this observation is in agreement with our earlier statement. Thus each equation in (4.4) may be interpreted as forcing a fitted one-variable subtable to coincide with its observed counter-part, except that the distribution described by the table may be that of a transformed variable.

The Deming-Stephan algorithm for solving these equations is an iterative procedure. One begins with a probability distribution belonging to the model being fitted; since the distribution which attached probability 2^{-d} to each outcome belongs to any model, this is usually chosen as the starting point. Each step of the iteration consists of a number of substeps, one for each marginal subtable which is constrained. The substep corresponding to a particular subtable consists of rescaling the distribution so as to satisfy the constraint, each of the probabilities which are summed to give one element of the subtable being rescaled by the same amount.

In fitting a hierarchical model, this procedure is applied to the minimal set of marginal subtables. For nonhierarchical models, however, this cannot be done. Since the nonhierarchical models we are interested in are hierarchical in terms of some transformed variables, one solution to this problem would be to transform the variables, that is permute the table, and then fit the corresponding hierarchical model. An alternative solution, in terms of the original variables, is to use the set of one-variable marginal subtables described above; this was the procedure used by the author. One disadvantage of this procedure is that it may fail to converge in a finite number of iterations when the more sophisticated version would, and in general it converges more slowly. To compensate for this, a modified version was used, in which instead of cycling through the set of constraints to be fitted, the one which is most seriously being violated is found, and then the fitted table is forced to fit it. It is not known how this procedure compares with the usual one when the model being fitted is, in fact, hierarchical.

The X^2 goodness of fit statistics referred to in the previous Section are calculated as minus twice the logarithm of the likelihood ratio, testing the model being fitted against the saturated model, in which all parameters are allowed to be nonzero. That is,

$$\tfrac{1}{2}X^2 = \sum_{\underset{\sim}{i}\in I_d} n_{\underset{\sim}{i}} \log n_{\underset{\sim}{i}} - n \log n - \sum_{\underset{\sim}{i}\in I_d} n_{\underset{\sim}{i}} \log p_{\underset{\sim}{i}} ,$$

where $n = \sum_{\underset{\sim}{i}\in I_d} n_{\underset{\sim}{i}}$ is the total number of observations, and $p_{\underset{\sim}{i}}$, $\underset{\sim}{i}\in I$, are the fitted probabilities under the model being tested. The last sum may be rewritten

$$\begin{aligned}\sum_{\underset{\sim}{i}\in I_d} n_{\underset{\sim}{i}} \log p_{\underset{\sim}{i}} &= \sum_{\underset{\sim}{i}\in I_d} n_{\underset{\sim}{i}} \sum_{\underset{\sim}{j}\in J} \lambda_{\underset{\sim}{j}} (-1)^{\underset{\sim}{i}^{\tau}\underset{\sim}{j}} \\ &= \sum_{\underset{\sim}{j}\in J} \lambda_{\underset{\sim}{j}} \sum_{\underset{\sim}{i}\in I_d} n_{\underset{\sim}{i}} (-1)^{\underset{\sim}{i}^{\tau}\underset{\sim}{j}} \\ &= \sum_{\underset{\sim}{j}\in J} \lambda_{\underset{\sim}{j}}\, a_{\underset{\sim}{j}} ,\end{aligned}$$

in the notation of (4.3) . Thus X^2 may be calculated from the data $n_{\underset{\sim}{i}}$, $\underset{\sim}{i}\in I$, and the fitted values of the parameters $\lambda_{\underset{\sim}{j}}$, $\underset{\sim}{i}\in J$. In the computer program used to fit the models discussed in the previous Section, the change in the value of X^2 was used both to select the parameter to be modified, and as a criterion for terminating the iteration.

Acknowledgements

The stimulus for this work was a lecture given by Professor D. R. Cox in the Spring of 1971. Gary Simon's comments have been invaluable during the subsequent development. This research was partially supported by the Office of Naval Research, under contract No. 0014-67-A-0151-0017, awarded to the Department of Statistics, Princeton University.

References

Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. In *Studies in Item Analysis and Prediction*, Solomon, H. (Ed), 158-168. Stanford University Press, Stanford, California.

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *J. Roy. Statist. Soc.* B **25**, 220-233.

Bishop, Y. M. M. (1969). Full contingency tables, logits, and split contingency tables. *Biometrics* **25**, 383-399.

Cox, D. R., and Lauh, E. (1967). A note on the graphical analysis of multidimensional contingency tables. *Technometrics* **9**, 481-488.

Deming, W. E., and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* **11**, 427-444.

Goodman, L. A. (1971). The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics* **13**, 33-63.

Ireland, C. T., and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika* **55**, 179-188.

Ries, P. N., and Smith, H. (1963). The use of chi-square for preference testing in multidimensional problems. *Chem. Eng. Progress Symposium Series* No. 42, 39-43.

Table 1

Preferences of Brand X over Brand M

		$X_2 = 0$		$X_2 = 1$	
		$X_1 = 0$	$X_1 = 1$	$X_1 = 0$	$X_1 = 1$
$X_4 = 0$	$X_3 = 0$	52	68	52	37
	$X_3 = 1$	30	42	43	24
$X_4 = 1$	$X_3 = 0$	53	63	49	57
	$X_3 = 1$	27	29	29	19

$$X_1 = \begin{cases} 0 & \text{preferred brand M} \\ 1 & \text{preferred brand X} \end{cases}$$

$$X_2 = \begin{cases} 0 & \text{previous non-user of M} \\ 1 & \text{previous user of M} \end{cases}$$

$$X_3 = \begin{cases} 0 & \text{low temperature} \\ 1 & \text{high temperature} \end{cases}$$

$$X_4 = \begin{cases} 0 & \text{hard water} \\ 1 & \text{soft water} \end{cases}$$

Table 2

Values of $\lambda_{\underset{\sim}{i}} = \lambda_{i_1, i_2, i_3, i_4}$, with rank

(in terms of absolute magnitude) in parentheses

		$i_2 = 0$		$i_2 = 1$	
		$i_1 = 0$	$i_1 = 1$	$i_1 = 0$	$i_1 = 1$
$i_4 = 0$	$i_3 = 0$	-2.8361 (16)	0.0216 (5)	0.0836 (12)	-0.1278 (14)
	$i_3 = 1$	0.2955 (15)	-0.0531 (10)	0.0148 (2)	0.0490 (8)
$i_4 = 1$	$i_3 = 0$	0.0492 (9)	0.0182 (4)	0.0174 (3)	-0,0633 (11)
	$i_3 = 1$	-0.0887 (13)	0.0313 (6)	0.0364 (7)	-0.0101 (1)

Table 3

Values of $\lambda'_{\underset{\sim}{i}} = \lambda'_{i_1, i_2, i_3, i_4}$, with rank

(in terms of absolute value) in parentheses

		$i_2 = 0$		$i_2 = 1$	
		$i_1 = 0$	$i_1 = 1$	$i_1 = 0$	$i_1 = 1$
$i_4 = 0$	$i_3 = 0$	-2.8361 (16)	-0.1278 (14)	0.0836 (12)	0.0216 (5)
	$i_3 = 1$	0.2955 (15)	0.0490 (8)	0.0148 (2)	-0.0531 (10)
$i_4 = 1$	$i_3 = 0$	0.0492 (9)	-0.0633 (11)	0.0174 (3)	0.0182 (4)
	$i_3 = 1$	-0.0887 (13)	-0.0101 (1)	0.0364 (7)	0.0313 (6)

Table 4

Values of $\underset{\sim}{\lambda''_i} = \lambda''_{i_1, i_2, i_3, i_4}$, with rank

(in terms of absolute magnitude) in parentheses

		$i_2 = 0$		$i_2 = 1$	
		$i_1 = 0$	$i_1 = 1$	$i_1 = 0$	$i_1 = 1$
$i_4 = 0$	$i_3 = 0$	-2.8361 (16)	0.0216 (5)	-0.1278 (14)	0.0836 (12)
	$i_3 = 1$	0.2955 (15)	-0.0531 (10)	0.0490 (8)	0.0148 (2)
$i_4 = 1$	$i_3 = 0$	0.0492 (9)	0.0182 (4)	-0.0633 (11)	0.0174 (3)
	$i_3 = 1$	-0.0887 (13)	0.0316 (6)	-0.0101 (1)	0.0364 (7)

Table 5

Values of $\lambda_i^* = \lambda_{i_1, i_2, i_3, i_4}^*$, with rank

(in terms of absolute magnitude) in parentheses

		$i_2 = 0$		$i_2 = 1$	
		$i_1 = 0$	$i_1 = 1$	$i_1 = 0$	$i_1 = 1$
$i_4 = 0$	$i_3 = 0$	-2.8361 (16)	-0.1278 (14)	0.0148 (2)	-0.0531 (10)
	$i_3 = 1$	0.2955 (15)	0.0490 (8)	0.0836 (12)	0.0216 (5)
$i_4 = 1$	$i_3 = 0$	0.0492 (9)	-0.0633 (11)	0.0364 (7)	0.0313 (6)
	$i_3 = 1$	-0.0887 (13)	-0.0101 (1)	0.0174 (3)	0.0182 (4)

Table 6

Sums of squares of fitted parameters,

grouped by level of interaction

	Number of variables in interaction			
	1	2	3	4
Number of such interactions	4	6	4	1
variables $\underset{\sim}{X}$	0.0972	0.0279	0.0087	0.0001
variables $\underset{\sim}{X}'$	0.1131	0.0253	0.0046	0.0010
variables $\underset{\sim}{X}''$	0.1065	0.0244	0.0016	0.0013
variables $\underset{\sim}{X}^*$	0.1063	0.0254	0.0019	0.0003