THREE-DIMENSIONAL TABLES

M. A. Toleman

So far we have only considered log-linear models for two-dimensional tables. Contingency tables of more than two-dimensions are very common. In this chapter we shall focus on three-dimensional tables and consider log-linear models for such tables. We shall also discuss the concept of collapsibility. In the past many investigators have opted to collapse over variables and examine the two-way tables so generated. Duncalfe (1980) explains that this may be a dangerous procedure and refers to Simpson's paradox. Tables of more than three-dimensions are discussed briefly. This is followed by an example of the analysis of a three-dimensional table.

1. THE GENERAL LOG-LINEAR MODEL

Consider an I x J x K contingency table. Let x_{ijk} be the observed count in the ith row (variable 1), jth column (variable 2) and kth layer (variable 3) of this table and let m_{ijk} be the corresponding expected value for that entry under some model. The general log-linear model for the table is given by

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)} \qquad (1.1)$$

where

$$\begin{aligned} \sum_i u_{1(i)} &= \sum_j u_{2(j)} = \sum_k u_{3(k)} = 0, \\ \sum_i u_{12(ij)} &= \sum_j u_{12(ij)} = \sum_i u_{13(ik)} = \sum_k u_{13(ik)} \\ &= \sum_j u_{23(jk)} = \sum_k u_{23(jk)} = 0, \qquad (1.2) \\ \sum_i u_{123(ijk)} &= \sum_j u_{123(ijk)} = \sum_k u_{123(ijk)} = 0. \end{aligned}$$

Note that the programs GLIM and GENSTAT use different constraints to those in (1.2). Both set the first level of a variable/factor to zero instead of having the levels sum to zero. This gives an identical analysis to above.

Interaction terms eg. $u_{12(ij)}$, have the same general meaning as for I x J tables ie. their presence in a model implies some dependence between the subscripted variables.

Four other types of models, corresponding to deletion of one or more terms from (1.1), are usually considered:

(a) $u_{12} = u_{13} = u_{23} = u_{123} = 0$, complete independence, ie. all three variables are independent of each other.

(b) $u_{12} = u_{13} = u_{123} = 0$, joint independence, ie. variable 1 is independent of variables 2 and 3 (there are three versions of this model).

(c) $u_{12} = u_{123} = 0$, conditional independence, ie. variables 1 and 2 are independent for all levels of variable 3 but each is associated with variable 3 (there are three versions of this model).

(d) $u_{123} = 0$, no second-order interaction, ie. no two variable interaction is affected by the third variable but all two variable interactions are present.

So far we have not considered the sampling distributions that are used for the collection of categorical data. Fienberg (1977) notes three sampling distributions, Poisson, multinomial and product-multinomial, that all give the same maximum likelihood estimates for expected cell counts. Duncalfe (1980) gives the derivation for each case.

These different sampling distributions arise because our variables are either response variables eg. disease rating, pregnancy, or design variables eg. treatment, age group. For a three-dimensional table there are three possible arrangements for our variables:

(i) three response, zero design;

(ii) two response, one design;

(iii) one response, two design.

Fienberg (1977) states that for (i) only Poisson and multinomial sampling distributions are appropriate, whereas, for (ii) and (iii) we could use a product-multinomial distribution in which the fixed marginal totals correspond to design variables.

Bishop, Fienberg and Holland (1975, p.70) show that certain interactions have to be included in models for cases (ii) and (iii). This will be discussed in more detail under Logit Models.

2. ESTIMATED EXPECTED VALUES

Consider model (a) from the previous section. We have

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} \qquad (2.1)$$

or

$$m_{ijk} = e^{u + u_{1(i)} + u_{2(j)} + u_{3(k)}} \qquad (2.2)$$

Then for all i, j and k and using the "+" notation to denote summation across all the levels of a variable we have

$$m_{i++} = e^{u + u_{1(i)}} \sum_{j,k} e^{u_{2(j)} + u_{3(k)}} \qquad (2.3)$$

$$m_{+j+} = e^{u + u_{2(j)}} \sum_{i,k} e^{u_{1(i)} + u_{3(k)}} \qquad (2.4)$$

$$m_{++k} = e^{u + u_{3(k)}} \sum_{i,j} e^{u1(i) + u_{2(j)}} \qquad (2.5)$$

$$m_{+++} = e^{u} \sum_{i,j,k} e^{u_{1(i)} + u_{2(j)} + u_{3(k)}} \qquad (2.6)$$

Dividing the product of (2.3), (2.4) and (2.5) by the square of (2.6) gives

$$m_{ijk} = \frac{m_{i++}\, m_{+j+}\, m_{++k}}{m_{+++}^{2}} \qquad (2.7)$$

with the corresponding maximum likelihood estimates given by

$$\hat{m}_{ijk} = \frac{x_{i++}\, x_{+j+}\, x_{++k}}{x_{+++}^{2}} \qquad (2.8)$$

In a similar fashion the maximum likelihood estimates for model (b) are given by

$$\hat{m}_{ijk} = \frac{x_{+jk}\, x_{i++}}{x_{+++}} \qquad (2.9)$$

and for model (c) by

$$\hat{m}_{ijk} = \frac{x_{i+k}\, x_{+jk}}{x_{++k}} \qquad (2.10)$$

All the estimates considered so far are direct estimates, ie. each $\hat{m}_{ijk}$ is a function of marginal totals. No such closed-form expression exists for the $\hat{m}_{ijk}$ for model (d). In this case an iterative procedure is necessary to evaluate the estimated expected values. Fienberg (1977, p.33-36) describes the procedure, known as the "iterative proportional fitting procedure" and adds that it will evaluate both indirect and the direct estimates.

3. HIERARCHICAL MODELS

All the models considered so far are hierarchical models in that higher order terms may be included only if related lower order terms are included. Nelder (1976) refers to this concept as "marginality". For example

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{123(ijk)}$$

with the usual constraints, is a non-hierarchical model since $u_{12(ij)}$, $u_{13(ik)}$ and $u_{23(jk)}$ are not included. The interpretation of non-hierarchical models is usually complex and hence they are avoided.

Bishop *et al* (1975) and Fienberg (1977) use the [ ] notation to represent hierarchical models, e.g.

[12] implies $\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$

[1][23] implies $\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)}$

[13][23] implies $\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)} + u_{23(jk)}$

4. COLLAPSING TABLES

The dimensionality of a contingency table may be reduced by collapsing across one or more of the variables in the table. Bishop et al (1975) state and prove the following theorem on collapsing three-dimensional tables:

Theorem 4.1 In a rectangular three-dimensional table a variable is collapsible with respect to the interaction between the other two variables if and only if it is at least conditionally independent of one of the other two variables given the third.

Thus, for example, we can measure u_{12} from the two-dimensional table of variable 1 and variable 2 if and only if $u_{13} = 0$ or $u_{23} = 0$. In this case we say the table is collapsible with respect to variable 3. Hence at least one two-factor term must be zero before we may collapse a three-dimensional table. If we collapse a table without considering these interactions then we risk drawing erroneous conclusions about the table. Simpson's paradox is explained by this theorem (Fienberg, 1977, p.45).

5. HIGHER-DIMENSIONAL TABLES

The extension of the theory of log-linear models to tables of more than three-dimensions is relatively straightforward. However, there are two practical problems that are not found with three-dimensional tables. The first problem relates to the selection of a model. The next chapter gives more detail on this but essentially the more dimensions, the more models, and then the harder it becomes to quickly pick a useful model. The second problem lies in the interpretation of a selected model. Bishop et al (1975, p.46) suggest that the verbal interpretation of many of these models is very cumbersome. They also suggest that one of the main purposes for a model is to help determine which variables can be collapsed across.

6. EXAMPLE OF THE ANALYSIS OF A THREE-DIMENSIONAL TABLE

Table 1 represents hypothetical data for 510 cows on the relationship between three variables, (1) breed of cow, (2) age of cow and (3) calf loss. There are two breeds, brahman and sahiwal, three age groups, 2 year old, 3 year old and 4 year old, and two categories of calf loss, loss and no loss.

Table 1 Frequency data on breed x age x calf loss

Breed	Age	Calf Loss Yes	No
Brahman	2	55	67
	3	16	44
	4	8	45
Sahiwal	2	48	66
	3	20	52
	4	18	71

Data in Table 2 are the expected values under two log-linear models; (a) complete independence, $u_{12} = u_{13} = u_{23} = u_{123} = 0$, (b) joint independence, $u_{12} = u_{13} = u_{123} = 0$. The estimated expected values are given by the following equations

(a) $$\hat{m}_{ijk} = \frac{x_{i++}\, x_{+j+}\, x_{++k}}{x_{+++}^{2}}$$

(b) $$\hat{m}_{ijk} = \frac{x_{+jk}\, x_{i++}}{x_{+++}}$$

where i refers to breed (2 levels), j to age (3 levels) and k to calf loss (2 levels).

Table 2 Observed values and expected values under two log-linear models for data in Table 1.

Cell (i,j,k)	Observed	Expected(a)	Expected(b)
1 , 1 , 1	55	35.18	47.46
1 , 1 , 2	67	73.56	61.28
1 , 2 , 1	16	19.68	16.59
1 , 2 , 2	44	41.15	44.24
1 , 3 , 1	8	21.17	11.98
1 , 3 , 2	45	44.26	53.45
2 , 1 , 1	48	41.17	55.54
2 , 1 , 2	66	86.08	71.72
2 , 2 , 1	20	23.03	19.41
2 , 2 , 2	52	48.15	51.76
2 , 3 , 1	18	24.77	14.02
2 , 3 , 2	71	51.80	62.55

Expected cell values are now obtained for all possible hierarchical log-linear models. The log-likelihood ratio statistic (G^2)is then calculated using these expected cell values and the observed cell frequencies. This test statistic is then compared with the 5% level of the χ^2 distribution with degrees of freedom given by

d.f. = number of cells - number of parameter in model

Table 3 lists all these models, their degrees of freedom and G^2 values.

Table 3 Log-linear model fits to data from table 1

Model	d.f.	G^2
[1][2][3]	7	37.14*
[12][3]	5	29.69*
[13][2]	6	36.82*
[1][23]	5	8.28
[12][13]	4	29.37*
[12][23]	3	0.83
[13][23]	4	7.96
[12][13][23]	2	0.83

* $P < 0.05$

The model with the fewest parameters and non-significant G^2 value is [1][23] or $u_{12} = u_{13} = u_{123} = 0$. Hence this is the simplest model that fits the data well. Estimates of the parameters for this model are given in Table 4.

Table 4 Estimates for model [1][23] using GLIM

Parameter	Estimate	Standard Error
u	3.860	0.110
$u_1(2)$	0.157	0.089
$u_2(2)$	-1.051	0.194
$u_2(3)$	-1.377	0.219
$u_3(2)$	0.256	0.131
$u_{23}(2)$	0.725	0.235
$u_{23}(3)$	1.240	0.254

Since $u_{13} = 0$ and using the theorem from Section 4 we can collapse across variable (1), "breed of cow" and produce Table 5. A later chapter on ordered categories will deal with further aspects of such tables.

Table 5 Data on age of cow x calf loss

Age	Calf Loss			
	Yes		No	
	no.	%	no.	%
2	103	20	133	26
3	36	7	96	19
4	26	5	116	23

7. REFERENCES

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975) Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass., The MIT Press.

Duncalfe, F. (1980) Some aspects of log-linear models for contingency tables. Unpublished M.Sc. St. report. University of Queensland.

Fienberg, S. E. (1977) The Analysis of Cross-Classified Categorical Data. Cambridge, Mass., The MIT Press.

Nelder, J. A. (1976) Hypothesis testing in linear models (letter to the Editor). American Statistician. 30:101.