Chapter 3

Three-Dimensional Tables

Just as a multinomial sample can be classified by the levels of two factors, a multinomial sample can also be classified by the levels of three factors.

EXAMPLE 3.0.1. Everitt (1977) considers a sample of 97 ten-year-old school children who were classified using three factors: classroom behavior, risk of home conditions, and adversity of school conditions. Classroom behavior was judged by teachers to be either nondeviant or deviant. Risk of home conditions either identify the child as not at risk (N) or at risk (R). Adversity of school condition was judged as either low, medium, or high. The observations are denoted as n_{ijk}, $i = 1, 2$, $j = 1, 2$, $k = 1, 2, 3$. The three-dimensional table of n_{ijk}'s is

		Adversity of School (k)						
		Low		Medium		High		
	Risk (j)	N	R	N	R	N	R	Total
Classroom	Nondeviant	16	7	15	34	5	3	80
Behavior (i)	Deviant	1	1	3	8	1	3	17
	Total	17	8	18	42	6	6	97

The totals at the right-hand margin are $n_{1..} = 80$ and $n_{2..} = 17$. The totals along the bottom margin are $n_{.11} = 17$, $n_{.21} = 8$, $n_{.12} = 18$, $n_{.22} = 42$, $n_{.13} = 6$, $n_{.23} = 6$, and $n_{...} = 97$.

In general, a three-dimensional table of counts is denoted n_{ijk}, $i =$

$1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$. Marginal totals are denoted

$$n_{ij\cdot} = \sum_{k=1}^{K} n_{ijk}\,, \quad n_{i\cdot k} = \sum_{j=1}^{J} n_{ijk}\,, \quad n_{\cdot jk} = \sum_{i=1}^{I} n_{ijk}\,,$$

$$n_{i\cdot\cdot} = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk} = \sum_{j=1}^{J} n_{ij\cdot} = \sum_{k=1}^{K} n_{i\cdot k}\,,$$

$$n_{\cdot j\cdot} = \sum_{i=1}^{I}\sum_{k=1}^{K} n_{ijk}\,, \quad n_{\cdot\cdot k} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ijk}\,,$$

and

$$n_{\cdots} = \sum_{ijk} n_{ijk}\,.$$

Similar notations are used for tables of probabilities p_{ijk}, tables of expected values m_{ijk}, and tables of estimates of the p_{ijk}'s and m_{ijk}'s.

Note that the values $n_{ij\cdot}$ define a two-dimensional $I \times J$ *marginal table*. The values $n_{i\cdot k}$ and $n_{\cdot jk}$ also define marginal tables.

Product-multinomial sampling is raised to a new level of complexity in three-dimensional tables. For example, we could have samples from I populations with each sample cross-classified into JK categories, or we could have samples from IJ (cross-classified) populations where each sample is classified into K categories.

Section 2 of this chapter discusses independence and odds ratio models for three-dimensional tables under multinomial sampling. Section 3 examines the iterative proportional fitting algorithm for finding estimates of expected cell counts. Section 4 introduces log-linear models for three-dimensional tables. Section 5 considers the modifications necessary for dealing with product-multinomial sampling and comments on other sampling schemes. Section 6 introduces model selection criteria and Section 7 introduces tables with four of more dimensions. We begin with a discussion of Simpson's paradox and the need for tables with more than two factors.

3.1 Simpson's Paradox and the Need for Higher-Dimensional Tables

It really is necessary to deal with three-dimensional tables; accurate information cannot generally be obtained by examining each of the three simpler two-dimensional tables. In fact, the conclusions from two-dimensional marginal tables can be contradicted by the accurate three-dimensional information. In this section, we demonstrate and examine the problem via an example.

EXAMPLE 3.1.1. Consider the outcome (success or failure) of two medical treatments classified by the sex of the patient. The data are given below.

		Patient Sex			
		Male		Female	
Outcome		Success	Failure	Success	Failure
Treatment	1	60	20	40	80
	2	100	50	10	30

Considering only the males, we have a two-way table of treatment versus outcome. The estimated probability of success under treatment 1 is $60/80 = .75$. For treatment 2, the estimated probability of success is $100/150 = .667$. Thus, for males, treatment 1 appears to be more successful.

Now consider the table of treatment versus outcome for females only. Under treatment 1, the estimated probability of success is $40/120 = .333$. Under treatment 2, the estimated probability of success is $10/40 = .25$. For women as for men, treatment 1 appears to be more successful.

Now examine the marginal table of treatment versus outcome. This is obtained by collapsing (summing) over the sexes. The table is given below.

		Outcome	
		Success	Failure
Treatment	1	100	100
	2	110	80

The estimated probability of success for treatment 1 is $100/200 = .50$, while the estimated probability of success for treatment 2 is $110/190 = .579$. The marginal table indicates that treatment 2 is better than treatment 1, whereas we know that treatment 1 is better than treatment 2 for both males and females! This contradiction is *Simpson's paradox*.

Simpson's paradox can occur because collapsing can lead to inappropriate weighting of the different populations. Treatment 1 was given to 80 males and 120 females, so the marginal table is indicating a success rate for treatment 1 that is a weighted average of the success rates for males and females with slightly more weight given to the females. Treatment 2 was given to 150 males and only 40 females, so the marginal success rate is a weighted average of the male and female success rates with most of the weight given to the male success rate. It is only a slight oversimplification to say that the marginal table is comparing a success rate for treatment 1 that is the mean of the male and female success rates, to a success rate for treatment 2 that is essentially the male success rate. Since the success rate for males is much higher than it is for females, the marginal table gives the illusion that treatment 2 is better.

The moral of all this is that one cannot necessarily trust conclusions drawn from marginal tables. It is generally necessary to consider all the dimensions of a table. Situations in which marginal (collapsed) tables yield valid conclusions are discussed in Section 5.3.

3.2 Independence and Odds Ratio Models

For multinomial sampling and a two-dimensional table, there was only one model of primary interest: independence of rows and columns. With three-dimensional tables, there are at least eight interesting models. Half of these are very easy to imagine. If we refer to the three dimensions of the table as rows, columns, and layers, we can have (0) rows, columns, and layers all independent, (1) rows independent of columns and layers (but columns and layers not necessarily independent), (2) columns independent of rows and layers, and (3) layers independent of rows and columns. Three of the remaining four models involve conditional independence: (4) given any particular layer, rows and columns are independent, (5) given any column, rows and layers are independent, and (6) given any row, columns and layers are independent. The last of the eight models is that certain odds ratios are equal. Section 4 discusses these models in relation to log-linear models.

We now examine the models in detail. The essential part of the log-likelihood for multinomial sampling is $\ell(p) \equiv \sum_{i,j,k} n_{ijk} \log(p_{ijk})$.

For all of the (conditional) independence models, the MLEs can be obtained using Lemma 2.4.1. The trick is to break $\ell(p)$ into a sum of terms, each of which can be maximized separately. The discussion below emphasizes a more general approach to finding MLEs.

3.2.1 The Model of Complete Independence

To put it briefly, the model of *complete independence* is that everything (rows, columns, and layers) is independent of everything else, cf. Example 1.1.3. Technically, the model is

$$M^{(0)}: p_{ijk} = p_{i\cdot\cdot} p_{\cdot j\cdot} p_{\cdot\cdot k}$$

where the superscript (0) is used to distinguish this model from the other models that will be considered.

The MLE of p_{ijk} under this model is

$$\begin{aligned} \hat{p}^{(0)}_{ijk} &= \hat{p}_{i\cdot\cdot} \hat{p}_{\cdot j\cdot} \hat{p}_{\cdot\cdot k} \\ &= (n_{i\cdot\cdot}/n_{\cdots})(n_{\cdot j\cdot}/n_{\cdots})(n_{\cdot\cdot k}/n_{\cdots}) \,. \end{aligned}$$

Since $m_{ijk} = n_{\cdots} p_{ijk}$, the MLE of m_{ijk} is

$$\hat{m}^{(0)}_{ijk} = n_{\cdots} \hat{p}^{(0)}_{ijk}$$

$$= \quad n_{i..}n_{.j.}n_{..k}/n_{...}^2 .$$

This is another application of the general result, discussed in Section 2.4, that the MLE of a function of the parameters is just the function applied to the MLEs. The Pearson chi-square statistic for testing lack of fit of $M^{(0)}$ is

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(n_{ijk} - \hat{m}_{ijk}^{(0)}\right)^2}{\hat{m}_{ijk}^{(0)}} .$$

The likelihood ratio test statistic is

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}\log\left(n_{ijk}/\hat{m}_{ijk}^{(0)}\right) .$$

An α level test is rejected if the test statistic is greater than $\chi^2(1-\alpha, IJK - I - J - K + 2)$. As in Section 2.4, the maximum likelihood estimate of m_{ijk} without any restrictions is $\hat{m}_{ijk} = n_{ijk}$; thus, n_{ijk} is used in the formula for G^2. Degrees of freedom for the tests given in this section will be discussed in Section 4

Chapter 10 establishes that the MLEs for $M^{(0)}$ are characterized by

$$\begin{aligned} \hat{m}_{ijk} &= n_{...}\hat{p}_{i..}\hat{p}_{.j.}\hat{p}_{..k} \\ &= \frac{\hat{m}_{i..}\hat{m}_{.j.}\hat{m}_{..k}}{n_{...}^2} \end{aligned}$$

and the marginal constraints $\hat{m}_{i..} = n_{i..}$, $\hat{m}_{.j.} = n_{.j.}$, and $\hat{m}_{..k} = n_{..k}$. In other words, any set of values $\hat{m}_{ijk}$ that satisfy the marginal constraints and satisfy model $M^{(0)}$ must be the maximum likelihood estimates. The estimates $\hat{m}_{ijk}^{(0)}$ given above are, under weak restrictions on the n_{ijk}'s, the unique values that satisfy both sets of conditions.

EXAMPLE 3.2.1. In the longitudinal study mentioned in Example 2.2.1, out of 3182 people without cardiovascular disease, 2121 neither exercised regularly nor developed cardiovascular disease during the $4\frac{1}{2}$-year study. We restrict our attention to these 2121 individuals. The subjects were cross-classified by three factors: Personality type (A,B), Cholesterol level (normal, high), and Diastolic Blood Pressure (normal, high). The data are

n_{ijk}		Diastolic Blood Pressure	
Personality	Cholesterol	Normal	High
A	Normal	716	79
	High	207	25
B	Normal	819	67
	High	186	22

The fitted values assuming complete independence are

$\hat{m}_{ijk}^{(0)}$		Diastolic Blood Pressure	
Personality	Cholesterol	Normal	High
A	Normal	739.9	74.07
	High	193.7	19.39
B	Normal	788.2	78.90
	High	206.3	20.65

Note that the $\hat{m}_{ijk}$'s satisfy the property that $n_{i\cdot\cdot} = \hat{m}_{i\cdot\cdot}$, $n_{\cdot j\cdot} = \hat{m}_{\cdot j\cdot}$, and $n_{\cdot\cdot k} = \hat{m}_{\cdot\cdot k}$ for all i, j, and k. For example, the Type A totals are $n_{1\cdot\cdot} = 716+79+207+25 = 1027$ and $\hat{m}_{1\cdot\cdot} = 739.9+74.07+193.7+19.39 = 1027.06$. The difference is roundoff error. Pearson's chi-square is

$$X^2 = \frac{(716 - 739.9)^2}{739.9} + \cdots + \frac{(22 - 20.65)^2}{20.65} = 8.730\,.$$

The likelihood ratio chi-square is

$$G^2 = 2\,[716\log(716/739.9) + \cdots + 22\log(22/20.65)] = 8.723\,.$$

The degrees of freedom for either chi-square test are

$$df = (2)(2)(2) - 2 - 2 - 2 + 2 = 4\,.$$

Since $\chi^2(.95, 4) = 9.49$, an $\alpha = .05$ level test will not reject the hypothesis of independence. In particular, the *P value* is .07. There is no clear evidence of any relationships among personality type, cholesterol level, and diastolic blood pressure level for these people who do not exercise regularly and do not have cardiovascular disease.

Although the test statistics give no clear evidence that complete independence does not hold, similarly they give no great confidence that complete independence is a good model. Deviations from independence can be examined using the Pearson residuals

$$\tilde{r}_{ijk} = \frac{n_{ijk} - \hat{m}_{ijk}}{\sqrt{\hat{m}_{ijk}}}\,.$$

The Pearson residuals for these data are

$\tilde{r}_{ijk}$		Diastolic Blood Pressure	
Personality	Cholesterol	Normal	High
A	Normal	−0.879	0.573
	High	0.956	1.274
B	Normal	1.097	−1.340
	High	−1.413	0.297

In particular, the residual for Type A, Normal, Normal is $-.879 = (716 - 739.9)/\sqrt{739.9}$. Relative to complete independence, high blood pressure and high cholesterol are overrepresented in Type A personalities (those showing signs of stress) and normal blood pressure and cholesterol are overrepresented in Type B personalities (relaxed individuals). These results agree well with conventional wisdom. Note also that for Type B personalities, individuals with only one high categorization are underrepresented.

The patterns in the residuals are interesting, but remember that there is no clear evidence (at this point) for rejecting the hypothesis of complete independence.

3.2.2 Models with One Factor Independent of the Other Two

With three factors, there are three ways in which one factor can be independent of the other two. For example, we can have rows independent of columns and layers, cf. Example 1.1.4. This model says nothing about the relationship between columns and layers. Columns and layers can either be independent or not independent. If they were not independent, typically we would be interested in examining how they differ from independence.

Specifically, the three models are rows independent of columns and layers,

$$M^{(1)}: p_{ijk} = p_{i\cdot\cdot}p_{\cdot jk} \ ,$$

columns independent of rows and layers,

$$M^{(2)}: p_{ijk} = p_{\cdot j\cdot}p_{i\cdot k} \ ,$$

and layers independent of rows and columns,

$$M^{(3)}: p_{ijk} = p_{\cdot\cdot k}p_{ij\cdot} \ .$$

All three of these models include the model of complete independence $M^{(0)}$ as a special case. If $M^{(0)}$ is true, then all three of these are true. The analyses for all three models are similar; we will consider only $M^{(1)}$ in detail.

Under $M^{(1)}$, no distinction is drawn between columns and layers. In fact, this model is equivalent to independence in an $I \times (JK)$ two-dimensional table where the columns of the two-dimensional table consist of all combinations of the columns and layers of the three-dimensional table.

Example 3.2.2. Consider again the classroom behavior data of Example 3.0.1. The test of $M^{(1)}$ is simply a test of the independence of the two rows: nondeviant, deviant, and the six columns: Low-N, Low-R, Medium-N, Medium-R, High-N, High-R.

From our results in Chapter 2, the MLE of p_{ijk} under $M^{(1)}$ is

$$\begin{aligned}\hat{p}_{ijk}^{(1)} &= \hat{p}_{i..}\hat{p}_{.jk} \\ &= (n_{i..}/n_{...})(n_{.jk}/n_{...})\,.\end{aligned}$$

The MLE of m_{ijk} is

$$\begin{aligned}\hat{m}_{ijk}^{(1)} &= n_{...}\hat{p}_{ijk}^{(1)} \\ &= n_{i..}n_{.jk}/n_{...}\,.\end{aligned}$$

The superscript (1) in $\hat{p}_{ijk}^{(1)}$ and $\hat{m}_{ijk}^{(1)}$ is used to indicate that these estimates are obtained assuming that $M^{(1)}$ is true. The Pearson chi-square test statistic for $M^{(1)}$ is

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{\left(n_{ijk}-\hat{m}_{ijk}^{(1)}\right)^2}{\hat{m}_{ijk}^{(1)}}\,.$$

The likelihood ratio test statistic is

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}\log\left(n_{ijk}/\hat{m}_{ijk}^{(1)}\right)\,.$$

These are compared to percentage points of a chi-square distribution with degrees of freedom

$$\begin{aligned}df &= (I-1)(JK-1) \\ &= IJK - I - JK + 1\,.\end{aligned}$$

Similar to $M^{(0)}$, the MLEs for model $M^{(1)}$ are any values $\hat{m}_{ijk}$ that satisfy $M^{(1)}$ and the marginal constraints

$$\hat{m}_{i..} = n_{i..} \quad \text{and} \quad \hat{m}_{.jk} = n_{.jk}\,.$$

Again, the values $\hat{m}_{ijk}^{(1)}$ given above are the unique MLEs under mild restrictions on the n_{ijk}'s. It is interesting to note that choosing $\hat{m}_{ijk} = n_{ijk}$ satisfies the marginal constraints but, except in the most bizarre cases, does not satisfy model $M^{(1)}$. On the other hand, taking the estimated cell counts from the complete independence model $\hat{m}_{ijk} = n_{i..}n_{.j.}n_{..k}/n_{...}^2$ satisfies $M^{(1)}$ but typically does not satisfy the marginal constraints.

EXAMPLE 3.2.2, CONTINUED. The table of $\hat{m}_{ijk}^{(1)}$'s is

$\hat{m}_{ijk}^{(1)}$		Adversity (k)						
		Low		Medium		High		
	Risk (j)	N	R	N	R	N	R	$\hat{m}_{i..}$
Classroom	Non.	14.02	6.60	14.85	34.64	4.95	4.95	80
Behavior (i)	Dev.	2.98	1.40	3.15	7.36	1.05	1.05	17
	$\hat{m}_{.jk}$	17	8	18	42	6	6	97

As displayed in the margins of the n_{ijk} and $\hat{m}_{ijk}^{(1)}$ tables, the MLEs satisfy the conditions $\hat{m}_{i\cdot\cdot}^{(1)} = n_{i\cdot\cdot}$ and $\hat{m}_{\cdot jk}^{(1)} = n_{\cdot jk}$.

The Pearson chi-square test statistic is

$$X^2 = 6.19 = \frac{(16 - 14.02)^2}{14.02} + \cdots + \frac{(3 - 1.05)^2}{1.05}.$$

The likelihood ratio test statistic is

$$G^2 = 5.56 = 2\left[16\log(16/14.02) + \cdots + 3\log(3/1.05)\right].$$

The degrees of freedom for the chi-square test are

$$\begin{aligned} df &= (2-1)[(2)(3)-1] \\ &= 5. \end{aligned}$$

The 95th percentile of a chi-square with 5 degrees of freedom is

$$\chi^2(.95, 5) = 11.07.$$

Both X^2 and G^2 are less than $\chi^2(.95, 5)$, so an $\alpha = .05$ level test provides no evidence against $M^{(1)}$. In other words, we have no reason to doubt that classroom behavior is independent of risk and adversity.

It is quite possible in this study that our primary interest would be in explaining classroom behavior in terms of risk and adversity. Unfortunately, classroom behavior seems to be independent of both of the variables with which we were trying to explain it. On the other hand, examining the relationship between risk and adversity becomes very simple. If classroom behavior is independent of risk and adversity, we can study the marginal table of risk and adversity without worrying about Simpson's paradox. The marginal table of counts is

$n_{\cdot jk}$		Adversity (k)			
		Low	Medium	High	$n_{\cdot j\cdot}$
Risk (j)	N	17	18	6	41
	R	8	42	6	56
	$n_{\cdot\cdot k}$	25	60	12	97

The model of independence for this marginal table is

$$M: \; p_{\cdot jk} = p_{\cdot j\cdot} p_{\cdot\cdot k}, \qquad j = 1, \ldots, J, \; k = 1, \ldots, K.$$

The expected counts for the marginal table under M are

$$\begin{aligned} \hat{m}_{\cdot jk} &= n_{\cdots} \hat{p}_{\cdot j\cdot} \hat{p}_{\cdot\cdot k} \\ &= n_{\cdot j\cdot} n_{\cdot\cdot k} / n_{\cdots}. \end{aligned}$$

The table of estimated expected counts is

		Adversity (k)			
		Low	Medium	High	$\hat{m}_{.j.}$
Risk (j)	N	10.57	25.36	5.07	41
	R	14.43	34.64	6.93	56
	$\hat{m}_{..k}$	25	60	12	97

yielding

$$X^2 = 10.78 \quad \text{and} \quad G^2 = 10.86$$

on

$$df = (2-1)(3-1) = 2\,.$$

Both X^2 and G^2 are significant at the $\alpha = .01$ level.

Either the residuals or the odds ratios can be used to explore the lack of independence. The table of residuals is

		Adversity (k)		
		Low	Medium	High
Risk (j)	N	1.98	−1.46	0.35
	R	−1.69	1.25	−0.35

For highly adverse schools, the residuals are near zero, so independence seems to hold. For schools with low adversity (i.e., good schools), the not-at-risk students are overrepresented and at-risk students are underrepresented. For schools with medium adversity, the at-risk students are overrepresented and the not-at-risk students are underrepresented. (I wonder if the criteria for determining whether a student is at risk may have been applied differently to students in high-adversity schools.)

Using odds ratios, we see that the odds of being not at risk for low-adversity schools (17/8) are about five times greater than for medium-adversity schools (18/42). In particular, the odds ratio is

$$\frac{(17)(42)}{(8)(18)} = 4.96\,.$$

The odds of being not at risk in a low-adversity school are only about twice as large as the odds of being not at risk in a high-adversity school [i.e., $(17)(6)/(8)(6) = 2.125$]. Finally, the odds of being not at risk in a medium-adversity school are only about half as large [$18(6)/42(6) = .429$] as the odds of being not at risk in a high-adversity school. Of course, the sample is small, so there is quite a bit of variability associated with these estimated odds ratios.

Before leaving this example, it is of interest to note a relationship between the two likelihood ratio test statistics that were considered. The models and statistics are

$$M^{(1)}: p_{ijk} \quad = \quad p_{i..}p_{.jk}\,, \qquad G^2 = 5.56\,, \qquad df = 5$$

$$M: p_{\cdot jk} = p_{\cdot j\cdot}p_{\cdot\cdot k}, \qquad G^2 = 10.86, \qquad df = 2.$$

Taken together, these models imply that

$$M^{(0)}: p_{ijk} = p_{i\cdot\cdot}p_{\cdot j\cdot}p_{\cdot\cdot k}$$

holds. The likelihood ratio test statistic for $M^{(0)}$ with these data is

$$G^2 = 16.42$$

with 7 degrees of freedom. As will be seen later, it is no accident that the test statistics G^2 satisfy

$$5.56 + 10.86 = 16.42$$

and that the degrees of freedom satisfy $5 + 2 = 7$.

EXERCISE 3.1. Examine the residuals from fitting $M^{(1)}$. Are any of them large enough to call in question the further analysis that was based on tentatively assuming that $M^{(1)}$ was true?

3.2.3 Models of Conditional Independence

Given that one is at a particular level of some factor, the other two factors could be independent. For example, for any given category in the levels, the rows and the columns may be independent, cf. Example 1.1.5. By the definition of conditional probability, the probability of row i and column j given that the layer is k is

$$\begin{aligned} &\Pr(\text{row} = i, \text{col} = j \mid \text{layer} = k) \\ &\quad = \Pr(\text{row} = i, \text{col} = j, \text{layer} = k) / \Pr(\text{layer} = k) \\ &\quad = p_{ijk}/p_{\cdot\cdot k}\,. \end{aligned} \tag{1}$$

Conditional independence of rows and columns for each layer means that for all i, j, and k

$$\begin{aligned} &\Pr(\text{row} = i, \text{col} = j \mid \text{layer} = k) \\ &\quad = \Pr(\text{row} = i|\text{layer} = k)\Pr(\text{col} = j|\text{layer} = k) \\ &\quad = (p_{i\cdot k}/p_{\cdot\cdot k})(p_{\cdot jk}/p_{\cdot\cdot k})\,. \end{aligned} \tag{2}$$

Assuming that every layer has a possibility of occurring (i.e., $p_{\cdot\cdot k} > 0$ for all k), then the model of conditional independence can be rewritten. Setting (1) and (2) equal and multiplying both sides by $p_{\cdot\cdot k}$ gives the requirement

$$p_{ijk} = p_{i\cdot k}\, p_{\cdot jk}/p_{\cdot\cdot k}$$

for independence of rows and columns given layers.

The nature of the conditional independence between rows and columns may or may not depend on the particular layer. For example, if rows, columns, and layers are all independent, then $M^{(0)}$ holds and for any layer k

$$\Pr(\text{row} = i,\ \text{col} = j \mid \text{layer} = k) = p_{i\cdot\cdot}p_{\cdot j\cdot}\,.$$

This does not depend on the layer. However, if rows are independent of columns and layers so that $M^{(1)}$ holds, then

$$\Pr(\text{row} = i,\ \text{col} = j \mid \text{layer} = k) = p_{i\cdot\cdot}(p_{\cdot jk}/p_{\cdot\cdot k})\,.$$

The column probabilities depend on the layer, but the row probabilities do not. Similarly, for $M^{(2)}$, the columns are independent of rows and layers; thus,

$$\Pr(\text{row} = i,\ \text{col} = j \mid \text{layer} = k) = (p_{i\cdot k}/p_{\cdot\cdot k})p_{\cdot j\cdot}\,.$$

The row structure depends on layers, but the column structure does not. Of course, the most interesting case of rows and columns independent given layers is when none of these simpler cases apply.

If two factors are to be independent given the third factor, there are three ways in which the conditioning factor can be chosen. This leads to three models: rows and columns independent given layers

$$M^{(4)}\colon p_{ijk} = p_{i\cdot k}p_{\cdot jk}/p_{\cdot\cdot k}\,,$$

rows and layers independent given columns

$$M^{(5)}\colon p_{ijk} = p_{ij\cdot}p_{\cdot jk}/p_{\cdot j\cdot}\,,$$

and columns and layers independent given rows

$$M^{(6)}\colon p_{ijk} = p_{ij\cdot}p_{i\cdot k}/p_{i\cdot\cdot}\,.$$

As in the previous subsection, the analyses for all three models are similar. We consider only $M^{(4)}$ in detail. The MLE for p_{ijk} is

$$\begin{aligned}\hat{p}^{(4)}_{ijk} &= \hat{p}_{i\cdot k}\hat{p}_{\cdot jk}/\hat{p}_{\cdot\cdot k}\\ &= (n_{i\cdot k}/n_{\cdots})(n_{\cdot jk}/n_{\cdots})\Big/(n_{\cdot\cdot k}/n_{\cdots})\\ &= n_{i\cdot k}n_{\cdot jk}/n_{\cdot\cdot k}n_{\cdots}\,.\end{aligned}$$

The MLE for $m_{ijk} = n_{\cdots}p_{ijk}$ is

$$\begin{aligned}\hat{m}^{(4)}_{ijk} &= n_{\cdots}\hat{p}_{i\cdot k}\hat{p}_{\cdot jk}/\hat{p}_{\cdot\cdot k}\\ &= n_{i\cdot k}n_{\cdot jk}/n_{\cdot\cdot k}\,.\end{aligned}$$

The MLEs are any numbers $\hat{m}_{ijk}$ that satisfy model $M^{(4)}$ and the marginal relations $\hat{m}_{i\cdot k} = n_{i\cdot k}$, $\hat{m}_{\cdot jk} = n_{\cdot jk}$, and (redundantly) $\hat{m}_{\cdot\cdot k} = n_{\cdot\cdot k}$. The test statistics are

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(n_{ijk} - \hat{m}_{ijk}^{(4)}\right)^2}{\hat{m}_{ijk}^{(4)}}$$

and

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}\log\left(n_{ijk}/\hat{m}_{ijk}^{(4)}\right) .$$

The tests actually pool the K separate tests for independence of rows and columns which are computed for each individual layer. There are $(I-1)(J-1)$ degrees of freedom for the test at each layer. The degrees of freedom for the pooled test is

$$df = (I-1)(J-1)K .$$

EXAMPLE 3.2.3. Consider again the data of Example 3.2.1. In that example, we found that the *P value* for testing complete independence of personality type, cholesterol level, and diastolic blood pressure level was .067. Our residual analysis pointed out some interesting differences between personality types. We now examine the model $M^{(6)}$ that cholesterol level and diastolic blood pressure level are independent given personality type. The test is really a simultaneous test of whether independence holds in each of the tables given below.

n_{1jk}	(Personality Type A) Diastolic Blood Pressure	
Cholesterol	Normal	High
Normal	716	79
High	207	25

n_{2jk}	(Personality Type B) Diastolic Blood Pressure	
Cholesterol	Normal	High
Normal	819	67
High	186	22

Each table has $(2-1)(2-1) = 1$ degree of freedom, so the overall test has 2 degrees of freedom. The table of estimated cell counts under conditional independence is

$\hat{m}_{ijk}^{(6)}$			Diastolic Blood Pressure	
		Cholesterol	Normal	High
Personality	A	Normal	714.5	80.51
		High	208.5	23.49
	B	Normal	813.9	72.08
		High	191.1	16.92

giving

$$X^2 = 2.188 \quad \text{and} \quad G^2 = 2.062$$

and

$$df = 2\,.$$

This is a very good fit. Given the personality type, there seems to be no relationship between cholesterol level and diastolic blood pressure level. As in the previous examples, observe that the estimates $\hat{m}_{ijk}^{(6)}$ satisfy the *likelihood equations* $\hat{m}_{i\cdot k}^{(6)} = n_{i\cdot k}$ and $\hat{m}_{ij\cdot}^{(6)} = n_{ij\cdot}$. (The likelihood equations are just the marginal constraints.)

Note that the odds of being normal in either cholesterol or blood pressure is higher for Type B personalities than for Type A personalities. If for each personality type, cholesterol and blood pressure are independent, we can examine the relationship between either personality and cholesterol or between personality and blood pressure from the appropriate marginal table, cf. Section 5.3. For example, to examine personality and cholesterol, the marginal table is

	Cholesterol	
Personality	Normal	High
A	795	232
B	886	208

The odds of having normal cholesterol for Type A personalities is $795/232 = 3.427$. The odds of having normal cholesterol for Type B personalities is $886/208 = 4.260$. The odds ratio is

$$\frac{\hat{p}_{11\cdot}\hat{p}_{22\cdot}}{\hat{p}_{12\cdot}\hat{p}_{21\cdot}} = \frac{795(208)}{232(886)} = .804\,.$$

The odds of having a normal cholesterol level with personality Type A are only about 80% as large as the odds for personality Type B.

A similar analysis shows that the odds of having a normal diastolic blood pressure level with personality Type A is 78.6% of the odds for personality Type B. Although this odds ratio of .786 is further from one than the odds ratio for cholesterol, it turns out to be less significant. The variabilities of these point estimates depend on the sample sizes in all the cells. The

personality–blood pressure marginal table has some smaller cells than the cholesterol–blood pressure table; thus, the personality–blood pressure odds ratio is subject to more variability. We will see in Example 3.4.1 that one could reasonably take personality and blood pressure to be independent, but personality and cholesterol are not independent.

3.2.4 A Final Model for Three-Way Tables

The last of the standard models for three-way tables is due to Bartlett (1935) and must be stated in terms of odds ratios. To look at an odds ratio in a three-way table, one fixes a factor and looks at the odds ratio relating the other two factors. For example, we can fix layers and look at the odds ratio $p_{11k}p_{ijk}/p_{1jk}p_{i1k}$. The last of our models is that these odds ratios are the same for every layer. In particular, the model is

$$M^{(7)}: \frac{p_{111}p_{ij1}}{p_{i11}p_{1j1}} = \frac{p_{11k}p_{ijk}}{p_{i1k}p_{1jk}}$$

for all $i = 2, \ldots, I$, $j = 2, \ldots, J$, and $k = 2, \ldots, K$. $M^{(7)}$ is stated as if layers are fixed, but, in fact, it is easily shown that the model is unchanged if stated for rows fixed or columns fixed.

There are no simple formulae for $\hat{p}_{ijk}^{(7)}$ or $\hat{m}_{ijk}^{(7)}$. Iterative computing methods (cf. Section 3) must be used to obtain the MLEs. It can be shown that the MLEs must satisfy the marginal constraints $\hat{m}_{ij\cdot} = n_{ij\cdot}$, $\hat{m}_{i\cdot k} = n_{i\cdot k}$, $\hat{m}_{\cdot jk} = n_{\cdot jk}$ and also the model; i.e., we need $\hat{m}_{111}\hat{m}_{ij1}/\hat{m}_{i11}\hat{m}_{1j1} = \hat{m}_{11k}\hat{m}_{ijk}/\hat{m}_{i1k}\hat{m}_{1jk}$ for $i, j, k \geq 2$. Given the MLEs, the test statistics are computed as usual.

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(n_{ijk} - \hat{m}_{ijk}^{(7)}\right)^2}{\hat{m}_{ijk}^{(7)}}$$

and

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}\log\left(n_{ijk}/\hat{m}_{ijk}^{(7)}\right) .$$

Although there is, at this point, no obvious reason for this figure, the degrees of freedom for the chi-square test is

$$df = (I-1)(J-1)(K-1) .$$

Example 3.2.4. Fienberg (1980) and Kihlberg, Narragon and Campbell (1964) report data on severity of drivers' injuries in auto accidents along with the type of accident and whether or not the driver was ejected from the vehicle during the accident. We consider the results only for small cars. The data are listed below.

n_{ijk}		Accident Type (k)			
		Collision		Rollover	
Injury (j)		Not Severe	Severe	Not Severe	Severe
Driver Ejected (i)	No	350	150	60	112
	Yes	26	23	19	80

Since $I = J = K = 2$, the model becomes

$$M^{(7)}: \frac{p_{111}p_{221}}{p_{121}p_{211}} = \frac{p_{112}p_{222}}{p_{122}p_{212}}.$$

Using the methods of Section 3, the MLEs of the m_{ijk}'s are found.

$\hat{m}_{ijk}^{(7)}$		Accident Type (k)			
		Collision		Rollover	
Injury (j)		Not Severe	Severe	Not Severe	Severe
Driver Ejected (i)	No	350.5	149.5	59.51	112.5
	Yes	25.51	23.49	19.49	79.51

The test statistics have $(2-1)(2-1)(2-1) = 1$ degree of freedom and are

$$X^2 = .04323 \quad \text{and} \quad G^2 = .04334\,.$$

The model of equality of odds ratios fits the data remarkably well. The reader can verify that none of the models involving independence or conditional independence fit the data.

Another way to examine $M^{(7)}$ is to look at the estimated odds ratios and see if they are about equal. For this purpose, we use the unrestricted estimates of the p_{ijk}'s, i.e., $\hat{p}_{ijk} = n_{ijk}/n_{...}$. The estimated odds ratios are

$$\begin{aligned} \hat{p}_{111}\hat{p}_{221}/\hat{p}_{121}\hat{p}_{211} &= 350(23)/26(150) \\ &= 2.064 \end{aligned}$$

and

$$\begin{aligned} \hat{p}_{112}\hat{p}_{222}/\hat{p}_{122}\hat{p}_{212} &= 60(80)/19(112) \\ &= 2.256\,. \end{aligned}$$

These values are quite close.

In summary, for both collisions and rollovers, the odds of a severe injury are about twice as large if the driver is ejected from the vehicle than if not. Equivalently, the odds of having a nonsevere injury are about twice as great if the driver is not ejected from the vehicle than if the driver is ejected. It should be noted that the odds of being severely injured in a rollover are consistently much higher than in a collision. What we have concluded in our analysis of $M^{(7)}$ is that the *relative* effect of the driver being ejected is the same for both types of accident and that being ejected substantially increases one's chances of being severely injured. So you see, it really does pay to wear seat belts.

3.2.5 Odds Ratios and Independence Models

For a two-dimensional table, Proposition 2.3.3 established that the model of independence was equivalent to the model that all odds ratios equal one. Similarly, all eight of the models discussed for three-dimensional tables can be written in terms of odds ratios. So far, only our characterization of $M^{(7)}$ is in terms of odds ratios. Since models $M^{(1)}$, $M^{(2)}$, and $M^{(3)}$ are similar, we will examine only $M^{(1)}$. Similarly, we will consider $M^{(4)}$ as representative of $M^{(4)}$, $M^{(5)}$, and $M^{(6)}$.

To begin, consider $M^{(4)}$. If $M^{(4)}$ is true, then $p_{ijk} = p_{i\cdot k}p_{\cdot jk}/p_{\cdot\cdot k}$; $M^{(4)}$ is the model that has rows and columns independent given layers. Let us consider an odds ratio with layers fixed. We can write a typical odds ratio as $p_{ijk}p_{i'j'k}/p_{ij'k}p_{i'jk}$. Assuming $M^{(4)}$ with positive p_{ijk}'s, we get

$$\begin{aligned}\frac{p_{ijk}p_{i'j'k}}{p_{ij'k}p_{i'jk}} &= \frac{(p_{i\cdot k}p_{\cdot jk}/p_{\cdot\cdot k})(p_{i'\cdot k}p_{\cdot j'k}/p_{\cdot\cdot k})}{(p_{i\cdot k}p_{\cdot j'k}/p_{\cdot\cdot k})(p_{i'\cdot k}p_{\cdot jk}/p_{\cdot\cdot k})} \\ &= \frac{p_{i\cdot k}p_{\cdot jk}p_{i'\cdot k}p_{\cdot j'k}}{p_{i\cdot k}p_{\cdot j'k}p_{i'\cdot k}p_{\cdot jk}} \\ &= 1\,.\end{aligned}$$

Thus, for a fixed value of k, the odds ratio is one.

Conversely, if $p_{ijk}p_{i'j'k}/p_{ij'k}p_{i'jk} = 1$ for all i, i', j, and j', then

$$\begin{aligned}p_{ijk}p_{\cdot\cdot k} &= \sum_{i',j'} p_{ijk}p_{i'j'k} = \sum_{i'j'} p_{ij'k}p_{i'jk} = \sum_{j'} p_{ij'k}\sum_{i'} p_{i'jk} \\ &= p_{i\cdot k}p_{\cdot jk},\end{aligned}$$

so $M^{(4)}$ holds.

It is also of interest to note that, under $M^{(4)}$, odds ratios with the row (column) fixed are equal for all rows (columns). In particular,

$$\begin{aligned}\frac{p_{ijk}p_{ij'k'}}{p_{ijk'}p_{ij'k}} &= \frac{(p_{i\cdot k}p_{\cdot jk}/p_{\cdot\cdot k})(p_{i\cdot k'}p_{\cdot j'k'}/p_{\cdot\cdot k'})}{(p_{i\cdot k'}p_{\cdot jk'}/p_{\cdot\cdot k'})(p_{i\cdot k}p_{\cdot j'k}/p_{\cdot\cdot k})} \\ &= \frac{p_{i\cdot k}p_{\cdot jk}p_{i\cdot k'}p_{\cdot j'k'}}{p_{i\cdot k'}p_{\cdot jk'}p_{i\cdot k}p_{\cdot j'k}} \\ &= \frac{p_{\cdot jk}p_{\cdot j'k'}}{p_{\cdot jk'}p_{\cdot j'k}}\,.\end{aligned} \tag{3}$$

Thus, the odds ratio does not depend on i and must be the same for each row. These facts imply that $M^{(4)}$ is a special case of $M^{(7)}$.

Perhaps the simplest way to examine odds ratios in relation to $M^{(1)}$ is to use the results obtained for $M^{(4)}$. The following proposition allows this.

Proposition 3.2.5. $M^{(1)}$ is true if and only if both $M^{(4)}$ and $M^{(5)}$ are true.

Proof. If $M^{(1)}$ is true, then $p_{ijk} = p_{i\cdot\cdot}p_{\cdot jk}$. Summing over j gives $p_{i\cdot k} = p_{i\cdot\cdot}p_{\cdot\cdot k}$; thus, $p_{i\cdot\cdot} = p_{i\cdot k}/p_{\cdot\cdot k}$. Substitution yields

$$p_{ijk} = p_{i\cdot\cdot}p_{\cdot jk} = p_{i\cdot k}p_{\cdot jk}/p_{\cdot\cdot k}\,;$$

thus, $M^{(4)}$ holds. A similar argument summing over k shows that $M^{(5)}$ holds.

Conversely, if both $M^{(4)}$ and $M^{(5)}$ are true, then

$$p_{ijk} = p_{i\cdot k}p_{\cdot jk}/p_{\cdot\cdot k}$$

and

$$p_{ijk} = p_{ij\cdot}p_{\cdot jk}/p_{\cdot j\cdot}\,.$$

It follows that

$$\frac{p_{ijk}}{p_{\cdot jk}} = \frac{p_{i\cdot k}}{p_{\cdot\cdot k}} = \frac{p_{ij\cdot}}{p_{\cdot j\cdot}}$$

and from the last equality,

$$p_{i\cdot k}p_{\cdot j\cdot} = p_{\cdot\cdot k}p_{ij\cdot}\,.$$

Summing over j gives

$$p_{i\cdot k}p_{\cdot\cdot\cdot} = p_{\cdot\cdot k}p_{i\cdot\cdot};$$

recalling that $p_{\cdot\cdot\cdot} = 1$ and rearranging terms gives

$$p_{i\cdot\cdot} = p_{i\cdot k}/p_{\cdot\cdot k}\,.$$

Substituting this into $M^{(4)}$ gives

$$\begin{aligned} p_{ijk} &= p_{i\cdot k}p_{\cdot jk}/p_{\cdot\cdot k} \\ &= p_{i\cdot\cdot}p_{\cdot jk} \end{aligned}$$

and $M^{(1)}$ holds. □

It follows from Proposition 3.2.5 and the discussion of $M^{(4)}$ that the model $M^{(1)}$ is equivalent to

$$p_{ijk}p_{i'j'k}/p_{ij'k}p_{i'jk} = 1 \qquad \text{(layers fixed)}$$

for all i, i', j, j', and k, and

$$p_{ijk}p_{i'jk'}/p_{ijk'}p_{i'jk} = 1 \qquad \text{(columns fixed)}$$

for all i, i', k, k', and j. In addition, all odds ratios with rows fixed will be equal (but not necessarily equal to one). $M^{(1)}$ is thus a special case of $M^{(7)}$.

Similar arguments establish that if $M^{(0)}$ is true, then all odds ratios equal one regardless of whether rows, columns, or layers have been fixed.

Finally, note that as in Chapter 2, odds ratios remain unchanged when the p_{ijk}'s are all replaced with m_{ijk}'s.

3.3 Iterative Computation of Estimates

To fit the model $M^{(7)}$ that all odds ratios are equal, we need to compute the $\hat{m}_{ijk}^{(7)}$'s. There are two standard algorithms for doing this: the *Newton-Raphson algorithm* and the *iterative proportional fitting algorithm*. Newton-Raphson amounts to doing a series of weighted regression analyses. It is commonly referred to as *iteratively reweighted least squares*. The Newton-Raphson algorithm is discussed in Section 10.5. Iterative proportional fitting was introduced by Deming and Stephan (1940) for purposes other than fitting models to discrete data, but the algorithm gives maximum likelihood estimates for the models discussed in the previous section and for the balanced ANOVA type log-linear models that will be discussed later. Meyer (1982) presents methods of transforming various other log-linear models so that iterative proportional fitting can be applied.

In this section, we describe the method of iterative proportional fitting for finding the $\hat{m}_{ijk}^{(7)}$'s. The method can be easily extended to find estimates for more complicated higher-dimensional tables. Under $M^{(7)}$, the $\hat{m}_{ijk}$'s are characterized by the model itself and the fitted margins $\hat{m}_{ij\cdot} = n_{ij\cdot}$, $\hat{m}_{i\cdot k} = n_{i\cdot k}$, and $\hat{m}_{\cdot jk} = n_{\cdot jk}$. The method is based on the fact that

$$1 = (n_{ij\cdot}/\hat{m}_{ij\cdot}) = (n_{i\cdot k}/\hat{m}_{i\cdot k}) = (n_{\cdot jk}/\hat{m}_{\cdot jk})$$

and, thus,

$$\begin{aligned} \hat{m}_{ijk} &= \frac{n_{ij\cdot}}{\hat{m}_{ij\cdot}}\hat{m}_{ijk}\,, \\ \hat{m}_{ijk} &= \frac{n_{i\cdot k}}{\hat{m}_{i\cdot k}}\hat{m}_{ijk}\,, \end{aligned}$$

and

$$\hat{m}_{ijk} = \frac{n_{\cdot jk}}{\hat{m}_{\cdot jk}}\hat{m}_{ijk}\,.$$

The iterative procedure begins with some initial guesses for the $\hat{m}_{ijk}$'s, say $\hat{m}_{ijk}^{[0]}$, and modifies the initial guess iteratively. Given estimates $\hat{m}_{ijk}^{[3t]}$, the modifications are

$$\begin{aligned} \hat{m}_{ijk}^{[3t+1]} &= \frac{n_{ij\cdot}}{\hat{m}_{ij\cdot}^{[3t]}}\hat{m}_{ijk}^{[3t]}\,, \\ \hat{m}_{ijk}^{[3t+2]} &= \frac{n_{i\cdot k}}{\hat{m}_{i\cdot k}^{[3t+1]}}\hat{m}_{ijk}^{[3t+1]}\,, \end{aligned}$$

and

$$\hat{m}_{ijk}^{[3(t+1)]} = \frac{n_{\cdot jk}}{\hat{m}_{\cdot jk}^{[3t+2]}}\hat{m}_{ijk}^{[3t+2]}\,.$$

The initial guesses can be any positive numbers that satisfy $M^{(7)}$. Typically, one takes

$$\hat{m}_{ijk}^{[0]} = 1 \qquad \text{for all} \qquad i,j,k\,.$$

The iterations continue until the estimates stop changing; i.e., for all i, j, k,

$$\hat{m}_{ijk}^{[3t]} \doteq \hat{m}_{ijk}^{[3t+1]} \doteq \hat{m}_{ijk}^{[3t+2]} \doteq \hat{m}_{ijk}^{[3(t+1)]} .$$

If convergence occurs to a set of values, say $\hat{m}_{ijk}$, then these must be the maximum likelihood estimates. To see this, we need to show two things: first, that $\hat{m}_{ij\cdot} = n_{ij\cdot}$, $\hat{m}_{i\cdot k} = n_{i\cdot k}$, and $\hat{m}_{\cdot jk} = n_{\cdot jk}$, and second, that the $\hat{m}_{ijk}$'s satisfy $M^{(7)}$.

Because of the nature of the iterative proportional fitting algorithm, at convergence we have

$$\hat{m}_{ijk} = \frac{n_{ij\cdot}}{\hat{m}_{ij\cdot}} \hat{m}_{ijk} ,$$

so

$$1 = \frac{n_{ij\cdot}}{\hat{m}_{ij\cdot}}$$

and

$$\hat{m}_{ij\cdot} = n_{ij\cdot} .$$

Similarly, $\hat{m}_{i\cdot k} = n_{i\cdot k}$ and $\hat{m}_{\cdot jk} = n_{\cdot jk}$.

To see that $M^{(7)}$ is satisfied, we must show that

$$\hat{m}_{111}\hat{m}_{ij1}/\hat{m}_{i11}\hat{m}_{1j1} = \hat{m}_{11k}\hat{m}_{ijk}/\hat{m}_{i1k}\hat{m}_{1jk}$$

for any values of i, j, and k greater than one. The key point here is that if $\hat{m}_{ijk}^{[3t]}$ satisfies $M^{(7)}$, then the modifications also satisfy $M^{(7)}$. Thus, if the initial values satisfy $M^{(7)}$, the result of the iterative procedure also satisfies $M^{(7)}$.

Specifically, assume that the $\hat{m}_{ijk}^{[3t]}$'s satisfy $M^{(7)}$. We will show that the $\hat{m}_{ijk}^{[3t+1]}$'s satisfy $M^{(7)}$. Since

$$\hat{m}_{ijk}^{[3t+1]} = \frac{n_{ij\cdot}}{\hat{m}_{ij\cdot}^{[3t]}} \hat{m}_{ijk}^{[3t]},$$

we have

$$\frac{\hat{m}_{111}^{[3t+1]}\hat{m}_{ij1}^{[3t+1]}}{\hat{m}_{i11}^{[3t+1]}\hat{m}_{1j1}^{[3t+1]}} = \left[\frac{\left(n_{11\cdot}\Big/\hat{m}_{11\cdot}^{[3t]}\right)\left(n_{ij\cdot}\Big/\hat{m}_{ij\cdot}^{[3t]}\right)}{\left(n_{i1\cdot}\Big/\hat{m}_{i1\cdot}^{[3t]}\right)\left(n_{1j\cdot}\Big/\hat{m}_{1j\cdot}^{[3t]}\right)}\right] \frac{\hat{m}_{111}^{[3t]}\hat{m}_{ij1}^{[3t]}}{\hat{m}_{i11}^{[3t]}\hat{m}_{1j1}^{[3t]}}$$

and

$$\frac{\hat{m}_{11k}^{[3t+1]}\hat{m}_{ijk}^{[3t+1]}}{\hat{m}_{i1k}^{[3t+1]}\hat{m}_{1jk}^{[3t+1]}} = \left[\frac{\left(n_{11\cdot}\Big/\hat{m}_{11\cdot}^{[3t]}\right)\left(n_{ij\cdot}\Big/\hat{m}_{ij\cdot}^{[3t]}\right)}{\left(n_{i1\cdot}\Big/\hat{m}_{i1\cdot}^{[3t]}\right)\left(n_{1j\cdot}\Big/\hat{m}_{1j\cdot}^{[3t]}\right)}\right] \frac{\hat{m}_{11k}^{[3t]}\hat{m}_{ijk}^{[3t]}}{\hat{m}_{i1k}^{[3t]}\hat{m}_{1jk}^{[3t]}}.$$

Since $M^{(7)}$ is satisfied for the $\hat{m}_{ijk}^{[3t]}$'s and the multipliers do not depend on k, clearly $M^{(7)}$ is satisfied for the $\hat{m}_{ijk}^{[3t+1]}$'s. Similar arguments show that the $\hat{m}_{ijk}^{[3t+2]}$'s and $\hat{m}_{ijk}^{[3(t+1)]}$'s also satisfy $M^{(7)}$, cf. Exercise 3.8.11.

In fact, iterative proportional fitting can be used to fit any of the standard models. To fit, say $M^{(6)}$, the iterative procedure chooses $\hat{m}_{ijk}^{[0]}$ to satisfy $M^{(6)}$ and then modifies estimates $\hat{m}_{ijk}^{[2t]}$ using the marginal conditions $\hat{m}_{ij\cdot} = n_{ij\cdot}$ and $\hat{m}_{i\cdot k} = n_{i\cdot k}$. Specifically,

$$\hat{m}_{ijk}^{[2t+1]} = \frac{n_{ij\cdot}}{\hat{m}_{ij\cdot}^{[2t]}} \hat{m}_{ijk}^{[2t]}$$

and

$$\hat{m}_{ijk}^{[2(t+1)]} = \frac{n_{i\cdot k}}{\hat{m}_{i\cdot k}^{[2t+1]}} \hat{m}_{ijk}^{[2t+1]} .$$

The equations for modifying the $\hat{m}$'s are determined by the marginal conditions. It is easily checked that if the sequence converges, then the $\hat{m}$'s satisfy both the marginal conditions and $M^{(6)}$. In fact, since there are closed form estimates for the $\hat{m}$'s, it takes only one set of modifications to obtain the MLEs.

It was mentioned earlier that the initial guesses are typically taken as

$$\hat{m}_{ijk}^{[0]} = 1 \qquad \text{for all} \qquad ijk .$$

The reason is that this initial guess satisfies all of the standard models. Thus, to use iterative proportional fitting for any model, one need only specify the marginal conditions and the algorithm automatically provides the estimates.

Finally, since the method is based on multiplication, any initial guess of zero will always remain zero. In other words, if for any ijk, $\hat{m}_{ijk}^{[0]} = 0$, then $\hat{m}_{ijk}^{[t]} = 0$ for all t. Thus, if it is known ahead of time that $m_{ijk} = 0$, iterative proportional fitting can handle the situation by taking $\hat{m}_{ijk}^{[0]} = 0$. On the other hand, if it is not known that $m_{ijk} = 0$, then we must choose $\hat{m}_{ijk}^{[0]} > 0$ (cf. Section 8.1).

3.4 Log-Linear Models for Three-Dimensional Tables

In a three-dimensional table for continuous data, the basic model is a three-way analysis of variance model with all interactions, i.e., $y_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)} + e_{ijk}$. For tables of counts, the same form model is used for the $\log(m_{ijk})$'s: A saturated model is written

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

We will be primarily interested in eight reduced versions of this model. Each of the eight reduced models corresponds to one of the eight independence – odds ratio models for three-dimensional tables. In particular, for

tables of positive probabilities, an odds ratio model is true if and only if the corresponding log-linear model is valid.

The eight submodels are

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)}, \tag{0}$$

$$\begin{aligned}
\log(m_{ijk}) &= u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)}, && (1)\\
\log(m_{ijk}) &= u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)}, && (2)\\
\log(m_{ijk}) &= u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)}, && (3)
\end{aligned}$$

$$\begin{aligned}
\log(m_{ijk}) &= u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)} + u_{23(jk)}, && (4)\\
\log(m_{ijk}) &= u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{23(jk)}, && (5)\\
\log(m_{ijk}) &= u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)}, && (6)
\end{aligned}$$

and

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}\,. \tag{7}$$

For $i = 0, \ldots, 7$, the log-linear model (i) holds if and only if $M^{(i)}$ holds. One way to see this is to go through a series of arguments similar to those used in Section 2.4; however, the equivalence can most easily be seen by examining odds ratios. For example, model (7) holds if and only if the $u_{123(ijk)}$ terms can be dropped from the full model. As in standard analysis of variance, the three-factor interaction terms can be dropped if and only if any set of $(I-1)(J-1)(K-1)$ linearly independent three-factor interaction contrasts are all zero. In general, a three-factor interaction contrast is

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} q_{ijk}u_{123(ijk)}$$

or, equivalently,

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} q_{ijk}\log(m_{ijk})$$

where $q_{ij\cdot} = q_{i\cdot k} = q_{\cdot jk} = 0$ for all i, j, k. The condition in $M^{(7)}$ is that

$$\frac{m_{111}m_{ij1}}{m_{i11}m_{1j1}} = \frac{m_{11k}m_{ijk}}{m_{i1k}m_{1jk}}$$

for all i, j, and k strictly greater than 1. Taking logs of both sides gives

$$\begin{aligned}
&\log(m_{111}) - \log(m_{i11}) - \log(m_{1j1}) + \log(m_{ij1})\\
&\qquad = \log(m_{11k}) - \log(m_{i1k}) - \log(m_{1jk}) + \log(m_{ijk}).
\end{aligned}$$

Rearranging terms, we see that $M^{(7)}$ is equivalent to

$$\log(m_{111}) - \log(m_{i11}) - \log(m_{1j1}) + \log(m_{ij1}) \\ - \log(m_{11k}) + \log(m_{i1k}) + \log(m_{1jk}) - \log(m_{ijk}) = 0$$

for each of the $(I-1)(J-1)(K-1)$ possible choices of i, j, and k greater than 1. All of these are three-factor interaction contrasts. Since these contrasts are linearly independent, $M^{(7)}$ holds if and only if the three-factor interaction terms can be dropped from the model, hence, if and only if model (7) holds.

Now consider $M^{(4)}$, that rows and columns are independent given layers. In terms of odds ratios, all odds ratios with any one factor fixed are equal, and all odds ratios with layers fixed equal one. To show that $M^{(4)}$ is equivalent to model (4), we need to show that $M^{(4)}$ is equivalent to having no three-factor interaction and no row-column interaction. As above, the fact that all odds ratios are equal is equivalent to having no three-factor interaction. We need to establish that all odds ratios with layers fixed equal one if and only if there is no row-column interaction. Under $M^{(4)}$, for all k, $m_{ijk}m_{i'j'k}/m_{ij'k}m_{i'jk} = 1$ or, taking logs,

$$\mu_{ijk} - \mu_{ij'k} - \mu_{i'jk} + \mu_{i'j'k} = 0$$

where $\mu_{ijk} \equiv \log m_{ijk}$. Recall that a contrast in the row-column interactions is an interaction contrast in the $\bar{\mu}_{ij\cdot}$'s. Averaging the odds ratio contrasts over k gives the row-column interaction contrast

$$\bar{\mu}_{ij\cdot} - \bar{\mu}_{ij'\cdot} - \bar{\mu}_{i'j\cdot} + \bar{\mu}_{i'j'\cdot} = 0\,.$$

If we take $i = 1$ and $j = 1$, we have $(I-1)(J-1)$ linearly independent contrasts in the row-column interaction equal to 0; the $u_{12(ij)}$ terms can be dropped from the full model. Conversely, if there is no three-factor interaction and no row-column interaction, then the contrasts $\mu_{ijk} - \mu_{ij'k} - \mu_{i'jk} + \mu_{i'j'k}$ are all equal and $\bar{\mu}_{ij\cdot} - \bar{\mu}_{ij'\cdot} - \bar{\mu}_{i'j\cdot} + \bar{\mu}_{ij\cdot} = 0$. Thus, all odds ratios are equal and those with layer fixed equal one.

Similarly, $M^{(1)}$ is true if and only if all odds ratios are equal and those with either columns or layers fixed equal one. All odds ratios equal is equivalent to no three-factor interaction; all odds ratios with layers fixed equal to one is equivalent to no row-column interaction (no $u_{12(ij)}$ terms); and all odds ratios with columns fixed equal to one is equivalent to no row-layer interaction (no $u_{13(ik)}$ terms).

Nearly all of models (0)-(7) are grossly overparametrized. For example, in model (1), the terms u, $u_{2(j)}$, and $u_{3(k)}$ are all totally redundant. The parameters $u_{1(i)}$ and $u_{23(jk)}$ are sufficient to explain everything. The u, $u_{2(j)}$, and $u_{3(k)}$ terms can take any values, yet by choosing the $u_{1(i)}$'s and $u_{23(jk)}$'s appropriately, model (1) holds. Rewriting the models in a less overparametrized fashion leads to a very convenient shorthand notation for the models

	MODEL	SHORTHAND
(0)	$\log(m_{ijk}) = u_{1(i)} + u_{2(j)} + u_{3(k)}$	[1][2][3]
(1)	$\log(m_{ijk}) = u_{1(i)} + u_{23(jk)}$	[1][23]
(2)	$\log(m_{ijk}) = u_{2(j)} + u_{13(ik)}$	[2][13]
(3)	$\log(m_{ijk}) = u_{3(k)} + u_{12(ij)}$	[3][12]
(4)	$\log(m_{ijk}) = u_{13(ik)} + u_{23(jk)}$	[13][23]
(5)	$\log(m_{ijk}) = u_{12(ij)} + u_{23(jk)}$	[12][23]
(6)	$\log(m_{ijk}) = u_{12(ij)} + u_{13(ik)}$	[12]][13]
(7)	$\log(m_{ijk}) = u_{12(ij)} + u_{13(ik)} + u_{23(jk)}$	[12][13][23]

In addition, the unrestricted (saturated) model, $\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$ can be rewritten $\log(m_{ijk}) = u_{123(ijk)}$ and abbreviated as [123].

The shorthand can also be used to remember the conditional independence interpretations of the models. For example, [1][2][3] has everything in different brackets so everything is independent. In [1][23], the rows ([1]) are in a different bracket from columns and layers ([23]); thus rows are independent of columns and layers. In [13][23], rows 1 and columns 2 are in different brackets but layers 3 is in both brackets. Thus, given layers, rows and columns are independent. No such separation of factors works for [12][13][23], and there is no interpretation in terms of independence for the corresponding model.

The shorthand identifies both the model and the margins that must be fitted to obtain MLEs. Thus, the shorthand provides all the information necessary for fitting the models using iterative proportional fitting (or Newton-Raphson). For example, [1][23] requires that the margins $\hat{m}_{i..} = n_{i..}$ and $\hat{m}_{.jk} = n_{.jk}$ be fitted and the model [12][23] requires that $\hat{m}_{ij.} = n_{ij.}$ and $\hat{m}_{.jk} = n_{.jk}$ be fitted. As discussed in Chapter 10, under mild restrictions, any values $\hat{m}_{ijk}$ that satisfy the fitted margins and the log-linear model are the MLEs. In particular, for model (1) the unique MLEs are $\hat{m}^{(1)}_{ijk} = n_{i..}n_{.jk}/n...$. These satisfy the marginal conditions and, because

$$\begin{aligned} \log\left(\hat{m}^{(1)}_{ijk}\right) &= \log(n_{i..}) + \log(n_{.jk}/n...) \\ &= \hat{u}_{1(i)} + \hat{u}_{23(jk)}, \end{aligned}$$

they satisfy the log-linear model (1).

3.4.1 Estimation

Estimation of the expected cell counts m_{ijk} has already been considered. Given the $\hat{m}_{ijk}$'s, estimation of the model parameters can proceed in a manner similar to analysis of variance.

Consider a standard ANOVA model, say

$$y_{ijk} = \xi + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + e_{ijk}$$

with $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, and the e_{ijk}'s independent $N(0, \sigma^2)$. A contrast in, for example, the $(\alpha\beta)$ interaction is determined by numbers q_{ij} $i = 1, \ldots, I$, $j = 1, \ldots, J$, that satisfy $q_{i\cdot} = q_{\cdot j} = 0$. The contrast is

$$\sum_{ij} q_{ij}(\alpha\beta)_{ij} \, .$$

The maximum likelihood estimate is

$$\sum_{ij} q_{ij}\bar{y}_{ij\cdot}$$

and

$$\text{Var}\left(\sum_{ij} q_{ij}\bar{y}_{ij\cdot}\right) = \frac{\sigma^2}{K}\sum_{ij} q_{ij}^2 \, .$$

The log-linear model

$$\log(m_{ijk}) = \xi + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik}$$

can be rewritten as

$$\mu_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} \, ,$$

where

$$\mu_{ijk} \equiv \log(m_{ijk}) \, .$$

Consider an interaction contrast

$$\sum_{ij} q_{ij} u_{12(ij)}$$

which is equivalent to

$$\sum_{ij} q_{ij}\bar{\mu}_{ij\cdot} \, .$$

The MLE of m_{ijk} is $\hat{m}_{ijk}$, so the MLE of μ_{ijk} is $\hat{\mu}_{ijk} = \log(\hat{m}_{ijk})$. Write

$$w_{ijk} \equiv \hat{\mu}_{ijk} = \log(\hat{m}_{ijk}) \, .$$

Then the estimated contrast is

$$\sum_{ij} q_{ij}\bar{w}_{ij\cdot} \, .$$

In general, the MLE of any function of the μ_{ijk}'s is just the same function applied to the $\hat{\mu}_{ijk}$'s. In particular, techniques from analysis of variance, when applied to the $\hat{\mu}_{ijk}$'s, give the estimates for contrasts in the corresponding log-linear models. In other words, whatever you would do to the

y_{ijk}'s in ANOVA to estimate a parameter, apply the same method to the $\hat{\mu}_{ijk}$'s to estimate the corresponding parameter in a log-linear model.

Unfortunately, computation of asymptotic variances is not straightforward. It requires the use of matrices and, even for contrasts, is similar in difficulty to finding the variance of a linear combination of regression coefficient estimates. Estimation is considered in detail in Section 10.2.

In the author's opinion, the most interesting aspects of estimation are those directly related to the m_{ijk}'s, odds, and odds ratios. Given the $\hat{m}_{ijk}$'s, estimates of odds and odds ratios are easy to obtain. Many examples of this have already been given. Again, the more difficult aspect of estimation is in obtaining asymptotic standard errors so that formal inferential procedures can be used.

3.4.2 Testing Models

In regression analysis it is well known that one can test a model against a larger model to see whether the smaller model is an inadequate explanation of the data. This technique is also used in analysis of variance but often it is not discussed explicitly because in balanced ANOVA it is possible to skirt the issue. For example, in a balanced ANOVA with two factors A and B and no interaction, the test for main effects in A does not depend on whether the main effects for B are included in the model. *The technique of testing models against larger models is fundamental in log-linear model analysis.* The sense in which one model is larger than another is illustrated below.

All of the tests discussed in Section 2 can be viewed as testing models against the saturated model. The test of $M^{(r)}$ was based on $\hat{m}_{ijk}^{(r)}$ and the n_{ijk}'s. The n_{ijk}'s are used because $n_{ijk} = \hat{m}_{ijk}$, where $\hat{m}_{ijk}$ is the unrestricted MLE of m_{ijk}. The unrestricted MLE of m_{ijk} is obtained by using a model that puts no restrictions on the m_{ijk}'s, namely, the saturated model

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)} . \quad (8)$$

Again, this model is grossly overparametrized; an equivalent model is

$$\log(m_{ijk}) = u_{123(ijk)} .$$

A saturated model has at least one parameter for every cell in the table, so the model always fits the data perfectly.

More generally, one can test any model against a strictly larger model. For instance, model (1)

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

can be tested against model (4)

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

because model (4) contains all of the terms in model (1) plus additional terms, the $u_{13(ik)}$'s. In other words, model (1) is a special case of model (4). The test is simply a test of whether the u_{13}'s are needed in model (4) (or equivalently a test of $M^{(1)}$ versus $M^{(4)}$).

To test [1][23] (model (1)) against [13][23] (model (4)), we use the $\hat{m}_{ijk}^{(1)}$'s, the $\hat{m}_{ijk}^{(4)}$'s, and the Pearson or likelihood ratio chi-squares. The Pearson chi-square is

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{\left(\hat{m}_{ijk}^{(4)} - \hat{m}_{ijk}^{(1)}\right)^2}{\hat{m}_{ijk}^{(1)}} .$$

The likelihood ratio chi-square is

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\hat{m}_{ijk}^{(4)}\log\left(\hat{m}_{ijk}^{(4)}/\hat{m}_{ijk}^{(1)}\right) .$$

Since this is a test of no row-layer interactions, the degrees of freedom are the degrees of freedom for row-layer interactions, i.e., $(I-1)(K-1)$, exactly as in analysis of variance.

Similarly, model (1): [1][23] can be tested against model (5): [12][23] because model (5) contains model (1). On the other hand, [1][23] cannot be tested against [12][13] because [1][23] contains $u_{23(jk)}$ terms, but [12][13] does not contain $u_{23(jk)}$ terms; thus, [12][13] is not strictly larger than [1][23]. In this case, we say that [1][23] and [12][13] are not comparable.

To perform tests, we need to be able to identify the degrees of freedom associated with each model. For standard analysis of variance type models, the degrees of freedom for a model are just the sum of the degrees of freedom for each term in the model. The degrees of freedom for terms are the same as in standard analysis of variance.

Term	Degrees of Freedom
u	1
u_1	$I-1$
u_2	$J-1$
u_3	$K-1$
u_{12}	$(I-1)(J-1)$
u_{13}	$(I-1)(K-1)$
u_{23}	$(J-1)(K-1)$
u_{123}	$(I-1)(J-1)(K-1)$

The degrees of freedom for testing [1][23] versus [13][23] are the degrees of freedom for [13][23] minus the degrees of freedom for [1][23]. Adding up the degrees of freedom for individual u terms, the degrees of freedom for [13][23] are $1+(I-1)+(J-1)+(K-1)+(I-1)(K-1)+(J-1)(K-1)$. The degrees

of freedom for [1][23] are $I + (I-1) + (J-1) + (K-1) + (J-1)(K-1)$. The degrees of freedom for the test are the difference in the model degrees of freedom, which is $(I-1)(K-1)$. As mentioned before, this is just the degrees of freedom for the terms that are in [13][23] but not in [1][23], i.e., the $u_{13(ik)}$'s.

In general, to test model (r) against model (s), where model (s) is strictly larger than model (r),

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(\hat{m}_{ijk}^{(s)} - \hat{m}_{ijk}^{(r)}\right)^2}{\hat{m}_{ijk}^{(r)}}$$

and

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \hat{m}_{ijk}^{(s)} \log\left(\hat{m}_{ijk}^{(s)}/\hat{m}_{ijk}^{(r)}\right) . \tag{9}$$

These values can be compared to a chi-square distribution. The degrees of freedom for the chi-square is the difference in the degrees of freedom for models (s) and (r).

One of the advantages of using G^2 instead of X^2 is that it simplifies the process of testing models against each other. Any of the usual tests of models can be obtained easily from the eight tests in Section 2. In fact, this is the standard way of performing tests on models. The tests in Section 2 are all tests of a reduced log-linear model against the saturated model (8). (Note that the saturated model is strictly larger than any of the other eight models.)

EXAMPLE 3.4.1. In Example 3.2.2 on classroom behavior, it was remarked that there were relationships between various likelihood ratio test statistics. Specifically, the following results were given:

Model	G^2	*df*
$M^{(0)}$: [1][2][3]	16.42	7
$M^{(1)}$: [1][23]	5.56	5
$p_{\cdot jk} = p_{\cdot j\cdot}p_{\cdot\cdot k}$	10.86	2

The test statistics for all the models are for testing against the saturated model. (In the third case, it is the saturated model for the two-dimensional table.)

The model [1][23] includes u_{23} terms in addition to the terms in [1][2][3]. Because the model [1][23] is strictly larger than [1][2][3], a test for the adequacy of the smaller model can be performed. Rather than using equation (9) directly, the test of the adequacy of the smaller log-linear model can be obtained by subtraction from the saturated model test statistics given above. Specifically, for testing [1][2][3] versus [1][23],

$$G^2 = 16.42 - 5.56 = 10.86$$

with $7 - 5 = 2$ degrees of freedom. *Not only can this be viewed as a test of the log-linear models but also as a test of the independence models, i.e., $M^{(0)}$ versus $M^{(1)}$.* Moreover, it is precisely the test given in Example 3.2.2 for

$$M: \quad p_{\cdot jk} = p_{\cdot j\cdot} p_{\cdot\cdot k}\,.$$

EXERCISE 3.2. Using the data of Example 3.2.2, compute the $\hat{m}_{ijk}^{(0)}$'s and use (9) to verify that $G^2 = 10.86$ for testing [1][2][3] versus [1][23].

We now develop these results in general. Consider testing each of models (r) and (s) against the saturated model. Recalling that for the saturated model $\hat{m}_{ijk} = n_{ijk}$, we get likelihood ratio test statistics of

$$G^2(\text{r vs. } 8) = 2\sum_{ijk} n_{ijk} \log\left(n_{ijk}/\hat{m}_{ijk}^{(r)}\right)$$

and

$$G^2(\text{s vs. } 8) = 2\sum_{ijk} n_{ijk} \log\left(n_{ijk}/\hat{m}_{ijk}^{(s)}\right)\,.$$

As shown at the end of Section 10.1, maximum likelihood estimates for log-linear models satisfy

$$\begin{aligned} G^2(\text{r vs. s}) &= 2\sum_{ijk} \hat{m}_{ijk}^{(s)} \log\left(\hat{m}_{ijk}^{(s)}/\hat{m}_{ijk}^{(r)}\right) \\ &= 2\sum_{ijk} n_{ijk} \log\left(\hat{m}_{ijk}^{(s)}/\hat{m}_{ijk}^{(r)}\right)\,. \end{aligned}$$

Given this property, it is a simple matter to check that

$$G^2(\text{r vs. s}) = G^2(\text{r vs. } 8) - G^2(\text{s vs. } 8)\,.$$

Moreover, the degrees of freedom for the tests satisfy

$$df(\text{r vs. s}) = df(\text{r vs. } 8) - df(\text{s vs. } 8)\,. \tag{10}$$

To see (10), note that (a) the degrees of freedom for the saturated model (8) are IJK, (b) $df(\text{r vs. s}) = df(\text{model (s)}) - df(\text{model (r)})$, (c) $df(\text{r vs. } 8) = IJK - df(\text{model (r)})$, and (d) $df(\text{s vs. } 8) = IJK - df(\text{model (s)})$. Substitution into (10) gives the correct result. *The methods of obtaining G^2 (r vs. s) and df (r vs. s) from G^2's and df's for testing against saturated models are basic to log-linear model practice.* Typically, computer programs only provide G^2's and df's for testing against saturated models, so reduced model tests must be constructed using this method.

In fact, this approach to testing (r) versus (s) using G^2's for the saturated model is not restricted to G^2's for the saturated model. It can be used with

any model (t) that is larger than both (r) and (s). The use of the saturated model is a convenience because it is strictly larger than *all* other models.

EXAMPLE 3.4.2. Once again, we consider the data on personality (1), cholesterol (2), and diastolic blood pressure (3) of Example 3.2.3. In the table below are given the degrees of freedom, values of X^2 and G^2, and the P *value* associated with G^2 for testing all eight of the standard models against the saturated model.

Model	*df*	X^2	G^2	P
[12][13][23]	1	0.617	0.613	.434
[12][13]	2	2.188	2.062	.358
[12][23]	2	2.985	2.980	.224
[13][23]	2	4.566	4.563	.100
[1][23]	3	7.102	7.101	.067
[2][13]	3	6.189	6.184	.102
[3][12]	3	4.543	4.601	.207
[1][2][3]	4	8.730	8.723	.067

Using a criterion of $\alpha = .05$, all of these models fit the data; however, consider testing [1][2][3] against [12][13]. The test statistic is

$$G^2 = 8.723 - 2.062 = 6.661$$

with

$$df = 4 - 2 = 2\,.$$

Because

$$\chi^2(.95, 2) = 5.99\,,$$

the model [12][13] fits significantly better than [1][2][3]. In other words, the model with cholesterol level and diastolic blood pressure level independent given personality type fits significantly better than the model of complete independence. The reader can also verify that the models [3][12] and [12][13][23] also fit significantly better than [1][2][3]. (The model [12][23] is almost significantly better than [1][2][3].) We are left with a sequence of hierarchical models [3][12], [12][13], and [12][13][23] that all fit better than complete independence. Testing [3][12] against [12][13] gives

$$\begin{aligned} G^2 &= 4.601 - 2.062 = 2.539\,, \\ df &= 3 - 2 = 1\,, \\ \chi^2(.95, 1) &= 3.84\,, \end{aligned}$$

so there seems to be no reason to take the larger model. Similarly, testing [3][12] against [12][13][23] gives

$$G^2 = 4.601 - 0.613 = 3.988$$

$$\begin{aligned} df &= 3 - 1 = 2 \\ \chi^2(.95, 2) &= 5.99\,, \end{aligned}$$

so again, the model [3][12] seems adequate. The model that posits blood pressure being independent of personality type and cholesterol is the smallest model that adequately fits the data. (It is interesting to note that based on Akaike's information criterion, cf. Section 6, model [12][13] is the best model.)

To complete the analysis, we need to examine the nature of the relationship between personality and cholesterol. This was done in Example 3.2.3. That analysis remains valid.

3.5 Product-Multinomial and Other Sampling Plans

In this section, we consider the implications of product-multinomial sampling and give a brief discussion of the effect of complex sampling plans involving stratified sampling and cluster sampling. In addition, the use of conditional distributions as a basis for statistical inference is mentioned.

Recall the data from Example 2.1.1 on opinions about legalized abortion.

	Support	Do Not Support	Total
Female	309	191	500
Male	319	281	600
Total	628	472	1100

This is product-multinomial sampling. A sample of 500 females was taken. An independent sample of 600 males was also taken. The results were combined into a 2×2 table. We consider two extensions of these data to illustrate product-multinomial sampling in three-dimensional tables.

EXAMPLE 3.5.1. Suppose that each population is further classified according to political affiliation. We might then get the table

Sex (i)	Party (j)	Opinion (k) Support	Do Not Support	Total
	Republican	79	40	119
Female	Democrat	132	71	203
	Independent	98	80	178
	Total	309	191	500
	Republican	65	94	159
Male	Democrat	141	95	236
	Independent	113	92	205
	Total	319	281	600

The totals for females and males are fixed. We know the female total is 500, so we must expect the female total to be 500. This means that

$$m_{1\cdot\cdot} = n_{1\cdot\cdot} = 500\,.$$

Similarly, for male totals,

$$m_{2\cdot\cdot} = n_{2\cdot\cdot} = 600\,.$$

More briefly, we write simply

$$m_{i\cdot\cdot} = n_{i\cdot\cdot}\,,$$

$i = 1, 2$. Any model that we fit must accommodate these facts. In other words, our estimates must satisfy the constraints

$$\hat{m}_{i\cdot\cdot} = n_{i\cdot\cdot}\,, \tag{1}$$

$i = 1, 2$. Fortunately, the estimates for all of the ANOVA type models that we have discussed satisfy this condition. Any model that includes the $u_{1(i)}$ terms (or their equivalents) will satisfy (1).

We now consider a slightly more complex sampling scheme.

EXAMPLE 3.5.2. Consider a three-factor table based on sex, socioeconomic status, and opinion about legalized abortion. Socioeconomic status has two categories: low and not low. The table of counts is

Sex (i)	Status (j)	Opinion (k) Support	Do Not Support	Total
	Low	171	79	250
Female	Not Low	138	112	250
	Total	309	191	500
	Low	152	148	300
Male	Not Low	167	133	300
	Total	319	281	600

In this table, four independent samples have been incorporated into the table. The samples are (1) a sample of 250 low-status females, (2) a sample of 250 females not of low status, (3) a sample of 300 low status males, and (4) a sample of 300 males not of low status. The sampling design has fixed the sex-status marginal totals, so the expected sex-status totals equal the observed totals, i.e., $m_{ij\cdot} = n_{ij\cdot}$. Any model that estimates expected cell counts must also incorporate the condition that

$$\hat{m}_{ij\cdot} = n_{ij\cdot}$$

for all i and j. In particular, any model that includes the $u_{12(ij)}$ terms will have these margins fixed. If we restrict attention to models that include the $u_{12(ij)}$ terms, we do not have to concern ourselves further with the product-multinomial nature of the sampling design.

The restriction that $u_{12(ij)}$ terms must be in the model reduces the possible number of models. The possible models are listed below.

Possible Models with $m_{ij\cdot}$ Fixed by the Sampling Design
[123]
[12][13][23]
[12][13]
[12][23]
[12][3]

Finally, these ideas extend easily to higher-dimensional tables. Suppose we have a four-dimensional table with indices h, i, j, k. If the sampling design fixes the margins

$$m_{h\cdot jk} = n_{h\cdot jk}\,,$$

then we restrict attention to log-linear models that include the $u_{134(hjk)}$ terms. If the sampling design fixes the margins

$$m_{\cdot i\cdot k} = n_{\cdot i\cdot k}\,,$$

then we consider only models with $u_{24(ik)}$ terms. Note that if the model includes, say, the $u_{234(ijk)}$ terms, then the $u_{24(ik)}$ terms are implicitly in the model. With the $u_{234(ijk)}$ terms in the model, the $u_{24(ik)}$ terms are redundant and it is irrelevant whether the $u_{24(ik)}$'s are explicitly stated as part of the model or not.

Examples 3.5.1 and 3.5.2 illustrate the two primary sampling schemes for response factors that were discussed in Section 2.3. In both examples, Opinion can be viewed as a response factor. In Example 3.5.2, both Sex and Status are explanatory factors. For every combination of the levels of the explanatory factors, there is an independent multinomial sample.

The categories for each multinomial are the levels of the response factor Opinion. This is the first of the sampling schemes discussed in Section 2.3. In Example 3.5.1, only the two levels of the explanatory factor Sex are used to define the independent multinomial populations. The categories of the multinomials are defined by all combinations of the levels of the response factor Opinion and the levels of the explanatory factor Party. This is the generalized sampling scheme discussed in Section 2.3. If Opinion is regarded as a response, it is not unusual to condition on all of the explanatory factors, i.e., Sex and Party, in the analysis. Thus, the data may be treated as if they were product-multinomial with an independent multinomial for each combination of the explanatory factors. These issues are discussed again in Chapter 4.

3.5.1 Other Sampling Models

As mentioned in Section 1.5, the other commonly used sampling scheme for log-linear models is Poisson sampling. In Poisson sampling, an independent Poisson random variable is observed for each cell in the table. It is easily seen that Poisson sampling leads to the same methods of analysis that are used for multinomial sampling, cf. Chapter 12.

Although Poisson, product-multinomial, and multinomial sampling are the only sampling models considered in this book, it should not be concluded that these are the only sampling schemes used to generate and analyze categorical data. The hypergeometric distribution is often taken as the appropriate sampling model. Also, most large sample surveys are not conducted so as to generate multinomial or Poisson data. Of course, these alternative sampling models may require substantial changes in the statistical analysis.

Hypergeometric sampling arises quite naturally in discrete data problems. For example, 2×2 tables in which both the row totals and the column totals are fixed can be generated by hypergeometric sampling. Hypergeometric sampling is easily generalized to $I \times J$ tables. The hypergeometric distribution is also appropriate if one conditions on the row and column totals. The reason for conditioning on the row and column totals is that they are sufficient statistics under the model of independence.

Tests for model adequacy based on the conditional distribution, given the sufficient statistics of the model, play a major role in Statistics generally and have a particularly important history in the analysis of categorical data. *Fisher's exact conditional test* for 2×2 tables, cf. Exercise 2.7.5 or Plackett (1981), may be the most famous single methodology in categorical data analysis. A key reason for using *conditional tests* is that they are appropriate even for small samples. The main problem with conditional tests is that for log-linear models, they are generally difficult to compute. McCullagh (1986) and Hirji, Mehta, and Patel (1987) use alternative approaches to examine conditional tests. McCullagh concentrates on the use of *asymp-*

totic conditional distributions with the idea that conditional asymptotics are more appropriate for small and moderate sample sizes than unconditional asymptotics. Hirji et al. *enumerate* all of the tables that have the same values for the sufficient statistics. This work also applies to logistic regression. While enumeration is a very demanding computational task, with modern algorithms and computing equipment it has become a realistic alternative to the methods discussed here. Haberman (1974a, p. 14-33) details conditional inference for categorical data. Balmer (1988), Bedrick and Hill (1990), Mehta and Patel (1980, 1983), Mehta, Patel, and Gray (1985), Mehta, Patel, and Tsiatis (1984), and Pagano and Taylor-Halvorson (1981) all address issues of conditional inference and table enumeration. Recent reviews of these methods have been given by Agresti (1992) and Mehta (1994). The computer programs StatXact and/or LogXact perform the necessary computations. Davison (1988) uses *saddlepoint methods* to approximate conditional distributions. In some cases, random samples of the tables can be used rather than enumerating all of the tables. Kreiner (1987) discusses model selection using conditional tests and, specifically, the use of random samples from the conditional distribution. The generation of such random samples is based on the work of Agresti, Wackerly and Boyett (1979), Boyett (1979), and Patefield (1981). Berkson (1978) and Kempthorne (1979) give alternative views of Fisher's exact test. For a general discussion of conditional tests see Lehmann (1986).

Large sample surveys typically involve the use of *stratification* and *cluster sampling*, cf. Kish (1965). Multinomial sampling corresponds to simple random sampling with replacement. Product-multinomial sampling involves independent samples on a number of subpopulations. This is just stratified sampling. As we have seen, if strata are included as a factor in the table, then stratified sampling causes no problems in a log-linear model analysis of the data. The difficulty with stratified sampling is that often the individual strata are not of interest in the analysis. The desired conclusions are for the population as a whole and must be arrived at by weighting results from the separate strata. In sampling theory, the point of selecting strata is to reduce the variability of the overall results. If this is accomplished and if data from a stratified sample are analyzed as though they are multinomial, the variability is overestimated and results appear to be less significant than they actually are.

The incorporation of cluster sampling is fundamentally more difficult to deal with. The whole point of cluster sampling is that the observations within a cluster are not independent. Typically, they display a positive correlation. All of our standard sampling plans assume independence between individual observations, so the standard plans are clearly inappropriate for cluster sampling. Inferences based on independence will underestimate the variability of data with a positive correlation. Thus, analyzing cluster sampling data as if they were multinomial will typically overstate the significance of results.

In a *complex survey* involving both stratification and cluster sampling, the tendencies to understate significance and to overstate significance will, to some extent, offset each other. While this is a positive sign for the standard multinomial analysis, it by no means ensures that any particular set of survey data can be analyzed accurately when the complex sampling structure is ignored. In all likelihood, one or the other tendency to misstate the variability will dominate. Any serious analysis of complex survey data must involve an evaluation of the effect of the survey design on the analysis. Some of the key references on the analysis of complex survey data are Koch, Freeman, and Freeman (1975), Fienberg (1979), Brier (1980), Holt, Scott, and Ewings (1980), Rao and Scott (1981, 1984, 1987), Bedrick (1983), Binder et al. (1984), Gross (1984), Fay (1985), and Thomas and Rao (1987). The collection of papers edited by Skinner, Holt, and Smith (1989) provides a useful summary of methods for analyzing data from complex surveys, including categorical data.

3.6 Model Selection Criteria

In analysis of variance and regression, the three measures most commonly used to evaluate the fit of models are R^2, Adjusted R^2, and Mallows' C_p. Each of these measures has natural analogues in log-linear models. R^2 measures how much of the total variation is being explained by the model. R^2 has the property that models must explain as much or more of the variation than their submodels (models in which some terms have been deleted). Adjusted R^2 modifies the definition of R^2 so that larger models are penalized for being larger. Mallows' C_p statistic is related to Akaike's information criterion. We will apply Akaike's information criterion to model selection for log-linear and logistic regression models and discuss the relation of Akaike's criterion to Mallows' C_p.

Discussions of model selection criteria in general settings are given by Akaike (1973) and Schwarz (1978). A good review and comparison is presented in Clayton, Geisser, and Jennings (1986).

3.6.1 R^2

In standard regression analysis, R^2 is defined as

$$R^2 = \frac{\text{SSReg}}{\text{SSTot} - \text{C}}$$

where SSReg is the sum of squares for regression and SSTot−C is the sum of squares total corrected for the grand mean. In fact, SSTot−C is just the error sum of squares for the model that includes only an intercept. If we denote SSE(X) as the error sum of squares for an arbitrary model called

X (e.g., with design matrix X) and $\text{SSE}(X_0)$ as the error sum of squares for a model with only an intercept, then

$$R^2 = \frac{\text{SSE}(X_0) - \text{SSE}(X)}{\text{SSE}(X_0)} .$$

$\text{SSE}(X_0)$ is the total variation and $\text{SSE}(X_0) - \text{SSE}(X)$ is the variability explained by the model X. The ratio of these two, R^2, is the proportion of the total variation explained by the model.

In general, there is no reason that $\text{SSE}(X_0)$ has to be the sum of squares for a model with just an intercept. In general, $\text{SSE}(X_0)$ could be the error sum of squares for the smallest interesting model. In regression analysis, the smallest interesting model is almost always the model with only an intercept. In log-linear models, the smallest interesting model may well be the model of complete independence. (Recall from Exercise 2.4 that the independence model for a two-way table is also the intercept-only model for logistic regression.)

In log-linear models, G^2 plays a role similar to that of SSE in regression. If X_0 indicates the smallest interesting model and X indicates the log-linear model of interest, we define

$$R^2 = \frac{G^2(X_0) - G^2(X)}{G^2(X_0)}$$

where $G^2(X)$ and $G^2(X_0)$ are the likelihood ratio test statistics for testing models X and X_0 against the saturated model.

If the X_0 model is the smallest interesting model, then $G^2(X_0)$ is a measure of the total variability in the data. (It tests X_0 against a model that fits the data perfectly.) It follows that $G^2(X_0) - G^2(X)$ measures the variability explained by the X model. R^2 is the proportion of the total variability explained by the X model. Alternative definitions of R^2 are available.

As in standard regression analysis, R^2 cannot be used to compare models that have different numbers of degrees of freedom. In regression analysis, this is caused by the fact that larger models have larger R^2's. Exactly the same phenomenon occurs with log-linear models. In fact, R^2 for the saturated model will always equal one because G^2 for the saturated model is zero.

3.6.2 Adjusted R^2

Having defined R^2 for log-linear models, the same adjustment for model size used in standard regression analysis can be used for log-linear models. The *adjusted* R^2 is

$$\text{Adj. } R^2 = 1 - \frac{q - r_0}{q - r}[1 - R^2]$$

where q is the number of cells in the table and r and r_0 are the degrees of freedom for the models X and X_0. Note that there are $q-r$ degrees of freedom for testing X against the saturated model and $q-r_0$ degrees of freedom for testing X_0.

A little algebra shows that

$$\text{Adj. } R^2 = 1 - \frac{G^2(X)/(q-r)}{G^2(X_0)/(q-r_0)} .$$

A large value of Adj. R^2 indicates that the model X fits well. The largest value of Adj. R^2 will occur for the model X with the smallest value of $G^2(X)/(q-r)$. Just as in regression analysis, the Adj. R^2 criterion suggests the inclusion of many (probably too many) explanatory terms.

3.6.3 Akaike's Information Criterion

We now consider Akaike's information criterion as a method for selecting log-linear models. After describing Akaike's method, we demonstrate its close relationship to standard regression model selection based on Mallow's C_p statistic.

Akaike (1973) proposed a criterion of the information contained in a statistical model. He advocated choosing the model that maximizes this information. For log-linear models, maximizing *Akaike's information criterion* (AIC) amounts to choosing the model X that *minimizes*

$$A_X = G^2(X) - [q - 2r] , \qquad \text{(log-linear)}$$

where $G^2(X)$ is the likelihood ratio test statistic for testing the X model against the saturated model, r is the number of degrees of freedom for the X model, and there are q degrees of freedom for the saturated model, i.e., q cells in the table.

Given a list of models to be compared along with their G^2 statistics and the degrees of freedom for the tests, a slight modification of A_X is easier to compute by hand.

$$\begin{aligned} A_X - q &= G^2(X) - 2[q-r] \\ &= G^2(X) - 2 \text{ (test degrees of freedom)} . \end{aligned}$$

Because q does not depend on the model X, minimizing $A_X - q$ is equivalent to minimizing A_X. Note that for the saturated model, $A - q = 0$.

Before continuing our discussion of the AIC, we give an example of the use of A, R^2, and Adj. R^2.

Example 3.6.1. For the personality (1), cholesterol (2), blood pressure (3) data of Examples 3.2.1 and 3.2.3, testing models against the saturated model gives

Model	df	G^2	$A-q$	R^2	Adj. R^2
[12][13][23]	1	0.613	−1.387	.885	.719
[12][13]	2	2.062	−1.938	.764	.527
[12][23]	2	2.980	−1.020	.658	.318
[13][23]	2	4.563	0.563	.477	−.046
[1][23]	3	7.101	1.101	.186	−.085
[2][13]	3	6.184	0.184	.291	.055
[3][12]	3	4.602	−1.398	.472	.297
[1][2][3]	4	8.723	0.723	0	0

In Example 3.4.2, we established that there were three eligible models: [3][12], [12][13], and [12][13][23] and that model [12][23] was almost eligible. The AIC criterion $A-q$ easily picks out all four of these models, with [12][13] the best of them. The adjusted R^2 criterion also identifies these four models, but the values seem rather strange. For example, [12][23], which is not significantly better than [1][2][3], has a higher Adj. R^2 than [3][12], which is significantly better than [1][2][3]. The R^2 values seem like reasonable measures. The author's inclination is to use the AIC, cf. Clayton, Geisser, and Jennings (1986).

With only three factors, it is easy to look at all possible models. Model selection criteria become more important when dealing with tables having more factors.

Relation to Mallows' C_p

The approach to using Akaike's information criterion outlined above is quite general. If we are considering a collection of models indexed by ξ, we can denote individual models as M_ξ. For any ξ, let p_ξ be the dimension of the parameter space of the model. This is the number of independent parameters in the model. For models with linear structure, this is the degrees of freedom for the model (rank of the design matrix) plus the number of any independent nonlinear parameters. Log-linear models do not involve any nonlinear parameters. Standard regression and ANOVA involve one nonlinear parameter, the variance σ^2.

Suppose that there exists a most general model M with s independent parameters. In other words, any model M_ξ is just a special case of M. For log-linear models, M is the saturated model and s is the number of cells in the table. For selecting variables in standard regression analysis, M is the full model that includes an intercept plus all $s-1$ available variables. Finally, let $\Lambda(\xi)$ be the likelihood ratio test statistic for testing M_ξ against the larger model M. Maximizing Akaike's information criterion is equivalent to choosing ξ to minimize

$$A_\xi = \Lambda(\xi) - (s - 2p_\xi) .$$

Consider applying this to the problem of variable selection in regression analysis. Let $\text{SSE}(F)$ be the sum of squares for error of the full model and $\text{SSE}(X)$ be the error sum of squares for a reduced model with design matrix X and $\text{rank}(X) = p$. Let s be the degrees of freedom for the full model. *If we assume that the variance σ^2 is known*, then

$$A_X = \frac{\text{SSE}(X) - \text{SSE}(F)}{\sigma^2} - (s - 2p) \; . \qquad \text{(regression)}$$

Because σ^2 will not really be known, an ad hoc procedure would be to estimate σ^2 with $\hat{\sigma}^2 = \text{SSE}(F)/(n-s)$, the mean squared error for the full model where n is the regression sample size. An estimate of A_X is

$$\begin{aligned} \hat{A}_X &= \frac{\text{SSE}(X)}{\hat{\sigma}^2} - \frac{\text{SSE}(F)}{\hat{\sigma}^2} - (s - 2p) \\ &= \frac{\text{SSE}(X)}{\hat{\sigma}^2} - (n - s) - (s - 2p) \\ &= \frac{\text{SSE}(X)}{\hat{\sigma}^2} - (n - 2p) \\ &= C_p \end{aligned}$$

where C_p is Mallows' well-known criterion for selecting regression models, cf. Christensen (1996b, Section 14.1).

3.7 Higher-Dimensional Tables

Log-linear models can easily be extended to tables with more than three factors. All of the basic principles from three-dimensional tables continue to apply. However the models, as well as independence and odds ratio relationships, become more complex. Independence relationships for high-dimensional tables are discussed in Chapter 5. With more factors, there are many more models to consider. Systematic methods of model selection are discussed in Chapter 6. In this section, we just illustrate some examples.

EXAMPLE 3.7.1. A study was performed on mice to examine the relationship between two drugs and muscle tension. For each mouse, a muscle was identified and its tension was measured. A randomly chosen drug was given to the mouse and the muscle tension was measured again. The muscle was then tested to identify which type of muscle it was. The weight of the muscle was also measured. Factors and levels are tabulated below.

Factor	Abbreviation	Levels
Change in Muscle Tension	T	High, Low
Weight of Muscle	W	High, Low
Muscle	M	Type 1, Type 2
Drug	D	Drug 1, Drug 2

The sampling is product-multinomial with the total count for each muscle type fixed. The data are

Tension(h)	Weight(i)	Muscle(j)	Drug(k) 1	Drug 2
High	High	Type 1	3	21
		Type 2	23	11
	Low	Type 1	22	32
		Type 2	4	12
Low	High	Type 1	3	10
		Type 2	41	21
	Low	Type 1	45	23
		Type 2	6	22

For illustration, we fit three log-linear models to this four-factor table: the model of all main effects

$$\log(m_{hijk}) = \gamma + \tau_h + \omega_i + \mu_j + \delta_k, \tag{1}$$

the model of all two-factor interactions

$$\begin{aligned}\log(m_{hijk}) &= \gamma + \tau_h + \omega_i + \mu_j + \delta_k + (\tau\omega)_{hi} + (\tau\mu)_{hj} + (\tau\delta)_{hk} \\ &\quad + (\omega\mu)_{ij} + (\omega\delta)_{ik} + (\mu\delta)_{jk},\end{aligned} \tag{2}$$

and the model of all three-factor interactions

$$\begin{aligned}\log(m_{hijk}) &= \gamma + \tau_h + \omega_i + \mu_j + \delta_k + (\tau\omega)_{hi} + (\tau\mu)_{hj} + (\tau\delta)_{hk} \\ &\quad + (\omega\mu)_{ij} + (\omega\delta)_{ik} + (\mu\delta)_{jk} \\ &\quad + (\tau\omega\mu)_{hij} + (\tau\omega\delta)_{hik} + (\tau\mu\delta)_{hjk} + (\omega\mu\delta)_{ijk}.\end{aligned} \tag{3}$$

Getting rid of some of the redundant parameters, these can be rewritten as

$$\log(m_{hijk}) = \tau_h + \omega_i + \mu_j + \delta_k, \tag{1}$$

$$\log(m_{hijk}) = (\tau\omega)_{hi} + (\tau\mu)_{hj} + (\tau\delta)_{hk} + (\omega\mu)_{ij} + (\omega\delta)_{ik} + (\mu\delta)_{jk}, \tag{2}$$

and

$$\log(m_{hijk}) = (\tau\omega\mu)_{hij} + (\tau\omega\delta)_{hik} + (\tau\mu\delta)_{hjk} + (\omega\mu\delta)_{ijk} \tag{3}$$

respectively, leading to the shorthand notations

$$[\mathrm{T}][\mathrm{W}][\mathrm{M}][\mathrm{D}], \tag{1}$$

$$[\mathrm{TW}][\mathrm{TM}][\mathrm{WM}][\mathrm{TD}][\mathrm{WD}][\mathrm{MD}], \tag{2}$$

$$[TWM][TWD][TMD][WMD]. \tag{3}$$

As discussed in Section 4, the shorthand provides all the information necessary for fitting the model (other than the actual cell counts).

The test statistics for testing these models against the saturated model are given below. Clearly, the only model that fits the data is the model of all three-factor interactions.

Model	*df*	G^2	*P*
[TWM][TWD][TMD][WMD]	1	0.11	.74
[TW][TM][WM][TD][WD][MD]	5	47.67	.00
[T][W][M][D]	11	127.4	.00

As before, we can also test any model against reduced models. For example the test of [TWM][TWD][TMD][WMD] versus the reduced model [TW][TM][WM][TD][WD][MD] has $G^2 = 47.67 - 0.11 = 47.56$ on $df = 5 - 1 = 4$.

The data of Example 3.7.1 and the following data will be used to illustrate techniques in subsequent chapters.

EXAMPLE 3.7.2. Consider a data set in which there are four factors defining a $2 \times 2 \times 3 \times 6$ table. The factors are

Factor	Abbreviation	Levels
Race	R	White, Nonwhite
Sex	S	Male, Female
Opinion	O	Yes = Supports Legalized Abortion
		No = Opposed to Legalized Abortion
		Und = Undecided
Age	A	18-25, 26-35, 36-45, 46-55, 56-65, 66+ years

The sex and opinion factors are reminiscent of Example 2.1.1, but the data are distinct. The data are given in Table 3.1. See also Exercise 3.8.10.

3.7.1 COMPUTER COMMANDS

The muscle tension data are listed in file 'tension.dat' as counts for each cell with indices for tension, weight, muscle type, and drug, respectively. The file is as given below.

```
 3 1 1 1 1
21 1 1 1 2
23 1 1 2 1
```

TABLE 3.1. Abortion Opinion Data

Race	Sex	Opinion	Age					
			18-25	26-35	36-45	46-55	56-65	66+
White	Male	Yes	96	138	117	75	72	83
		No	44	64	56	48	49	60
		Und	1	2	6	5	6	8
	Female	Yes	140	171	152	101	102	111
		No	43	65	58	51	58	67
		Und	1	4	9	9	10	16
Nonwhite	Male	Yes	24	18	16	12	6	4
		No	5	7	7	6	8	10
		Und	2	1	3	4	3	4
	Female	Yes	21	25	20	17	14	13
		No	4	6	5	5	5	5
		Und	1	2	1	1	1	1

```
11 1 1 2 2
22 1 2 1 1
32 1 2 1 2
 4 1 2 2 1
12 1 2 2 2
 3 2 1 1 1
10 2 1 1 2
41 2 1 2 1
21 2 1 2 2
45 2 2 1 1
23 2 2 1 2
 6 2 2 2 1
22 2 2 2 2
```

We can fit the log-linear model [WMD][TWM][TWD][TMD] using SAS PROC GENMOD.

```
options ps=60 ls=72 nodate;
data tension;
   infile 'tension.dat';
   input n T W M D;
proc genmod data=tension;
   class T W M D;
   model n = W*M*T T*W*M T*W*D T*M*D / link=log
                                       dist=poisson;
```

```
run;
```

The main differences between these commands and those given in Subsection 2.6.1 for logistic regression are that now "link=log" and "dist=poisson". These change GENMOD from fitting logistic regression to fitting log-linear models. To fit other specific models such as [TM][WM][MD] or [T][WM][D], the model statement uses T*M W*M M*D or T W*M D, respectively. The "class" command used above specifies that a variable is not acting like a predictor variable in regression but rather that it gives indices for specifying the levels of an analysis of variance type factor.

Similarly, we can use GENMOD to fit the abortion data. The data file 'abort.dat' has five columns, the first four are indices for race, sex, age, and opinion. The last column has the counts for each cell. The SAS commands for fitting the model [RSO][RSA][ROA][SOA] are

```
options ps=60 ls=72 nodate;
data abort;
   infile 'abort.dat';
   input R S A O N;
proc genmod data=abort;
   class R S A O;
   model N = R*S*O R*S*A R*O*A S*O*A / link=log
                                       dist=poisson;
run;
```

GLIM uses commands that are similar to GENMOD. Interactions are specified with a period rather than an asterisk. GLIM begins by specifying the number of cells in the table, i.e., the "units."

```
$units 72$
$data r s a o n$
$factor r 2 s 2 a 6 o 3$
$dinput 6$
$yvar n$
$error poisson$
$fit r.s.o + r.s.a + r.o.a + s.o.a$
$display e$
$stop$
```

The "factor" command specifies that a variable is not acting like a predictor variable in regression but rather that it gives indices for specifying the levels of an analysis of variance type factor. For GLIM, the user needs to specify the number of levels for each factor. After the 'dinput 6' command, the DOS version of GLIM prompts the user for the name of the data file.

GENMOD and GLIM use the Newton-Raphson algorithm. BMDP-4F uses iterative proportional fitting. For the model [RSO][RSA][ROA][SOA],

the BMDP-4F commands are as follows:

```
/ INPUT     FILE = 'C:\LOGLIN\ABORT.DAT'.
          FORMAT = FREE.
          VARIABLES = 5.
/ VARIABLE  NAMES = R, S, A, O, N.
/ TABLE     INDICES = R, S, A, O.
            COUNT = N.
/ STAT      ALL.
/ FIT       MODEL = RSO, RSA, ROA, SOA.
/ PRINT     LINE = 79.
/ END
```

BMDP-4F is the most powerful program I am aware of for fitting analysis of variance type log-linear models. In addition to GENMOD, SAS has a procedure called CATMOD. CATMOD will not be discussed. (I'll leave the reasons to your imagination.)

3.8 Exercises

EXERCISE 3.8.1. Complete an analysis similar to that of Example 3.4.2 for the classroom behavior data of Example 3.0.1.

EXERCISE 3.8.2. Complete an analysis similar to that of Example 3.4.2 for the auto accident data of Example 3.2.4.

EXERCISE 3.8.3. Radelet (1981) gives data on the relationship between race and the imposition of the death penalty. The data are given in Table 3.2. Analyze the data.

TABLE 3.2. Race and the Death Penalty

Defendant's Race	Victim's Race	Death Penalty Yes	No
Black	Black	6	97
	White	11	52
White	Black	0	9
	White	19	132

EXERCISE 3.8.4. The data on graduate admissions at Berkeley given in Exercise 2.6.1 was actually collapsed over the six largest departments within the university. The possibility exists that the data may display Simpson's

TABLE 3.3. Graduate Admissions at Berkeley

	Male		Female	
Dept.	Admitted	Rejected	Admitted	Rejected
A	512	313	89	19
B	353	207	17	8
C	120	205	202	391
D	138	279	131	244
E	53	138	94	299
F	22	351	24	317

paradox. The full data are given in Table 3.3. Analyze the three-dimensional table and comment on Simpson's paradox relative to these data.

EXERCISE 3.8.5. Discuss Simpson's paradox in terms of the following probability inequalities.

$$\Pr(A|B \text{ and } C) < \Pr(A|\text{ not } B \text{ and } C),$$

$$\Pr(A|B \text{ and not } C) < \Pr(A|\text{ not } B \text{ and not } C),$$

and

$$\Pr(A|B) > \Pr(A|\text{ not } B).$$

EXERCISE 3.8.6. Reevaluate your analysis of the data discussed in Exercise 2.6.3 in light of Simpson's paradox. Are there other factors that need to be accounted for in a correct analysis of these data?

EXERCISE 3.8.7. For the data of Example 3.2.4, do the first step of the iterative proportional fitting algorithm for $\hat{m}_{ijk}^{(7)}$ using a hand calculator. Use starting values of $\hat{m}_{ijk}^{[0]} = 1$. Compare the results after one step to the fully iterated estimates.

EXERCISE 3.8.8. Consider the model $\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)}$.

(a) Show that the maximum likelihood estimate of $u_{3(1)} - u_{3(2)}$ is $\log(n_{..1}) - \log(n_{..2})$.

(b) Show that maximum likelihood estimation gives

$$\log\left[\frac{\hat{u}_{12(11)}\hat{u}_{12(22)}}{\hat{u}_{12(12)}\hat{u}_{12(21)}}\right] = \log(n_{11.}) - \log(n_{12.}) - \log(n_{21.}) + \log(n_{22.}).$$

EXERCISE 3.8.9. *The Mantel-Haenszel Statistic.*
In biological and medical applications, it is not uncommon to be confronted with a series of 2×2 tables that examine the same effect under different

conditions. If there are K such tables, the data can be combined to form a $2 \times 2 \times K$ table. Because each 2×2 table examines the same effect, it is often assumed that the odds ratio for the effect is constant over tables. This is equivalent to assuming the no three-factor interaction model. To test for the existence of the effect, one tests whether the common log odds ratio is zero while adjusting for the various circumstances under which data were collected. In terms of log-linear models, this is a one degree of freedom test of conditional independence given the layer k. Prior to the development of log-linear model theory, Mantel and Haenszel (1959) proposed a statistic for testing this hypothesis. The statistic, apart from a continuity correction factor, is

$$\frac{\left[\sum_k (n_{11k} - \hat{m}_{11k})\right]^2}{\sum_k \left[\hat{m}_{11k}\hat{m}_{22k}\right] \Big/ \left[n_{\cdot\cdot k} - 1\right]},$$

where the $\hat{m}$'s are obtained from the conditional independence model. This statistic has an asymptotic $\chi^2(1)$ distribution under the conditional independence model.

The Berkeley graduate admission data of Exercise 3.8.4 and Table 3.3 is a set of six 2×2 tables. In each table we are interested in the effect of sex on admission; the six departments constitute various conditions under which this effect is being investigated.

a) Give a justification for whether or not use of the Mantel-Haenszel statistic is appropriate for these data.

b) If appropriate, use both G^2 and the Mantel-Haenszel statistic to test whether there is an effect of sex on admission.

c) Show that the denominator of the Mantel-Haenszel statistic can be written as $\sum_k \left[\hat{m}_{12k}\hat{m}_{21k}\right] / \left[n_{\cdot\cdot k} - 1\right]$.

EXERCISE 3.8.10. Using the data of Table 3.1 fit the all main effects model, the all two-factor effects model, and the all three-factor effects model. Perform all of the tests possible among these three models. Discuss your results.

EXERCISE 3.8.11. With regard to Section 3, show that the $\hat{m}_{ijk}^{[3t+2]}$'s and $\hat{m}_{ijk}^{[3(t+1)]}$'s also satisfy $M^{(7)}$.

EXERCISE 3.8.12. As can be seen from the iterative proportional fitting algorithm, the $\hat{m}$'s for the model of no three-factor interaction depend only on the 3 two-dimensional marginal tables. Discuss how this fact can be used to develop a more complete analysis for the Gilby data of Exercise 2.7.3. What assumptions must be made and what techniques should be used? What problems will Standard VIII cause?

PROJECT 3.8.13. Write a computer program to fit the model of no three-factor interaction to the Gilby data of Exercise 2.6.3. This can be done

using iterative proportional fitting. Assuming that this model fits, do any submodels fit adequately?