

2 Structural Models for Counted Data

2.1 Introduction

As soon as a problem is clearly defined, its solution is often simple. In this chapter we show how complex qualitative data may be described by a mathematical model. Questions that the data were designed to answer may then be stated precisely in terms of the parameters of the model.

In multivariate qualitative data each individual is described by a number of attributes. All individuals with the same description are enumerated, and this count is entered into a cell of the resulting contingency table. Descriptive models with as many independent parameters as the table has cells are called “saturated.” They are useful in reducing complexity only if the parameters can be readily interpreted as representing “structural” features of the data, because most of the questions of importance may be interpreted as being questions about the data structure.

The complexity of the data is reflected by the number of parameters in the model describing its structure. When the structure is simple, the model has few parameters. Whenever the model has fewer parameters than the number of data cells, we say that the model is “unsaturated.” For some unsaturated models we can reduce the number of cells in the table without distorting the structure. Such reduction we refer to as “collapsing” and we give theorems defining those structures that are collapsible. Before proceeding to describe models for the simplest four-cell table, we enlarge on this concept of structure and on the development and uses of models.

2.1.1 Structure

If every individual in the population under study can be classified as falling into one and only one of t categories, we say that the categories are mutually exclusive and exhaustive. A randomly selected member of the population will fall into one of the t categories with probability p_i , where $\{p_i\}$ is the vector of cell probabilities

$$\{p_i\} = (p_1, p_2, \dots, p_t) \quad (2.1-1)$$

and

$$\sum_{i=1}^t p_i = 1.$$

Here the cells are strung out into a line for purposes of indexing only; their arrangement and ordering does not reflect anything about the characteristics of individuals falling into a particular cell. The p_i reflect the relative frequency of each category in the population.

When the cells are defined in terms of the categories of two or more variables, a structure relating to the nature of the data is imposed. The natural structure for two variables is often a rectangular array with columns corresponding to the categories of one variable and rows to categories of the second variable; three variables create layers of two-way tables, and so on. As soon as this structure is imposed, the position of the cells tells us something about the characteristics of individuals falling into them: For instance, individuals in a specific cell have one characteristic in common with individuals in all cells of the same row, and another characteristic in common with all individuals in cells in the same column. A good mathematical model should reflect this structure.

As soon as we consider more than one randomly selected individual we must consider the sampling plan. If the second and all subsequent individuals are sampled “with replacement,” that is, the first is replaced in the population before the second is randomly drawn, and so on, then the vector of probabilities (2.1-1) is unchanged for each individual. Alternatively, the vector of probabilities is unchanged if the population is infinitely large. In either of these circumstances, if we take a simple random sample of size N , we obtain a sample of counts $\{x_i\}$ such that

$$\{x_i\} = (x_1, x_2, \dots, x_t), \quad (2.1-2)$$

where

$$\sum x_i = N.$$

The corresponding expected counts are $\{m_i\}$, such that

$$\{m_i\} = (m_1, m_2, \dots, m_t), \quad (2.1-3)$$

where

$$\begin{aligned} E(x_i) &= m_i \quad \text{for } i = 1, \dots, t, \\ m_i &= Np_i. \end{aligned}$$

In Chapter 3 we deal with estimating the $\{m_i\}$ from the $\{x_i\}$ under a variety of sampling schemes and for different models. In Chapter 13 we consider different sampling distributions and the relationships between the $\{m_i\}$ and the $\{p_i\}$. In this chapter we are not concerned with the effects of sampling, but only with the underlying data structure. Thus we are interested in models that specify relationships among the cell probabilities $\{p_i\}$ or among the expected counts $\{m_i\}$. Some sampling schemes impose restrictions on the $\{m_i\}$, and so we also discuss situations where these constraints occur without considering how they arise. The constraints occur in situations where we are in effect taking several samples, each drawn from one segment of the population. We then have probabilities for each segment summing to 1, but we cannot relate probabilities in different segments to the population frequency in different segments unless we know the relative size of the segments.

2.1.2 Models

The smallest rectangular table is based on four cells, and the saturated model describing it has four independent parameters. In Section 2.2 we give a four-term

model for this table that is linear in the logarithmic scale, and we give an interpretation of each of the four terms. In Section 2.3 we extend this four-term model to larger two-dimensional tables by enlarging the number of parameters encompassed by each term of the model.

Log-linear models are not new; they are implicit in the conventional χ^2 test for independence in two-way contingency tables. The notation of Birch [1963] is convenient for such models, as the number of terms depends on the dimension and the interdependencies between dimensions, rather than on the number of cells. Each term encompasses as many parameters as are needed for the total number of independent parameters in the saturated model to equal the number of cells in the table. When the model is unsaturated, the reduction is generally achieved by removing one or more terms completely, because the terms rather than the parameters correspond to effects of interest. In Section 2.4 we show that an s -dimensional table of any size is described by a model with 2^s terms. Thus the models reflect the structure imposed on the data, and the terms are closely related to hypotheses of interest.

2.1.3 *Uses of structural models*

The interpretation of the terms of saturated models that fully specify an array leads to interpretation of models with fewer terms. The investigator faced with data of an unknown structure may wish to determine whether they are fitted well by a particular unsaturated model, that is he may wish to test a particular hypothesis. Alternatively, he may wish to obtain good estimates for some or all of the cells and may obtain such estimates by fitting an unsaturated model. Using unsaturated models to obtain stable cell estimates is akin to fitting an approximate response curve to quantitative data; the investigator gains knowledge of important underlying trends by reducing the number of parameters to less than that required for perfect fit. Thus comprehension is increased by focusing on the most important structural features.

If the data can be described by models with few terms, it may be possible to condense the data without either obscuring important structural features or introducing artifactual effects. Such condensation is particularly pertinent when the data are sparse relative to the magnitude of the array. In addition to focusing on parameter and model interpretation, we look in each section of this chapter at the problem of determining when such condensation is possible without violating important features of the underlying structure.

In this chapter we do not discuss fitting models; we discuss procedures that yield maximum likelihood estimates in Chapter 3 and assessment of goodness of fit in Chapter 4. The concern here is with such questions as:

1. What do we mean by “independence” and “interaction”?
2. Why is it necessary to look at more than two dimensions at a time?
3. How many variables should be retained in a model and which can safely be removed?

2.2 Two Dimensions—The Fourfold Table

The simplest contingency table is based on four cells, and the categories depend on two variables. The four cells are arranged in a 2×2 table whose two rows correspond to the categorical variable A and whose two columns correspond to

the second categorical variable B . We consider first the different constraints that we may use to specify the cell probabilities, then the effect of rearranging the cells. This leads to formulation of a model, the log-linear model, that we can readily interpret in terms of the constraints and apply to any arrangement of the four cells.

We discuss features of the log-linear model for the 2×2 table in detail. Important features that also apply to larger tables are:

1. Only one parameter of the model is changed when it is used to describe expected cell counts m instead of probabilities p ;
2. the model is suitable for a variety of sampling schemes;
3. the ready interpretability of the terms of the model is not shared by models that are linear in the arithmetic scale.

In Section 2.7 we give a geometric interpretation of the 2×2 table and show how the parameters of the log-linear model are related to the structural features of a three-dimensional probability simplex.

2.2.1 Possible constraints for one arrangement

Double subscripts refer to the position of the cells in our arrangement. The first subscript gives the category number of variable A , the second of variable B , and the two-dimensional array is displayed as a grid with two rows and two columns:

$$\begin{array}{cc}
 & \begin{array}{c} B \\ 1 \quad 2 \end{array} \\
 \begin{array}{c} A \\ 1 \\ 2 \end{array} & \begin{array}{|c|c|} \hline p_{11} & p_{12} \\ \hline p_{21} & p_{22} \\ \hline \end{array}
 \end{array} \quad (2.2-1)$$

We consider first a simple random sample such that the cell probabilities sum to 1, that is, we have the linear constraint

$$\sum_{i=1}^2 \sum_{j=1}^2 p_{ij} = 1. \quad (2.2-2)$$

By displaying the cells as in expression (2.2-1), we introduce a structure to the corresponding probabilities, and it is natural for us to examine the row and column marginal totals:

$$p_{i+} = \sum_{k=1}^2 p_{ik} \quad i = 1, 2 \quad (2.2-3)$$

$$p_{+j} = \sum_{k=1}^2 p_{kj} \quad j = 1, 2. \quad (2.2-4)$$

These totals give the probabilities of an individual falling in categories i and j of variables A and B , respectively. (Throughout this book, when we sum over a subscript we replace that subscript by a “+”.) At this point, we can expand our

tabular display (2.2-1) to include the marginal totals and the basic constraint (2.2-2):

		<i>B</i>		
		1	2	Totals
<i>A</i>	1	p_{11}	p_{12}	p_{1+}
	2	p_{21}	p_{22}	p_{2+}
Totals		p_{+1}	p_{+2}	1

(2.2-5)

The marginal probabilities p_{i+} and p_{+j} are the unconditional probabilities of belonging to category i of variable A and category j of variable B , respectively. Each set of marginal probabilities must sum to 1. As we have only two categories, once we know one row total, p_{1+} , we also know the other row total, p_{2+} , because $p_{2+} = 1 - p_{1+}$, and similarly for column totals. Thus if we know the values of p_{1+} and p_{+1} , the two linear constraints on the marginal probabilities lead to a complete definition of all the marginal probabilities. We need only one further constraint involving the internal cells to specify completely the structural relationships of the table.

We refer to the internal cells as “elementary” cells. The probability p_{ij} is the probability of an individual being in category i of variable A and category j of variable B . Most questions of interest related to the fourfold table are concerned with differences between such internal probabilities and the marginal probabilities. A variety of functions of the probabilities are commonly used, and others can readily be devised, that will produce the further constraint needed for complete specification of the table. Commonly used are:

- (i) the difference in column proportions

$$\frac{p_{11}}{p_{+1}} - \frac{p_{12}}{p_{+2}};$$

- (ii) the difference in row proportions

$$\frac{p_{11}}{p_{1+}} - \frac{p_{21}}{p_{2+}};$$

- (iii) the cross-product ratio

$$\alpha = \frac{p_{11}p_{22}}{p_{12}p_{21}}.$$

A natural choice if we wish to continue to use linear constraints is:

- (iv) the diagonal sum

$$S_d = p_{11} + p_{22}.$$

Finally, we can choose:

- (v) the ratio of an elementary cell probability to the product of its row and column probabilities

$$\frac{p_{11}}{p_{1+}p_{+1}}.$$

Other measures appear in Chapter 11. Specifying the value of any one of the five statistics in this list is equivalent to specifying the remaining four, given p_{1+} and p_{+1} . Such specification completely determines the values of the cell probabilities $\{p_{ij}\}$. The third function, α , has desirable properties not possessed by the others. We consider its properties in detail because they lead us to the formulation of our model for the fourfold table.

Properties of the cross-product ratio

Since the rows of the table correspond to one variable, A , and the columns to a second variable, B , it is natural for us to be interested in the relationship between these underlying categorical variables. We first consider the behavior of the statistics (i)–(v) under independence. If the state of A is independent of the state of B , then

$$p_{ij} = p_{i+}p_{+j} \quad i = 1, 2; \quad j = 1, 2; \quad (2.2-6)$$

but this relationship is not satisfied for all i and j if A and B are dependent.

As any of the functions, when combined with the marginal totals, completely specify the table, they also measure dependence between the underlying variables. For instance, the independence relationship (2.2-6) is equivalent to stating that the proportional difference (i) or (ii) is 0, or that the measure (v) has the value 1. The measure (iv) becomes a less attractive function of the marginal probabilities, namely,

$$S_d = 1 - p_{+1} - p_{1+} + 2p_{1+}p_{+1}.$$

When we focus on the product relationship (2.2-6), it is reasonable for us to choose the cross-product ratio instead of the linear functions. The cross-product ratio α , like measure (v), attains the value 1 when the condition of independence holds, and it has two properties not possessed by measure (v), or any of the other measures:

1. α is invariant under the interchange of rows and columns;
2. α is invariant under row and column multiplications. That is, suppose we multiply the probabilities in row 1 by $r_1 > 0$, those in row 2 by $r_2 > 0$, those in column 1 by $c_1 > 0$, and those in column 2 by $c_2 > 0$. Then we get

$$\alpha = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{(r_1c_1p_{11})(r_2c_2p_{22})}{(r_1c_2p_{12})(r_2c_1p_{21})}. \quad (2.2-7)$$

This result holds regardless of whether we normalize so that the new cell entries sum to 1. An important implication is that we obtain the same value of α when we use either the cell probabilities or the expected counts in each cell.

Interpretation of cross-product ratio

The cross-product ratio α is also known as the “odds ratio.” For the first level of variable A , the odds on being in the first level of variable B are p_{11}/p_{12} , and for

the second level of variable A they are p_{21}/p_{22} . The cross-product ratio is the ratio of these odds,

$$\alpha = \frac{p_{11}/p_{12}}{p_{21}/p_{22}} = \frac{p_{11}p_{22}}{p_{12}p_{21}}.$$

This definition is also invariant under interchange of the variables. It should not be confused with another measure used by epidemiologists, the relative risk r , defined as the ratio of the row proportion $p_{11}/(p_{11} + p_{12})$ to the corresponding row proportion $p_{21}/(p_{21} + p_{22})$. Thus we have

$$r = \frac{p_{11}(p_{21} + p_{22})}{p_{21}(p_{11} + p_{12})} = \frac{p_{11}p_{2+}}{p_{21}p_{1+}}. \quad (2.2-8)$$

We can define r similarly in terms of column proportions, but then we obtain a different value. The relative risk does not have the invariance properties possessed by the relative odds, although its interpretation when dealing with the risk of contracting disease in two population groups makes it a useful parameter.

The logarithm of the relative odds is also a linear contrast of the log-probabilities of the four elementary cells, namely

$$\log \alpha = \log p_{11} - \log p_{12} - \log p_{21} + \log p_{22}, \quad (2.2-9)$$

and when $\log \alpha = 0$ we have independence between variables A and B .

The cross-product ratio and bivariate distributions

As soon as we consider the cross-product ratio as a measure of departure from independence, the question of its relationship to the correlation coefficient arises. Mosteller [1968] takes bivariate normals with different selected values of ρ and shows that the value of α differs according to the breaking point chosen. Thus α is not easily related to ρ for bivariate normals, but Plackett [1965] shows that it is possible to construct a class of distributions where the value of α is unchanged by the choice of breaking point.

2.2.2 Effect of rearranging the data

Suppose that the two underlying variables A and B for the 2×2 table actually have the same categories and simply represent measurements on one variable at two points in time. We can then refer to them as A_1 and A_2 . There are three different arrays that may be of interest:

1. the basic table

$$\begin{array}{cc}
 & A_2 \\
 & 1 \quad 2 \\
 A_1 \quad 1 & \begin{array}{|c|c|} \hline p_{11} & p_{12} \\ \hline \end{array} \\
 2 & \begin{array}{|c|c|} \hline p_{21} & p_{22} \\ \hline \end{array}
 \end{array} \quad (2.2-10)$$

2. the table measuring changes from the first measurement to the second. This table preserves the margins for the first variable:

		Same	Different	
	1	p_{11}	p_{12}	
A_1	2	p_{22}	p_{21}	(2.2-11)

3. the table measuring changes going back from the second measurement to the first. This table preserves the margins for the second variable:

		A_2	
		1	2
Same	p_{11}	p_{22}	
Different	p_{21}	p_{12}	(2.2-12)

For each of these 2×2 tables we have a cross-product ratio. Taking tables (2.2-10)–(2.2-12) in order, these ratios are

$$\alpha_3 = \frac{p_{11}p_{22}}{p_{12}p_{21}}, \quad (2.2-13)$$

$$\alpha_2 = \frac{p_{11}p_{21}}{p_{12}p_{22}}, \quad (2.2-14)$$

$$\alpha_1 = \frac{p_{11}p_{12}}{p_{22}p_{21}}. \quad (2.2-15)$$

The reason for this ordering of the subscripts will become apparent shortly. For the moment we note that these three expressions suggest a class of structural models based on α_3 , α_2 , and α_1 , rather than on one of the cross products together with the margins of one of the tables.

Taking logarithms of the $\{\alpha_i\}$, we get three linear contrasts

$$\log \alpha_3 = \log p_{11} - \log p_{12} - \log p_{21} + \log p_{22}, \quad (2.2-16)$$

$$\log \alpha_2 = \log p_{11} - \log p_{12} + \log p_{21} - \log p_{22}, \quad (2.2-17)$$

$$\log \alpha_1 = \log p_{11} + \log p_{12} - \log p_{21} - \log p_{22}. \quad (2.2-18)$$

If we specify values for these three contrasts and recall that

$$\sum p_{ij} = 1, \quad (2.2-19)$$

we have completely defined the four cell probabilities. This formulation suggests that we look for a model that is linear in the log scale.

2.2.3 The log-linear model

A simple way to construct a linear model in the natural logarithms of the cell

probabilities is by analogy with analysis of variance (ANOVA) models. We write

$$\log p_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \quad i = 1, 2; j = 1, 2, \quad (2.2-20)$$

where u is the grand mean of the logarithms of the probabilities :

$$u = \frac{1}{4}(\log p_{11} + \log p_{12} + \log p_{21} + \log p_{22}), \quad (2.2-21)$$

$u + u_{1(i)}$ is the mean of the logarithms of the probabilities at level i of first variable :

$$u + u_{1(i)} = \frac{1}{2}(\log p_{i1} + \log p_{i2}) \quad i = 1, 2, \quad (2.2-22)$$

and similarly for the j th level of the second variable :

$$u + u_{2(j)} = \frac{1}{2}(\log p_{1j} + \log p_{2j}) \quad j = 1, 2. \quad (2.2-23)$$

Since $u_{1(i)}$ and $u_{2(j)}$ represent deviations from the grand mean u ,

$$u_{1(1)} + u_{1(2)} = u_{2(1)} + u_{2(2)} = 0. \quad (2.2-24)$$

Similarly, $u_{12(ij)}$ represents a deviation from $u + u_{1(i)} + u_{2(j)}$, so that

$$u_{12(11)} = -u_{12(12)} = -u_{12(21)} = u_{12(22)}. \quad (2.2-25)$$

We note that the additive properties (2.2-24) and (2.2-25) imply that each u -term has one absolute value for dichotomous variables. Thus we introduce no ambiguity by writing, for instance, $u_1 = 0$ without specifying the second subscript.

If we define $l_{ij} = \log p_{ij}$, then by analogy with ANOVA models we can write the grand mean as

$$u = \frac{l_{++}}{4} = \sum_{i,j} \frac{l_{ij}}{4}. \quad (2.2-26)$$

Similarly, the main effects are

$$u_{1(i)} = \frac{l_{i+}}{2} - \frac{l_{++}}{4}, \quad (2.2-27)$$

$$u_{2(j)} = \frac{l_{+j}}{2} - \frac{l_{++}}{4}, \quad (2.2-28)$$

and the interaction term becomes

$$u_{12(ij)} = l_{ij} - \frac{l_{i+}}{2} - \frac{l_{+j}}{2} + \frac{l_{++}}{4}. \quad (2.2-29)$$

We note that the main effects are functions of the marginal sums of the logarithms but do not correspond to the marginal sums p_{i+} and p_{+j} in the original scale.

We now consider properties of the log-linear model.

Relationship of u -terms to cross-product ratios

From equations (2.2-18) and (2.2-27) we have

$$\begin{aligned} u_{1(1)} &= \frac{1}{4}(\log p_{11} + \log p_{12} - \log p_{21} - \log p_{22}) \\ &= \frac{1}{4} \log \alpha_1. \end{aligned} \quad (2.2-30)$$

Similarly, from expressions (2.2-17) and (2.2-28) we have

$$\begin{aligned} u_{2(1)} &= \frac{1}{4}(\log p_{11} - \log p_{12} + \log p_{21} - \log p_{22}) \\ &= \frac{1}{4} \log \alpha_2, \end{aligned} \quad (2.2-31)$$

and from (2.2-16) and (2.2-29) we have

$$\begin{aligned} u_{12(11)} &= \frac{1}{4}(\log p_{11} - \log p_{12} - \log p_{21} + \log p_{22}) \\ &= \frac{1}{4} \log \alpha_3. \end{aligned} \quad (2.2-32)$$

Thus the main effects in the log-linear u -term model are directly related to the two cross-product ratios described above, u_1 to α_1 and u_2 to α_2 . The choice of subscripts for the α_i now becomes apparent. We note that for $u_{1(1)}$ the terms in p appear with positive sign whenever variable 1 is at level 1, and similarly for $u_{2(1)}$ and variable 2 at level 1. For $u_{12(11)}$, the positive sign appears whenever both variables are on the same level. Thus the u -terms can be regarded as measures of departure from independence for the three different data arrangements.

Effect of imposing constraints

To assess the effect on the u -terms of imposing constraints on the $\{p_{ij}\}$, we need to revert to the arithmetic scale.

We can rewrite the model (2.2-20) for cell (1, 1) as

$$\log p_{11} = u + \frac{1}{4} \log \alpha_1 + \frac{1}{4} \log \alpha_2 + \frac{1}{4} \log \alpha_3, \quad (2.2-33)$$

and hence

$$p_{11} = \lambda \alpha'_1 \alpha'_2 \alpha'_3, \quad (2.2-34)$$

where $\log \lambda = u$ and $\alpha'_i = (\alpha_i)^{1/4}$ for $i = 1, 2, 3$. Then the basic table can be rewritten as

		A_2		
		1	2	Totals
A_1	1	$\lambda \alpha'_1 \alpha'_2 \alpha'_3$	$\frac{\lambda \alpha'_1}{\alpha'_2 \alpha'_3}$	$\lambda \alpha'_1 \left(\alpha'_2 \alpha'_3 + \frac{1}{\alpha'_2 \alpha'_3} \right)$
	2	$\frac{\lambda \alpha'_2}{\alpha'_1 \alpha'_3}$	$\frac{\lambda \alpha'_3}{\alpha'_1 \alpha'_2}$	$\frac{\lambda}{\alpha'_1} \left(\frac{\alpha'_2}{\alpha'_3} + \frac{\alpha'_3}{\alpha'_2} \right)$
Totals		$\lambda \alpha'_2 \left(\alpha'_1 \alpha'_3 + \frac{1}{\alpha'_1 \alpha'_3} \right)$	$\frac{\lambda}{\alpha'_2} \left(\frac{\alpha'_1}{\alpha'_3} + \frac{\alpha'_3}{\alpha'_1} \right)$	1

Setting $p_{1+} = p_{2+} = 1/2$ implies that the $\{\alpha_i\}$ must satisfy the relationship

$$\alpha_1^{1/2} - \alpha_2^{1/2} - \alpha_3^{1/2} + (\alpha_1 \alpha_2 \alpha_3)^{1/2} = 0. \quad (2.2-35)$$

If we set $\alpha'_1 = 1$, which is equivalent to setting $u_1 = 0$, the condition (2.2-35) becomes

$$\left(\alpha'_2 - \frac{1}{\alpha'_2} \right) \left(\alpha'_3 - \frac{1}{\alpha'_3} \right) = 0, \quad (2.2-36)$$

which is satisfied by either $\alpha'_2 = 1$ or $\alpha'_3 = 1$, or both. Equivalently, we must have $u_2 = 0$ or $u_{12} = 0$, or both.

This result also holds in larger tables; constant marginal probabilities do not imply that $u_1 = 0$ unless we also have $u_2 = 0$ or $u_{12} = 0$, or both. Consequently, when we move from simple random sampling to sampling different segments of the population independently, we cannot specify that a margin is fixed by placing constraints on a single u -term.

Model describes probabilities or expected counts

So far we have dealt entirely with a table of probabilities that sum to 1. If we consider instead a table of expected counts $\{m_{ij}\}$ that sum to a grand total $N = \sum_{i,j} m_{ij}$, we have $m_{ij} = Np_{ij}$, and hence

$$\begin{aligned} \log m_{ij} &= \log N + \log p_{ij} \\ &= u' + (u_{1(i)} + u_{2(j)} + u_{12(ij)}), \end{aligned} \quad (2.2-37)$$

where $u' = u + \log N$. Thus for the single sample we can describe the structure of the expected counts instead of the structure of the probabilities by changing the value of u from the mean of the logarithms of the $\{p_{ij}\}$ to the mean of the logarithms of the $\{m_{ij}\}$, and henceforth we denote the constant by u in both cases. In other words, the equations (2.2-26)–(2.2-29) are applicable if we define $l_{ij} = \log m_{ij}$ instead of $l_{ij} = \log p_{ij}$.

It follows that α can be defined similarly as the cross-product ratio of expected counts instead of probabilities.

Model applicable in varied sampling situations

So far we have considered taking a single sample of size N , with p_{ij} the probability of an individual falling into the cell (i, j) . This is the simple random sampling scheme. A fourfold table can also be generated by other sampling schemes. Suppose that we take a sample of N_1 individuals from the first category of variable A and N_2 from the second category, and then count how many fall into the different categories of variable B . Our table of expected counts becomes

		<i>B</i>		
		1	2	Totals
<i>A</i>	1	m_{11}	m_{12}	N_1
	2	m_{21}	m_{22}	N_2
Totals		m_{+1}	m_{+2}	N

(2.2-38)

and we have

$$m_{11} + m_{12} = N_1,$$

$$m_{21} + m_{22} = N_2,$$

$$N_1 + N_2 = N.$$

Corresponding to this table, there is a table of probabilities $P_{j(i)}$ the probability of being in category j for sample i . Thus

$$\begin{aligned} N_1 P_{j(1)} &= m_{1j}, \\ N_2 P_{j(2)} &= m_{2j}, \end{aligned} \quad (2.2-39)$$

for $j = 1, 2$. We write these probabilities with capital letters, as they are no longer the probabilities giving the frequency of occurrence of the four types of individuals in the population. Instead of the four probabilities summing to 1, we have

$$\begin{aligned} P_{1(1)} + P_{2(1)} &= 1, \\ P_{1(2)} + P_{2(2)} &= 1. \end{aligned} \quad (2.2-40)$$

We have taken two independent samples from different segments of the population and cannot get back to the population p_{ij} unless we know the relative magnitude of the two segments of the population.

Our log-linear model is still applicable to the table of expected counts (2.2-38), but the restriction (2.2-35) derived for equal row margins applies, so the relative magnitudes of the u -terms are constrained. In other sampling plans the restrictions on the probabilities differ in other ways. For simplicity, in the rest of this chapter we discuss log-linear models in terms of expected counts, not probabilities.

Before comparing the log-linear model with other models, we give an example of sampling that gives a 2×2 table with a fixed margin.

Example 2.2-1 Sensitivity, specificity, and predictive value

The problem of evaluating a new laboratory procedure designed to detect the presence of disease affords an example not only of sampling so that a 2×2 table has a fixed margin, but also of rearranging four cells for three different purposes.

1. Natural arrangement for laboratory data

To determine how effectively the laboratory procedure identifies positives and negatives, the investigator evaluates N_1 persons known to have the disease and N_2 persons known to be free of the disease. The results are designed to estimate the expected counts in array (2.2-41). In this array we no longer enclose every elementary cell in a box, but the arrangement of cells is the same as in array (2.2-10).

True State	Laboratory Procedure		Totals	
	Disease	No Disease		
Disease	m_{11}	m_{12}	N_1	(2.2-41)
No Disease	m_{21}	m_{22}	N_2	

A perfect laboratory procedure correctly identifies as diseased all those persons who are truly diseased and none of those who are not diseased; this situation corresponds to $m_{21} = m_{12} = 0$. Thus $\alpha_3 = m_{11}m_{22}/m_{21}m_{12}$ tells us whether the laboratory procedure is of any value. Unless α_3 is large, the laboratory procedure is abandoned.

2. Measuring sensitivity and specificity

When the evaluation of the laboratory procedure is described, laboratory results indicating disease are considered positive, the others negative. The term "sensi-

tivity” is used for the proportion of positive results that agree with the true state, and the term “specificity” for the proportion of negative results that agree with the true state. These are the proportions described by the rearranged array:

True State	Laboratory Procedure		Totals	
	Correct	Incorrect		
Disease	m_{11}	m_{12}	N_1	(2.2-42)
No Disease	m_{22}	m_{21}	N_2	

Now each row yields one of the proportions of interest:

$$\begin{aligned}\text{sensitivity} &= P_{1(1)} = \frac{m_{11}}{N_1} = 1 - P_{2(1)} = 1 - \frac{m_{12}}{N_1}, \\ \text{specificity} &= P_{2(2)} = \frac{m_{22}}{N_2} = 1 - P_{1(2)} = 1 - \frac{m_{21}}{N_2}.\end{aligned}\quad (2.2-43)$$

The relative magnitude of the sensitivity and specificity is measured by

$$\alpha_1 = \frac{m_{11}m_{21}}{m_{22}m_{12}}.$$

Such laboratory procedures are often used on large populations to find diseased persons. When a choice is to be made between two competitive procedures for screening a population, the prevalence and nature of the disease determines which characteristic, sensitivity or specificity, should be maximized.

3. Assessing predictive value

The third arrangement of the array does not preserve the fixed margins N_1 and N_2 :

Agrees with True State	Laboratory Procedure		
	Disease	No Disease	
Yes	m_{11}	m_{22}	(2.2-44)
No	m_{21}	m_{12}	

Unless the sample sizes N_1 and N_2 are proportional to the prevalence of the disease in the population where the laboratory procedure is to be used as a screening device, $\alpha_2 = m_{11}m_{12}/m_{21}m_{22}$ does not measure the relative odds on a correct prediction according to the outcome of the laboratory procedure.

To assess whether the cost of screening a population is worthwhile in terms of the number of cases detected, the health official needs to know the positive predictive value $PV+$ and the negative predictive value $PV-$. To compute predictive values we need to know the proportion D of diseased persons in the population to be screened. Then we multiply the first row of the original table (2.2-41) by D/N_1 and the second row by $(1 - D)/N_2$ to obtain

True State	Laboratory Procedure		
	Disease	No Disease	
Disease	$DP_{1(1)}$	$DP_{2(1)}$	(2.2-45)
No Disease	$(1 - D)P_{1(2)}$	$(1 - D)P_{2(2)}$	

The cross-product ratio α_3 is the same in array (2.2-45) as in array (2.2-41). Similarly, if we rearrange array (2.2-45) to correspond with array (2.2-42), we obtain the same values for sensitivity and specificity. When we rearrange array (2.2-45) to correspond with array (2.2-44), a difference occurs. We obtain

Agrees with True State	Laboratory Procedure	
	Disease	No Disease
Yes	$DP_{1(1)}$	$(1 - D)P_{2(2)}$
No	$(1 - D)P_{1(2)}$	$DP_{2(1)}$

(2.2-46)

The cross product in array (2.2-46) differs from that in array (2.2-44) by the factor $D^2/(1 - D)^2$ and measures the relative odds in the population of having the disease according to the results of the laboratory procedure. For the positive laboratory results we have

$$\begin{aligned}
 PV+ &= \frac{DP_{1(1)}}{(1 - D)P_{1(2)} + DP_{1(1)}} \\
 &= \frac{1}{1 + \frac{1 - D}{D} \frac{N_1}{N_2} \frac{m_{21}}{m_{11}}}, \quad (2.2-47)
 \end{aligned}$$

and for the negative laboratory results

$$PV- = \frac{1}{1 + \frac{D}{1 - D} \frac{N_2}{N_1} \frac{m_{12}}{m_{22}}}. \quad (2.2-48)$$

When the two predictive values are equal we have independence in array (2.2-46).

Thus we have shown that rearranging tables has practical applications. It is helpful in assessing the relationships between predictive values, and between sensitivity and specificity for particular disease prevalences, as discussed by Vechio [1966]. (See exercises 1 and 2 in Section 2.6 for further details.)* ■■

2.2.4 Differences between log-linear and other models

Models other than the log-linear have been proposed for describing tables of counts. We now discuss two contenders and show that the logit model can be regarded as a different formulation of the log-linear model, but models that are linear in the arithmetic scale have different advantages and disadvantages.

Logit models

Suppose that the row totals m_{1+} and m_{2+} are fixed and that we are interested in the relative proportions in the rows. We have, as before, $P_{1(i)} = m_{i1}/m_{1+}$ for $i = 1, 2$.

Then the logit for the i th row is defined as

$$L_i = \log \frac{P_{1(i)}}{1 - P_{1(i)}} = \log \frac{m_{i1}}{m_{i2}}. \quad (2.2-49)$$

From the saturated model

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}, \quad (2.2-50)$$

* The symbol ■■ marks the end of an example.

we find that

$$\begin{aligned} L_i &= u_{2(1)} - u_{2(2)} + u_{12(i1)} - u_{12(i2)} \\ &= 2u_{2(1)} + 2u_{12(i1)}, \end{aligned}$$

and letting $w = 2u_{2(1)}$ and $w_{1(i)} = 2u_{12(i1)}$, we get

$$L_i = w + w_{1(i)}, \quad (2.2-51)$$

with $w_{1(1)} + w_{1(2)} = 0$. Thus we have transformed the log-linear model for the expected cell counts into a linear model for the logits.

We can now compare the linear logit model with the linear model for the one-way analysis of variance, because we can think of the row variable A as being an independent variable and the column variable B as being a dependent variable. As $w_{1(ij)}$ measures the structural relationship between A and B (i.e., because $u_{12(ij)}$ measures this relationship), we can speak of the effect of A on B .

We discuss other aspects of logits in Section 2.3.5, where we show that the logit model is appropriate primarily for stratified samples. It is unduly restrictive for a simple random sample, as it requires that one margin be fixed. In Chapter 10, Section 10.4, we discuss uses that have been made of the logistic model for mixtures of quantitative and qualitative variables.

Additive models

It is natural to explore the possibility of using a linear model in the cell probabilities instead of their logarithms. Suppose we let

$$p_{ij} = \mu + \beta_i + \gamma_j + \varepsilon_{ij} \quad i = 1, 2; \quad j = 1, 2, \quad (2.2-52)$$

with

$$\beta_+ = \gamma_+ = \varepsilon_{i+} = \varepsilon_{+j} = 0.$$

Since the $\{p_{ij}\}$ must sum to 1, $\mu = \frac{1}{4}$. By examining the marginal totals, we also have

$$\begin{aligned} \beta_i &= \frac{1}{2}(p_{i+} - \frac{1}{2}) \quad i = 1, 2, \\ \gamma_j &= \frac{1}{2}(p_{+j} - \frac{1}{2}) \quad j = 1, 2. \end{aligned} \quad (2.2-53)$$

Thus, unlike the u -terms, the β_i and γ_j are directly interpretable in terms of the marginal totals p_{i+} and p_{+j} . This advantage brings with it the range restrictions

$$\begin{aligned} -\frac{1}{4} &\leq \beta_i \leq \frac{1}{4}, \\ -\frac{1}{4} &\leq \gamma_j \leq \frac{1}{4}, \\ -\frac{1}{4} &\leq \varepsilon_{ij} \leq \frac{1}{4}. \end{aligned} \quad (2.2-54)$$

The major problem comes in the interpretation of ε_{11} , which we can write as

$$\begin{aligned} \varepsilon_{11} &= \frac{1}{4}(p_{11} + p_{22} - p_{12} - p_{21}) \\ &= \frac{1}{4}(4p_{11} - 2p_{1+} - 2p_{+1} + 1). \end{aligned} \quad (2.2-55)$$

Setting $\varepsilon_{11} = 0$ does not imply independence of the underlying variables unless $p_{1+} = \frac{1}{2}$ or $p_{+1} = \frac{1}{2}$, nor does setting $p_{ij} = p_{i+}p_{+j}$ imply that ε_{11} takes on any specific value.

We have found that the cross-product ratio α_3 is a simple function of u_{12} and has useful invariance properties. We cannot express α_3 as a simple function of the $\{\varepsilon_{ij}\}$, nor can we find a simple alternative function of the $\{\varepsilon_{ij}\}$ that has the invariance properties. We conclude that the difficulty of relating the additive model to the concept of independence makes it less attractive than the log-linear model.

2.3 Two Dimensions—The Rectangular Table

The log-linear model used to describe the structure of the 2×2 table is unaltered in appearance when applied to larger two-way tables. The number and interpretation of parameters differ for larger tables. The applicability of the model to expected cell counts or to probabilities and its suitability for a variety of sampling schemes persist, as does the relationship to logit models. The rearrangement of cells demonstrated for the 2×2 table is not as useful for interpreting parameters in larger tables, except when the arrays, instead of being rectangular, take some other shape such as triangular.

2.3.1 The log-linear model

Suppose the cells from a single sample of size N form a rectangular array with I rows and J columns, corresponding to the I categories of variable 1 and J categories of variable 2. We have already seen that the log-linear model describes either probabilities or expected counts. As we wish to consider later a variety of sampling schemes, we define the model in terms of expected counts. A given sampling scheme places constraints on the expected counts, but for every scheme we have

$$\sum_{i,j} m_{ij} = N, \quad (2.3-1)$$

and define $l_{ij} = \log m_{ij}$ for $i = 1, \dots, I; j = 1, \dots, J$.

The log-linear model is unchanged from the form used for the 2×2 table:

$$l_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}. \quad (2.3-2)$$

The number of parameters contained in each u -term is a function of I and J , but the constraints are unaltered:

$$\sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_i u_{12(ij)} = \sum_j u_{12(ij)} = 0. \quad (2.3-3)$$

By analogy with analysis of variance, we define

$$\text{overall mean, } u = \frac{l_{++}}{IJ}, \quad (2.3-4)$$

$$\text{main effect of variable 1, } u_{1(i)} = \frac{l_{i+}}{J} - \frac{l_{++}}{IJ}, \quad (2.3-5)$$

$$\text{main effect of variable 2, } u_{2(j)} = \frac{l_{+j}}{I} - \frac{l_{++}}{IJ}, \quad (2.3-6)$$

two-factor effect between variables,

$$u_{12(ij)} = l_{ij} - \left(\frac{l_{+j}}{I} + \frac{l_{i+}}{J} \right) + \frac{l_{++}}{IJ}. \quad (2.3-7)$$

Degrees of freedom

The constraints (2.3-3) reduce the number of independent parameters represented by each u -term. Thus the value of $u_{1(i)}$ differs for each of the I categories of variable 1, but the constraints reduce the number of independent parameters to $I - 1$. Similarly, u_{12} is an $I \times J$ array of parameters that sum to zero across each row and column, and so has $(I - 1)(J - 1)$ independent values. To verify that the total number of independent parameters equals the total number of elementary cells, we add the contributions from each u -term. The numbers of parameters for each term are listed under “degrees of freedom” because this is how we view parameters when we fit models to data in later chapters.

u -Term	Degrees of Freedom	
u	1	
u_1	$I - 1$	(2.3-8)
u_2	$J - 1$	
u_{12}	$IJ - I - J + 1$	
Total	IJ	

The sum IJ of all parameters in the saturated model matches the number of elementary cells (i, j) .

Constructing a table

We can substantiate the claim that a log-linear model describes the structure of a table by using a model to build a table. Table 2.3-1 helps us illustrate the process

Table 2.3-1 Construction of Artificial 2×3 Table

Cell	e^u	e^{u_1}	e^{u_2}	$e^{u_{12}}$	$e^{u_1 + u_2 + u_{12}}$	m_{ij}
1, 1	60*	2*	6*	4*	48	2,880
2, 1	60	1/2	6	1/4	3/4	45
1, 2	60	2	1/2*	5*	5	300
2, 2	60	1/2	1/2	1/5	1/20	3
1, 3	60	2	1/3	1/20	1/30	2
2, 3	60	1/2	1/3	20	10/3	200
Total						3,430

*Selected values.

for a 2×3 table. As there are six cells, we select values for six parameters (indicated by an asterisk in the table), and the constraints enable us to derive the other parameters from these six. For the main effect u_1 we put $e^{u_{1(1)}} = 2$; then its reciprocal is the value for $e^{u_{1(2)}}$, and the u_1 -term is completely defined. The second main effect u_2 has two degrees of freedom, so we select $e^{u_{2(1)}} = 6$, $e^{u_{2(2)}} = 1/2$, and derive $e^{u_{2(3)}}$ as the reciprocal of their product. Similarly for u_{12} we select $e^{u_{12(11)}} = 4$, derive $e^{u_{12(21)}} = 1/4$, and then select $e^{u_{12(12)}} = 5$ and derive the remaining parameters.

Multiplying e^{u_1} , e^{u_2} , and $e^{u_{12}}$ gives $e^{u_1 + u_2 + u_{12}}$. For a simple example, we use integers for the $\{m_{ij}\}$, and we can get them by selecting $e^u = 60$. The $\{m_{ij}\}$ are given in the last column. If we wish to construct values for p_{ij} instead of m_{ij} , we sum the values for $e^{u_1 + u_2 + u_{12}}$ and define e^u as the reciprocal of this sum to ensure that the $\{p_{ij}\}$ sum to 1.

Extension of this construction process to larger tables follows the same procedure. Such dummy tables of known structure are used for checks of computing procedures, empirical investigations of the effects of collapsing tables, and theoretical investigations utilizing Monte Carlo methods.

2.3.2 Cross-product ratios and linear contrasts

When the number of categories per variable exceeds 2, we find that we can still express each u -term as a function of cross-product ratios, or equivalently as a linear contrast. We consider first increasing the number of categories for one variable only.

The $2 \times J$ table

With two rows and J columns, the log odds for any column j are

$$\log(m_{1j}/m_{2j}) = l_{1j} - l_{2j} = 2(u_{1(1)} + u_{12(1j)}). \quad (2.3-9)$$

If we let $\alpha_{r,s}$ denote the cross-product ratio for columns r and s , we have

$$\log \alpha_{r,s} = \log(m_{1r}/m_{2r}) - \log(m_{1s}/m_{2s}). \quad (2.3-10)$$

For the first two columns,

$$\log(\alpha_{1,2}) = 2(u_{12(11)} - u_{12(12)}), \quad (2.3-11)$$

and for the first and j th column,

$$\log(\alpha_{1,j}) = 2(u_{12(11)} - u_{12(1j)}). \quad (2.3-12)$$

Taking the logarithm of the product of all $J - 1$ such α -terms yields

$$\begin{aligned} \log(\alpha_{1,2} \alpha_{1,3} \cdots \alpha_{1,J}) &= 2(J - 1)u_{12(11)} - 2(u_{12(12)} + u_{12(13)} + \cdots + u_{12(1J)}) \\ &= 2Ju_{12(11)}. \end{aligned} \quad (2.3-13)$$

Thus each parameter of u_{12} is a function of $J - 1$ cross products. It is a matter of rearrangement to obtain the linear contrast

$$\begin{aligned} u_{12(11)} &= \frac{1}{2J} \sum_{j=2}^J \log(\alpha_{1,j}) \\ &= \frac{J-1}{2J} (l_{11} - l_{21}) + \frac{1}{2J} \sum_{j=2}^J (l_{2j} - l_{1j}). \end{aligned} \quad (2.3-14)$$

Alternatively, we can write

$$e^{u_{12(11)}} = \prod_{j=2}^J \left(\frac{m_{11}m_{2j}}{m_{21}m_{1j}} \right)^{1/2J}. \quad (2.3-15)$$

The contrast (2.3-14) could have been derived more directly from expression (2.3-7). Using this more direct approach for the other terms, we have, from expression (2.3-5),

$$u_{1(1)} = \frac{\sum_j l_{1j}}{J} - \frac{\sum_{i,j} l_{ij}}{2J} = \sum_j \frac{(l_{1j} - l_{2j})}{2J}, \quad (2.3-16)$$

or the corresponding relationship

$$e^{u_{1(1)}} = \prod_j \left(\frac{m_{1j}}{m_{2j}} \right)^{1/2J}. \quad (2.3-17)$$

The other main-effect term is similarly defined, from (2.3-6), as

$$u_{2(1)} = \frac{(J-1)}{2J}(l_{11} + l_{21}) - \sum_{j=2}^J \frac{(l_{1j} + l_{2j})}{2J}, \quad (2.3-18)$$

or can be written in product form

$$e^{u_{2(1)}} = \prod_{j=2}^J \left(\frac{m_{11}m_{21}}{m_{1j}m_{2j}} \right)^{1/2J}, \quad (2.3-19)$$

with each term a rearrangement of the four cells used to define $\alpha_{1 \cdot j}$.

In any rectangular table we can use the definitions (2.3-4)–(2.3-7) to express parameters of subscripted u -terms as linear contrasts of the $\{l_{ij}\}$ or to give the corresponding multiplicative form.

2.3.3 Effect of combining categories

As soon as a table has more than two categories per variable, the possibility of amalgamating categories arises. We need to consider when we can combine categories without changing the structure of the table.

The 2×3 table

Consider taking a 2×2 table and subdividing the second category of the second variable to give a 2×3 table. The new table has cell entries m'_{ij} , where

$$m_{i2} = m'_{i2} + m'_{i3}, \quad (2.3-20)$$

$$m_{i1} = m'_{i1} \quad (2.3-21)$$

for $i = 1, 2$, and is described by the new model

$$\log m'_{ij} = u' + u'_{1(i)} + u'_{2(j)} + u'_{12(ij)}. \quad (2.3-22)$$

Although cell (1, 1) is unchanged, the parameters in the new model will in general differ from those in the original model. For instance, in the original table from (2.3-15) we obtain

$$4u_{12(11)} = \log \left(\frac{m_{11}m_{22}}{m_{21}m_{12}} \right), \quad (2.3-23)$$

and in the expanded table

$$6u'_{12(11)} = \log \left[\left(\frac{m'_{11}}{m'_{21}} \right)^2 \left(\frac{m'_{22}m'_{23}}{m'_{12}m'_{13}} \right) \right]. \quad (2.3-24)$$

Comparing expressions (2.3-23) and (2.3-24) shows us that in general $u_{12(11)} \neq u'_{12(11)}$.

Putting all three odds ratios in the new table equal to one another, i.e., setting

$$\frac{m'_{11}}{m'_{21}} = \frac{m'_{12}}{m'_{22}} = \frac{m'_{13}}{m'_{23}},$$

gives $u'_{12(11)} = 0$, and we find that for the original table

$$\frac{m_{12}}{m_{22}} = \frac{m'_{12} + m'_{13}}{m'_{22} + m_{23}} = \frac{m_{11}}{m_{21}}.$$

Thus we also have $u_{12(11)} = 0$ and have shown that independence in the expanded table implies independence in the condensed table. The converse does not hold: a smaller table fitting the independence model can be derived from a larger table with more complex structure, as we show in the following example.

Example 2.3-1 Condensing categories

Consider three tables:

Table A			Table B		Table C		
4	2	6	4	8	4	1	7
6	3	9	6	12	6	6	6

In table A the row odds m_{1j}/m_{2j} are constant for all j , so the table fits the independence model. Pooling the last two columns gives table B, with the same constant row odds. Thus both tables are fitted by the independence model. A new partitioning of the second column of table B gives table C, and the row odds are no longer constant. In other words, table C does not fit the independence model, but in the condensed table B we have independence. ■■

The independent rectangular table

For any rectangular array, we can express the independence of rows and columns in two equivalent forms:

$$m_{ij} = \frac{m_{i+}m_{+j}}{N}, \quad (2.3-25)$$

$$l_{ij} = u + u_{1(i)} + u_{2(j)}. \quad (2.3-26)$$

The independence model (2.3-26) is derived from the model (2.3-2) by putting $u_{12(ij)} = 0$ for all i and j , or, more briefly, by putting $u_{12} = 0$. Independent tables have special properties.

From model (2.3-26), the difference between the logarithms of any two cells in the same column is

$$l_{ij} - l_{rj} = u_{1(i)} - u_{1(r)}. \quad (2.3-27)$$

Changing from the log scale to the original scale, this yields

$$\begin{aligned}\frac{m_{ij}}{m_{rj}} &= e^{u_{1(i)} - u_{1(r)}} \\ &= \frac{m_{i+}}{m_{r+}}\end{aligned}\quad (2.3-28)$$

for all j . Thus the ratio of internal cells in a given column is the same as the ratio of corresponding row sums. By symmetry, this is also true for row elements and column sums. For tables described by the independence model (2.3-26), we have the following conclusions:

1. independence is not lost if we combine some of the categories of either variable;
2. the parameters of u_1 can be determined from the vector of row sums $\{m_{i+}\}$, and similarly for u_2 from the column sums.

We have thus established that two-way tables with independent structures are *collapsible*. We can combine one or more categories of either variable, and the same model describes the structure of the reduced table. If we combine all the categories of one variable, say variable 2, then we have a string of I cells, and from these cells we can compute the parameters of u_1 . The values of u_1 that we obtain from the reduced table are identical to those obtained, using expression (2.3-20) for instance, from the full table. Conversely, if the structure is not independent, combining categories gives a reduced table with parameter values that differ from those of the parent table; the parameters of u_{12} may even become zero, thus giving a reduced table of different structural form from the parent table.

Identifying collapsible structures in more than two dimensions is a useful tool for handling large tables. When the structures are not collapsible, the analyst who inspects only two-way tables of sums can be led to false conclusions about the interaction patterns between the variables.

2.3.4 Different sampling schemes

When we constructed a table by specifying values of the parameters we found that the size N was dependent on the value u selected for the overall mean. This constraint is imposed by simple random sampling. When we have stratified sampling, the size of the sample selected from each stratum is fixed. We can still use model (2.3-2) to describe the structure of the table of expected counts, but further constraints are imposed on the u -terms.

Consider I strata with N_i observations in the i th stratum, where $\sum_i N_i = N$. If the J categories for each stratum are the same, putting

$$\log m_{ij} = l_j^{(i)} = v^{(i)} + v_{2(j)}^{(i)} \quad (2.3-29)$$

with $\sum_j v_{2(j)}^{(i)} = 0$ defines each stratum in terms of the log mean of the stratum and deviations from it.

We now compare the expressions obtained from model (2.3-29) for average values of the $l_j^{(i)}$ with values obtained from the u -term model (2.3-2) applied to the

whole table. Taking the mean of all cells gives

$$u = \sum_i \frac{v^{(i)}}{I}. \quad (2.3-30)$$

Using this relation, we obtain from the mean of the j th column

$$u_{2(j)} = \sum_i \frac{v_{2(j)}^{(i)}}{I}, \quad (2.3-31)$$

from the mean of the i th row

$$u_{1(i)} = v^{(i)} - u, \quad (2.3-32)$$

and finally, from the expressions for single cells and relations (2.3-30)–(2.3-32),

$$u_{12(ij)} = v_{2(j)}^{(i)} - u_{2(j)}. \quad (2.3-33)$$

Thus we can relate the two-dimensional log-linear model (2.3-2) to the one-dimensional analogue (2.3-29). The only constraint on the two-dimensional model introduced by stratification is that the magnitudes of the N_i restrict the magnitudes of the $u + u_{1(i)}$.

2.3.5 The logit model

The logit formulation is equivalent to the log-linear model for describing the structure of the $2 \times J$ table with one fixed margin $\{m_{+j}\}$. Thus we can write $m_{+j} = N_j$. By definition, we have

$$\text{logit}(j) = \log \left(\frac{m_{1j}}{m_{2j}} \right) = l_{1j} - l_{2j} \quad (2.3-34)$$

and can write $\text{logit}(j) = 2(u_{1(1)} + u_{12(1j)})$, or equivalently,

$$\text{logit}(j) = w + w_{2(j)}, \quad (2.3-35)$$

where $w = 2u_{1(1)}$ and $w_{2(j)} = 2u_{12(1j)}$.

When the sampling determines the J marginal totals N_j , we have only J degrees of freedom among the $2J$ cells, and this is adequately reflected by the two-term model (2.3-35). For a single sample with only N fixed, model (2.3-35) does not describe the structure of the full array because it gives no information about the margins $\{m_{+j}\}$.

Although we most often use the logit model for stratified sample schemes where each stratum has only two categories, we can also use it for the multiple-category case by defining

$$\text{logit}(ij) = \log \frac{m_{ij}}{\sum_{r \neq i} m_{rj}}$$

for $i = 1, \dots, (I - 1)$. We can thus regard the logit model as a special case of the more general log-linear model, suitable only for stratified sampling schemes. In Section 10.4 we discuss useful extensions of the logit model for handling mixtures of discrete and continuous variables.

2.3.6 Irregular arrays

When rows and columns correspond to two variables it is sometimes impossible to observe particular category combinations. If this logical impossibility occurs in the (i, j) cell, we say that we have a *structural zero* and $p_{ij} = m_{ij} = 0$. We discuss tables with structural zeros in Chapter 5 and show that we can use log-linear models to describe the structure of the cells where $m_{ij} \neq 0$. We refer to the array of nonzero cells as an *incomplete* table.

In two dimensions, incomplete tables can often be arranged in a triangular array. Triangular arrays arise commonly from

1. measurements on paired individuals where no other distinction is made between members of each pair, e.g., measurements on two eyes of an individual that are not distinguished as right eye and left eye but only sorted by the measurement itself as better eye or worse eye;
2. paired measurements on a single individual, where the value of one measurement sets bounds on the possible values of the second. Commonly, we first assign each individual to a particular category of an ordered set, and any subsequent assignment must place the individual in a higher category; e.g., individuals graded by severity of disease on admission to hospital are not discharged alive until the disease grade has improved.

We can also create triangular tables by folding square tables along a diagonal if this helps in our analysis. We discuss the usefulness of this device in Chapter 8. Further discussion of incomplete tables is deferred to Chapters 5 and 8.

2.4 Models for Three-Dimensional Arrays

As the number of variables measured on each individual increases, the resulting multidimensional contingency tables become more unwieldy. The investigator is apt to give up on multiway tables and look instead at large numbers of two-way tables derived by adding over the categories of all except two variables. In this section we show for three-dimensional tables the dangers inherent in examining only such tables of sums and advocate instead construction of models that describe the full array. We discuss first the $2 \times 2 \times 2$ table and then proceed to general $I \times J \times K$ rectangular tables, with the main focus on

1. interpreting each parameter of the saturated model;
2. interpreting unsaturated models as descriptions of hypotheses;
3. determining when the size of the table may be reduced without distorting the structural relationships between variables of interest.

We extend the notation describing the cells of a two-dimensional array to encompass multiway tables simply by adding more subscripts. The number of subscripts normally matches the number of variables, but exceptions occur. It may sometimes be convenient to split or combine the categories of a single variable, while in other instances it may be useful to fold a two-dimensional array so that it forms an irregular three-dimensional shape. We therefore define subscripts to match the dimension of a particular arrangement of the cells. In three dimensions, the probability of a count falling in cell (i, j, k) is p_{ijk} and the expected count is m_{ijk} , where $i = 1, \dots, I$; $j = 1, \dots, J$; $k = 1, \dots, K$.

In complete three-dimensional arrays we have a rectangular parallelepiped of size $I \times J \times K$ containing IJK cells. The subscript “+” denotes addition across all cells with a common value for one or more subscripts. Thus $m_{+jk} = \sum_{i=1}^I m_{ijk}$, the sum of all cells with a common value for j and k .

2.4.1 The $2 \times 2 \times 2$ table

When we take measurements on three dichotomous variables, we have eight possible combinations of outcome. We can arrange the eight cells in a $2 \times 2 \times 2$ cube with each dimension corresponding to one variable. For display in two dimensions, we split the cube into two 2×2 arrays. If we make the split by dividing the categories of the third variables we have:

		First Category of Variable 3		Second Category of Variable 3	
		Variable 2		Variable 2	
		1	2	1	2
Variable 1	1	m_{111}	m_{121}	m_{112}	m_{122}
	2	m_{211}	m_{221}	m_{212}	m_{222}

We can now describe each 2×2 array by a separate log-linear model: for array k , we have

$$l_{ijk} = v^{(k)} + v_{1(i)}^{(k)} + v_{2(j)}^{(k)} + v_{12(ij)}^{(k)} \quad k = 1, 2, \quad (2.4-1)$$

with all subscripted v -terms summing to zero across each subscript variable as usual. We combine these models into a single linear expression by taking means across the tables. Remembering that in this array $K = 2$, we have the following mean effects:

overall mean,

$$u = \frac{1}{K} \sum_k v^{(k)},$$

main effect of variable 1,

$$u_{1(i)} = \frac{1}{K} \sum_k v_{1(i)}^{(k)}, \quad (2.4-2)$$

main effect of variable 2,

$$u_{2(j)} = \frac{1}{K} \sum_k v_{2(j)}^{(k)},$$

interaction between variables 1 and 2,

$$u_{12(ij)} = \frac{1}{K} \sum_k v_{12(ij)}^{(k)}.$$

Thus u_{12} , sometimes called the “partial association,” is the average interaction between variables 1 and 2.

The deviations from these means depend on the third variable and are defined as follows :

main effect of variable 3,

$$u_{3(k)} = v^{(k)} - u,$$

interactions with variable 3,

$$u_{13(ik)} = v_{1(i)}^{(k)} - u_{1(i)}, \quad (2.4-3)$$

$$u_{23(jk)} = v_{2(j)}^{(k)} - u_{2(j)},$$

three-factor effect,

$$u_{123(ijk)} = v_{12(ij)}^{(k)} - u_{12(ij)}.$$

We can now write the single linear model for the whole $2 \times 2 \times 2$ cube,

$$l_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}. \quad (2.4-4)$$

The subscripted u -terms of equations (2.4-2) are derived by taking means of u -terms that sum to zero within tables ; this derivation preserves the property that subscripted terms sum to zero. The terms with subscript k are all deviations and so also have this property. For instance,

$$\sum_i u_{123(ijk)} = \sum_j u_{123(ijk)} = \sum_k u_{123(ijk)} = 0. \quad (2.4-5)$$

Consequently, we have one absolute value for the parameters of each u -term in the $2 \times 2 \times 2$ table. Thus each of the eight u -terms in the saturated model contributes one degree of freedom, and the total number of degrees of freedom matches the total number of cells as required.

We now consider interpretations of the parameters of model (2.4-4) which can be extended to $I \times J \times K$ tables of any size.

Interpretation of parameters and the hierarchy principle

a. Two-factor effects

By splitting the cube according to the categories of the third variable, we introduced an apparent difference in the definitions of the two-factor effects that disappears when we consider different partitions. We defined the two-factor effect u_{12} as representing the interaction between variables 1 and 2 averaged over tables, and it was defined solely in terms of v_{12} . By partitioning the cube differently, we get the same interpretation as an average for the other two-factor terms u_{13} and u_{23} , instead of defining them as deviations as in (2.4-3). Conversely, the symmetry permits us to define u_{12} as a deviation from the overall effect of a single variable, similar to the definitions given for the other two-factor terms in (2.4-3). We can also define two-factor effects as products of cross-product ratios. If we define

$$\alpha^{(k)} = \frac{m_{11k}m_{22k}}{m_{12k}m_{21k}}, \quad (2.4-6)$$

we can use expression (2.4-6) to get

$$u_{12} = \frac{1}{8} \log(\alpha^{(1)}\alpha^{(2)}). \quad (2.4-7)$$

b. *Three-factor effect*

The three-dimensional model has one new feature that the two-dimensional model did not have, namely, the three-factor effect u_{123} . We derived u_{123} as the difference between the average value of u_{12} across tables and the particular value exhibited by table k . The symmetry of this definition becomes apparent when we write u_{123} in terms of cross-product ratios:

$$u_{123(111)} = \frac{1}{8} \log \left(\frac{\alpha^{(1)}}{\alpha^{(2)}} \right) = \frac{1}{8} \log \left(\frac{m_{111}m_{221}m_{122}m_{212}}{m_{121}m_{211}m_{112}m_{222}} \right). \quad (2.4-8)$$

All the cells whose subscripts sum to an odd number appear in the numerator and those whose subscripts sum to an even number in the denominator. This formulation is independent of the direction of partitioning; α -terms corresponding to u_{13} or u_{23} give the same ratio (2.4-8).

The three-factor effect is sometimes called the “second-order interaction.” It measures the difference in the magnitude of the two-factor effect between tables for any of the three partitions of the cube into two 2×2 tables. Such an interpretation of the meaning of the three-factor effect has the natural converse that if any two-factor effect is constant between tables, then the three-factor effect is zero. In particular, setting any two-factor effect equal to zero implies that the three-factor effect is zero. This leads us to a definition of the hierarchy principle, but before we can state this principle in general terms we need a more formal definition of the relationships between u -terms.

c. *Alternate interpretation of two-factor effects*

When dealing with two-dimensional tables we showed how to interpret all the subscripted u -terms in the log-linear model as functions of cross-product ratios. In particular, for a 2×2 table we showed how these cross-product ratios arise naturally from rearrangements of the table. For $2 \times 2 \times 2$ tables we can also show that all subscripted u -terms can be written as functions of ratios of cross-product ratios, and there are rearrangements of tables such that each subscripted u -term takes the form (2.4-8) and corresponds to the three-factor term for the rearrangement.

d. *Relationships between terms*

Consider two u -terms, one with r subscripts and the other with s subscripts, where $r > s$. We say that the terms are *relatives* if the r subscripts contain among them all the s subscripts. Thus u_{123} is a higher-order relative of all the other u -terms in the three-dimensional model, and u_{12} is a higher-order relative of both u_1 and u_2 .

e. *The hierarchy principle*

The family of hierarchical models is defined as the family such that if any u -term is set equal to zero, all its higher-order relatives must also be set equal to zero. Conversely, if any u -term is not zero, its lower-order relatives must be present in the log-linear model. Thus if $u_{12} = 0$, we must have $u_{123} = 0$; also, if u_{13} is present in the model, then u_1 and u_3 must be present also.

f. *Linear contrasts*

We showed for the 2×2 table that every u -term can be written as a linear contrast of the logarithms of the four cells. By rearranging the cells of the table, all the

subscripted terms can also be written as cross-product ratios. For the $2 \times 2 \times 2$ table we can similarly write every u -term as a linear contrast of the eight cells, or as a function of cross-product ratios.

Remembering that every u -term has one absolute value, in table 2.4-1 we assume that each term is positive in cell (1, 1, 1). Thus the first row has all plus signs. We can fill in the columns corresponding to each u -term so that each term sums to zero over each of its subscripts. These columns give us the linear contrasts, for instance,

$$u_{12(11)} = \frac{1}{8}(l_{111} + l_{221} - l_{121} - l_{211} + l_{112} + l_{222} - l_{122} - l_{212}). \quad (2.4-9)$$

This contrast has positive sign for all terms with subscripts i and j adding to an even number and negative for the remaining terms.

Table 2.4-1 Sign of u -terms of Fully Saturated Model for Three Dichotomous Variables

Cell	u	u_1	u_2	u_3	u_{12}	u_{13}	u_{23}	u_{123}
1, 1, 1	+	+	+	+	+	+	+	+
2, 1, 1	+	-	+	+	-	-	+	-
1, 2, 1	+	+	-	+	-	+	-	-
2, 2, 1	+	-	-	+	+	-	-	+
1, 1, 2	+	+	+	-	+	-	-	-
2, 1, 2	+	-	+	-	-	+	-	+
1, 2, 2	+	+	-	-	-	-	+	+
2, 2, 2	+	-	-	-	+	+	+	-

2.4.2 The $I \times J \times K$ model

We derived model (2.4-4) for the cube of side 2 by averaging across 2×2 tables. We can use a similar procedure for any three-dimensional array, namely, averaging across K tables of size $I \times J$. Expressions (2.4-1)–(2.4-3) are unaltered, and model (2.4-4) is the appropriate saturated model for any rectangular array in three dimensions.

When the sample is such that $\sum_{i,j,k} m_{ijk} = N$, the only further constraints are that each u -term sums to zero over each variable listed among its subscripts. We then have the following degrees of freedom associated with each level of u -terms:

Number in Level	Level	Degrees of Freedom	
1	overall mean	1	
3	one-factor terms	$(I - 1) + (J - 1) + (K - 1)$	
3	two-factor terms	$(I - 1)(J - 1) + (I - 1)(K - 1) + (J - 1)(K - 1)$	(2.4-10)
1	three-factor term	$(I - 1)(J - 1)(K - 1)$	
Total		IJK	

Thus the number of independent parameters corresponds to the number of elementary cells.

The interpretation of the individual u -terms is the same for any $I \times J \times K$ table as for the $2 \times 2 \times 2$ table except that when a u -term has more than one degree of freedom it is possible for some of its parameters to be zero while others are large. When this occurs, we need to consider whether some categories can be combined without violating the structure by introducing spurious interaction effects or masking effects that are present. Sometimes we are interested in combining only two categories of a multicategory variable, sometimes in combining more than two. In the extreme case we collapse the variable by combining all its categories. In theorem 2.4-1 below we consider when we can collapse a variable without violating the structure of the three-dimensional array. If the structure is such that we can safely collapse a variable, then we can also combine any subset of its categories without violating the structure.

Before turning to the problem of collapsing, we discuss briefly the effect of sampling design in fixing parts of the structure. We also consider the interpretation of the set of hierarchical models. In any particular data set, our decisions regarding the desirability of collapsing must be governed by what is meaningful in terms of the sampling design and the interpretation—as well as by the model structure.

Constraints imposed by sampling design

When we take a simple random sample and arrange the counts in a three-dimensional array, only the total sample size N is fixed. Stratified samples that yield a three-dimensional array of counts usually fall into two main types, each with a different structure of fixed subtotals:

- (i) if one of the three variables, say variable 1, determines the strata, then the components of the one-dimensional array $\{m_{i++}\}$ are fixed, and we have $m_{i++} = N_i$ for all i , and $\sum_i N_i = N$;
- (ii) if two of the variables, say variables 1 and 2, are required to describe the strata, then only the third variable, sometimes called the response variable, is measured on each individual. This scheme fixes the two-dimensional array $\{m_{ij+}\}$, and we have $m_{ij+} = N_{ij}$, for all i, j , and $\sum_{i,j} N_{ij} = N$. We show in Chapter 3 that only the hierarchical models that include u_{12} are appropriate for this design.

These are the two types of design that arise most frequently, but more complex designs can occur. For instance, it is possible to have more than one two-dimensional array of sums fixed by the sampling plan, and it is only appropriate to use a subset of the hierarchical models to describe them. We note that the saturated model describes all such arrays, but the constraints on the $\{n_{ijk}\}$ impose further constraints on the u -terms.

For two-dimensional tables we showed that the parameters of the logit model where one margin was fixed correspond to the relevant u -terms in the log-linear model. We can also use logit models for either of the designs (i) and (ii), and again the logit parameters correspond to the relevant u -terms in the log-linear model (see exercise 4 in Section 2.6).

Interpretation of models

In the following discussion of the interpretation of three-dimensional models, we assume that the sampling scheme is such that all the models are meaningful. As each subscripted u -term can be interpreted as measuring the deviations from lower-order terms, removal of higher-order terms has the effect of simplifying the structure. Starting with the most complex unsaturated model, we discuss each of the models that conform to the hierarchy principle defined in Section 2.4.1.

a. *Three-factor effect absent*

As u_{123} measures the difference between two-factor effects attributable to the third variable, putting $u_{123} = 0$ enables us to describe a table with constant two-factor effects. Thus the model

$$l_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} \quad (2.4-11)$$

states that there is “partial association” between each pair of variables, to use the terminology of Birch [1963].

In other chapters we fit such unsaturated models to observed data and compute measures of goodness of fit. These measures indicate whether the model is an adequate description of the structure of the population yielding the observed data. To test the null hypothesis that there is no three-factor effect or “second-order interaction” we fit model (2.4-11). The measure of goodness of fit is our test statistic, with $(I - 1)(J - 1)(K - 1)$ degrees of freedom.

b. *Three-factor and one two-factor effect absent*

There are three versions of the model with the three-factor effect and one two-factor effect missing. Selecting u_{12} as the absent two-factor effect gives

$$l_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)} + u_{13(ik)}. \quad (2.4-12)$$

This model states that variables 1 and 2 are independent for every level of variable 3, but each is associated with variable 3. In other words, variables 1 and 2 are *conditionally independent*, given the level of variable 3.

c. *Three-factor and two two-factor effects absent*

There are also three versions of the model with the three-factor effect and two two-factor effects missing. Selecting

$$u_{123} = u_{12} = u_{13} = 0$$

gives

$$l_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)}. \quad (2.4-13)$$

Variable 1 is now *completely independent* of the other two variables; variables 2 and 3 are associated.

Later theorems show we can collapse any variable that is independent of all other variables without affecting any of the remaining parameters of the subscripted u -terms. For the simple model (2.4-13) we can prove this directly by writing

$$\log(m_{+jk}) = w + u_{2(j)} + u_{3(k)} + u_{23(jk)}, \quad (2.4-14)$$

where

$$w = u + \log\left(\sum_i e^{u_{1(i)}}\right) \quad (2.4-15)$$

is a constant independent of any variable category. Equation (2.4-14) shows that the table of sums $\{m_{+jk}\}$ has the same subscripted u -terms as the overall model (2.4-13).

d. *Three-factor and all two-factor effects absent*

The model that represents *complete independence* of all three variables is

$$l_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)}. \quad (2.4-16)$$

In this model none of the two-dimensional faces $\{m_{ij+}\}$, $\{m_{i+k}\}$, or $\{m_{ij+}\}$ exhibit any interaction. Summing over two variables gives

$$\begin{aligned} m_{i++} &= \exp(u + u_{1(i)}) \sum_{j,k} \exp(u_{2(j)} + u_{3(k)}) \\ &= \exp(u + u_{1(i)}) \sum_j \exp(u_{2(j)}) \sum_k \exp(u_{3(k)}), \end{aligned} \quad (2.4-17)$$

and similarly for each of the other one-dimensional sums. Summing over all three variables, we have for the total number of counts:

$$N = \sum_i m_{i++} = \exp(u) \sum_i \exp(u_{1(i)}) \sum_j \exp(u_{2(j)}) \sum_k \exp(u_{3(k)}). \quad (2.4-18)$$

Combining the three variants of (2.4-17) with (2.4-18), we find

$$m_{ijk} = \frac{m_{i++}m_{+j+}m_{++k}}{N^2}. \quad (2.4-19)$$

e. *Noncomprehensive models*

Proceeding further with deletion of u -terms gives models that are independent of one or more variables, which we call *noncomprehensive* models. Suppose we put $u_3 = 0$. Then the model becomes

$$l_{ijk} = u + u_{1(i)} + u_{2(j)}. \quad (2.4-20)$$

It is apparent that

$$m_{ijk} = \frac{m_{ij+}}{K}$$

and we have the same $I \times J$ array for all k . We can always sum over any variables not included in the model and describe the condensed structure by a *comprehensive* model that includes all the remaining variables in the resulting lower-dimensional array.

f. *Nonhierarchical models*

We defined the hierarchy principle in the Section 2.4.1. Most arrays can be described by the set of hierarchical models. Exceptions do occur, but generally the interpretation of nonhierarchical models is complex. In example 3.7-4 of Chapter 3 there is a $2 \times 2 \times 2$ model with $u_{12} = u_{23} = u_{13} = 0$, but $u_{123} \neq 0$. We show there that this model can be related to the concept of “synergism,” where a response occurs when two factors are present together but not when either occurs alone.

In larger tables with nonhierarchical structure a possible strategy is to partition and look at smaller sections of data. If $u_{123} \neq 0$ but $u_{12} = 0$ when we partition according to the categories of variable 3, then we have some tables with interaction

between variables 1 and 2 in one direction and others with interaction in the opposite direction. We can either consider the structure of each set of tables separately or rearrange the cells to form new compound variables.

Reduction to two dimensions

We derived the three-dimensional model by averaging across a set of two-dimensional tables, each relating to a category of the third variable. We found that all the u -terms involving variables 1 and 2 were averages of the corresponding terms in the two-way tables, as in expression (2.4-2), and all the u -terms involving variable 3 were deviations from these averages, as in expression (2.4-3). We now consider when we can obtain valid estimates of the two-factor terms u_{12} , u_{13} , and u_{23} from the two-way table of sums $\{m_{ij+}\}$, $\{m_{i+k}\}$, and $\{m_{+ik}\}$ formed by adding across such sets of two-way tables.

The rows of table 2.4-1 yield expressions for the sums of the l_{ijk} . For example, adding the first two rows gives

$$l_{+11} = 2(u + u_{2(1)} + u_{3(1)} + u_{23(11)}), \quad (2.4-21)$$

with no u -terms involving variable 1 remaining. These u -terms do not disappear, however, when we sum m_{ijk} over variable 1:

$$\begin{aligned} m_{+jk} &= \exp(u + u_{2(j)} + u_{3(k)} + u_{23(jk)}) \\ &\quad \times \sum_i \exp(u_{1(i)} + u_{12(ij)} + u_{13(ik)} + u_{123(ijk)}) \\ &= \exp(\text{terms independent of variable 1}) \\ &\quad \times \sum_i \exp(\text{terms dependent on variable 1}). \end{aligned} \quad (2.4-22)$$

Consequently, if we describe the table of sums by a saturated log-linear model

$$\log m_{+jk} = v + v_{2(j)} + v_{3(k)} + v_{23(jk)}, \quad (2.4-23)$$

we find that, in general,

$$v_{23(jk)} \neq u_{23(jk)}.$$

If $v_{23(jk)} = u_{23(jk)}$ for all j, k , then we say that in the three-dimensional table, variable 1 is *collapsible* with respect to the two-factor effect u_{23} . We now prove that variable 1 is only collapsible with respect to u_{23} if the $I \times J \times K$ table is described by an unsaturated model with $u_{123} = 0$ and either $u_{12} = 0$ or $u_{13} = 0$ or both.

THEOREM 2.4-1 *In a rectangular three-dimensional table a variable is collapsible with respect to the interaction between the other two variables if and only if it is at least conditionally independent of one of the other two variables given the third.*

Proof Without loss of generality we can consider the model with one interaction absent. We choose the model with $u_{12} = u_{123} = 0$, and so we have

$$l_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)} + u_{23(jk)}. \quad (2.4-24)$$

We can write the logarithms of the marginal sums as

$$\log m_{ij+} = u + u_{1(i)} + u_{2(j)} + \lambda_{ij}, \quad (2.4-25)$$

$$\log m_{i+k} = u + u_{1(i)} + u_{3(k)} + u_{13(ik)} + \lambda_k, \quad (2.4-26)$$

$$\log m_{+jk} = u + u_{2(j)} + u_{3(k)} + u_{23(jk)} + \lambda'_k, \quad (2.4-27)$$

where

$$\begin{aligned}\lambda_{ij} &= \log \left(\sum_k \exp(u_{3(k)} + u_{13(ik)} + u_{23(jk)}) \right), \\ \lambda_k &= \log \left(\sum_j \exp(u_{2(j)} + u_{23(jk)}) \right), \\ \lambda'_k &= \log \left(\sum_i \exp(u_{1(i)} + u_{13(ik)}) \right).\end{aligned}\tag{2.4-28}$$

The subscripts on the terms λ indicate which variables λ depends on. For m_{i+k} and m_{+jk} only one variable is involved, so we can also describe these sums by a saturated four-term model. For instance, if we consider collapsing over variable 2,

$$\log(m_{i+k}) = w + w_{1(i)} + w_{3(k)} + w_{13(ik)}.\tag{2.4-29}$$

We now compare (2.4-26) and (2.4-29). Summing over both i and k gives

$$w = u + \sum_k \frac{\lambda_k}{K}.\tag{2.4-30}$$

Summing over k only, we have

$$w_{1(i)} = u_{1(i)},\tag{2.4-31}$$

and summing over i only gives

$$w_{3(k)} = u_{3(k)} + \lambda_k - \sum_k \frac{\lambda_k}{K}.\tag{2.4-32}$$

From (2.4-30) through (2.4-32), we have

$$w + w_{1(i)} + w_{3(k)} = u + u_{1(i)} + u_{3(k)} + \lambda_k,\tag{2.4-33}$$

and so from (2.4-26) and (2.4-29) we have

$$w_{13(ik)} = u_{13(ik)}.\tag{2.4-34}$$

The analogous result for m_{+jk} is

$$w_{23(jk)} = u_{23(jk)}.\tag{2.4-35}$$

This proves that summing over either of the unrelated variables 1 or 2 gives a table of sums that does not distort the two-factor effects present in the complete array.

Proof of Converse We need only observe from expression (2.4-25) that $\log m_{ij+}$ has a term λ_{ij} dependent on both variables 1 and 2. The table of sums $\{m_{ij+}\}$ thus exhibits a two-factor effect not present in the structural model for the elementary cells. We conclude that summing over a variable associated with both the other variables yields a table of sums that exhibits a different interaction than that shown by its component tables.* ■

* The symbol ■ marks the end of a proof.

Notes on Theorem 2.4-1

From expression (2.4-31) we observe that collapsing over variable 2 when $u_{12} = u_{123} = 0$ preserves the value of u_1 as well as the value of u_{13} . Contrastingly, from (2.4-30) and (2.4-32) we find that u and u_3 are changed by collapsing. Thus if a three-dimensional array is collapsed over a single variable (say variable 2), those u -terms involving a specific remaining variable (say u_1) are unchanged if and only if the variable collapsed over is independent of the specific variable (i.e., $u_{12} = 0$).

Clearly, if variable 2 is independent of variables 1 and 3 jointly, none of u_1 , u_3 , and u_{13} is changed by collapsing over variable 2.

Theorem 2.4-1 in a weaker form is analogous to a result in the theory of partial correlation. We recall a well-known formula from standard multivariate statistical analysis regarding partial correlation coefficients:

$$\rho_{12 \cdot 3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{(1 - \rho_{13}^2)(1 - \rho_{23}^2)}},$$

where $\rho_{12 \cdot 3}$ is the partial correlation between variables 1 and 2, controlling for variable 3, and ρ_{12} , ρ_{23} , and ρ_{13} are the simple correlations. If $\rho_{13} = 0$ or $\rho_{23} = 0$, then $\rho_{12 \cdot 3}$ is a scalar multiple of the ρ_{12} , and we can test the hypothesis $\rho_{12 \cdot 3} = 0$ by testing for $\rho_{12} = 0$.

For three-dimensional contingency tables, theorem 2.4-1 says that the term u_{12} in the three-dimensional log-linear model is the same as the u_{12} -term in the log-linear model for the two-dimensional table $\{m_{ij+}\}$, provided that u_{13} or $u_{23} = 0$ in the three-dimensional model.

Example 2.4-1 Collapsing a table

The data of table 2.4-2, analyzed by Bishop [1969], have been used for class exercises at the Harvard School of Public Health, but the original source is unfortunately lost. They relate to the survival of infants (variable 1) according to the amount of prenatal care received by the mothers (variable 2). The amount of care is classified as "more" or "less." The mothers attend one of two clinics, denoted here as *A* and *B*. Thus we have a three-dimensional array with the clinic as the third variable.

Table 2.4-2 Three-Dimensional Array Relating Survival of Infants to Amount of Prenatal Care Received in Two Clinics

Place where Care Received	Amount of Prenatal Care	Infants' Survival		Mortality Rate (%)
		Died	Survived	
Clinic A	Less	3	176	1.7
	More	4	293	1.4
Clinic B	Less	17	197	7.9
	More	2	23	8.0

Source: Bishop [1969].

The table for mothers who attended clinic *A* has the cross-product ratio

$$\frac{(3)(293)}{(4)(176)} = 1.2$$

and that for mothers who attended clinic *B*,

$$\frac{(17)(23)}{(2)(197)} = 1.0.$$

Both of these values are close to 1, and we conclude that u_{12} is very small. Thus the data could reasonably be considered to be a sample from a population where $u_{12} = 0$; in other words, survival is unrelated to amount of care.

Table 2.4-3 Two-Dimensional Array Relating Survival of Infants to Amount of Prenatal Care; Array Obtained by Pooling Data from Two Clinics

Amount of Prenatal Care	Infants' Survival		Mortality Rate (%)
	Died	Survived	
Less	20	373	5.1
More	6	316	1.9

If we collapse over the third variable (clinic), we obtain table 2.4-3, and we have the cross product

$$\frac{(20)(316)}{(6)(373)} = 2.8,$$

which does not reflect the magnitude of u_{12} . If we were to look only at this table we would erroneously conclude that survival was related to the amount of care received. Theorem 2.4-1 tells us that we cannot evaluate u_{12} from the collapsed table because both u_{13} and u_{23} are nonzero. ■■

2.5 Models for Four or More Dimensions

We proceeded from two dimensions to three dimensions by writing a model for each two-way table defined by the categories of the third variable. Similarly, when we have a four-dimensional array we can write a model in $w^{(l)}$ -terms for each three-way array defined by the L categories of the fourth variable. The averages across three-way arrays give the corresponding u -terms for the overall four-dimensional model, and the deviations from these averages give new terms with subscripts that include variable 4, as shown in table 2.5-1. We can continue this process for any number of dimensions, say s . The expected counts in elementary cells of a four-dimensional array have four subscripts i, j, k, l , and in general in s dimensions they have s subscripts. In going from two to three dimensions we doubled the number of u -terms from 4 (i.e., u, u_1, u_2 , and u_{12}) to 8. When we go to four dimensions another doubling brings the number of u -terms to 16. In general, for s dimensions the log-linear model has 2^s u -terms.

For the saturated s -dimensional model we have s single-factor u -terms, $s - 1$ of them from the first $s - 1$ dimensions and a term representing deviations from the overall mean due to the last variable to be added. All possible combinations

of two variables at a time give $\binom{s}{2}$ two-factor terms. Proceeding thus we find that in general the number of r -factor terms is $\binom{s}{r}$ for $r = 0, \dots, s$ if we define $\binom{s}{0}$ as 1.

When all variables have only two categories, we can readily interpret the parameters of the u -terms as functions of cross-product ratios involving four cells. In larger tables this formulation becomes more difficult, but the interpretation of unsaturated models is independent of the number of categories per variable. Thus it is sufficient for us to consider the interpretation of parameters in the 2^s table and discuss models that are often encountered in any dimension. We follow with a theorem defining in general terms when we can collapse multidimensional tables to fewer dimensions and still assess the magnitude of u -terms of interest from the condensed array.

Table 2.5-1 Relationship of Three-Dimensional w -terms to Four-Dimensional u -terms in an $I \times J \times K \times L$ Array

Order	w-terms for Each of L Three-Way Tables	u -terms	
		Average of w-terms	Deviations from Average ^a
Mean	w	u	u_4
Single-factor	w_1, w_2, w_3	u_1, u_2, u_3	u_{14}, u_{24}, u_{34}
Two-factor	w_{12}, w_{23}, w_{13}	u_{12}, u_{23}, u_{13}	$u_{124}, u_{234}, u_{134}$
Three-factor	w_{123}	u_{123}	u_{1234}

^a Note each deviation is one order higher than other terms on same line.

2.5.1 Parameters in the 2^s table

The additive constraints on all subscripted u -terms ensure that the number of independent parameters in the saturated model is the same as the total number of elementary cells. For the 2^s table each of the 2^s u -terms has one absolute value, and each of these can be expressed as a function of the log odds.

Suppose the array is split into 2^{s-2} two-dimensional tables relating variables 1 and 2. The cross-product ratio in each of these 2×2 tables is $\alpha^{(r)}$, where $r = 1, \dots, 2^{s-2}$, and the tables are ordered so that the subscripts corresponding to the third variable change before the subscripts corresponding to the fourth, and so on. We can now define u_{12} and all higher-order relatives.

Two-factor terms

The two-factor term relating variables 1 and 2 is the average of the two-factor effects in each table:

$$u_{12} = \frac{1}{2^s} \log \left(\prod_r \alpha^{(r)} \right). \quad (2.5-1)$$

Three-factor terms

The definition of the three-factor term u_{123} varies according to the number of dimensions. Using superscript notation, the definition given in expression (2.4-8) for three dimensions becomes

$$u_{123} = \frac{1}{8} \log \frac{\alpha^{(1)}}{\alpha^{(2)}}, \quad (2.5-2)$$

the ratio of cross-product ratios. In four dimensions u_{123} is the average of two such three-dimensional terms:

$$\begin{aligned} u_{123} &= \frac{1}{16} \left(\log \frac{\alpha^{(1)}}{\alpha^{(2)}} + \log \frac{\alpha^{(3)}}{\alpha^{(4)}} \right) \\ &= \frac{1}{16} \log \left(\frac{\alpha^{(1)}\alpha^{(3)}}{\alpha^{(2)}\alpha^{(4)}} \right). \end{aligned} \quad (2.5-3)$$

We can continue for further dimensions, and in general, for s dimensions

$$u_{123} = \frac{1}{2^s} \log \left(\frac{\prod_{i \text{ odd}} \alpha^{(i)}}{\prod_{j \text{ even}} \alpha^{(j)}} \right). \quad (2.5-4)$$

The α -terms derived from tables corresponding to the first category of variable 3 are in the numerator and those corresponding to the second category in the denominator.

Other three-factor terms u_{124} , u_{125} , etc., are similarly defined by selecting α -terms according to the category of the third variable.

Four-factor terms

In four dimensions the term u_{1234} measures differences of two three-factor terms from their average; thus, from (2.5-2) and (2.5-3),

$$\begin{aligned} u_{1234} &= \frac{1}{8} \log \frac{\alpha^{(1)}}{\alpha^{(2)}} - \frac{1}{16} \log \frac{\alpha^{(1)}\alpha^{(3)}}{\alpha^{(2)}\alpha^{(4)}} \\ &= \frac{1}{16} \log \frac{\alpha^{(1)}\alpha^{(4)}}{\alpha^{(2)}\alpha^{(3)}}. \end{aligned} \quad (2.5-5)$$

This expression is a cross-product ratio of the $\{\alpha^{(r)}\}$, themselves cross-product ratios. In five dimensions two such terms are averaged to give

$$u_{1234} = \frac{1}{32} \log \frac{\alpha^{(1)}\alpha^{(4)}\alpha^{(5)}\alpha^{(8)}}{\alpha^{(2)}\alpha^{(3)}\alpha^{(6)}\alpha^{(7)}}. \quad (2.5-6)$$

In s dimensions the general form is

$$u_{1234} = \frac{1}{2^s} \log \left(\prod_r \frac{\alpha^{(4r-3)}\alpha^{(4r)}}{\alpha^{(4r-2)}\alpha^{(4r-1)}} \right), \quad (2.5-7)$$

where the product is taken from $r = 1$ to $r = 2^{s-4}$. In the numerator the $\{\alpha^{(r)}\}$ are derived from those tables where the categories of variables 3 and 4 are the same, while the remaining $\{\alpha^{(r)}\}$ are in the denominator.

Higher-order terms

Hierarchical models require that if a term is not set equal to zero, none of its lower-order relatives are set equal to zero. This requirement is reasonable when we consider expressing u -terms as functions of cross-product ratios. By continuing the process described previously, we can express any u -term involving variables 1 and 2 as a function of the $\{\alpha^{(r)}\}$. Similarly, by partitioning the array in different directions we can express all related terms as functions of a set of cross products.

Even within the set conforming to the hierarchical requirements we have a large choice of models available. We consider in more detail only a few of the many possibilities.

2.5.2 *Uses and interpretation of models*

We can divide the hierarchical models into two broad classes, those with all two-factor effects present and those with at least one two-factor effect absent.

All two-factor effects present

The hypotheses most frequently encountered relate to the independence of variables. Even conditional independence requires that at least one two-factor term is absent. Thus models with all two-factor effects present are more likely to be used not for hypothesis testing but for obtaining elementary cell estimates that are more stable than the observed cell counts. Successively higher-order terms can be regarded as deviations from the average value of related lower-order terms, and so models with only the higher-order terms removed are useful in describing the gross structure of an array. Such models describe general trends, and hence can be regarded as “smoothing” devices.

In other chapters we show that these models are primarily used for

1. obtaining cell estimates for every elementary cell in a sparse array. In Chapter 3 we show that fitting unsaturated models gives estimates for elementary cells that have a positive probability but a zero observed count.
2. detecting outliers. In some circumstances the detection of sporadic cells that are unduly large may be of importance. For example, in some investigations it may be desirable to determine what combination of variable categories gives an excessive number of deaths. In Chapter 4 we describe how to detect cells that show large deviations from a hierarchical model applied to the whole array.

Hierarchical models with one two-factor effect absent

Restriction to the class of hierarchical models still permits many structures with one two-factor effect absent and all others present.

If we put $u_{12} = 0$, in five dimensions the hierarchy principle requires that u_{123} , u_{124} , u_{125} , u_{1234} , u_{1245} , u_{1235} , and u_{12345} are also set equal to zero. Thus the total of 32 terms is reduced by one-fourth to 24. Those remaining fall into three groups: Eight terms involve neither variable 1 nor variable 2, eight involve variable 1 but not variable 2, and eight involve variable 2 but not variable 1. These three groups appear in the first three columns of table 2.5-2, and the fourth column gives u_{12} and its higher-order relatives. We define the sums of the four groups as A , B , C , and D . Thus we have

$$A = u + u_3 + u_4 + u_5 + u_{34} + u_{35} + u_{45} + u_{345}, \quad (2.5-8)$$

and similarly for the other columns. When $u_{12} = 0$, the model describing the five-dimensional structure is

$$l_{ijklm} = A + B + C. \quad (2.5-9)$$

Table 2.5-2 Grouping Terms of Fully Saturated Five-Dimensional Model

Row	Groups			
	A Includes Neither Variable 1 nor 2	B Includes Variable 1, but Not 2	C Includes Variable 2, but Not 1	D Includes Both Variables 1 and 2
1	u	u_1	u_2	u_{12}
2	u_3	u_{13}	u_{23}	u_{123}
3	u_4	u_{14}	u_{24}	u_{124}
4	u_5	u_{15}	u_{25}	u_{125}
5	u_{34}	u_{134}	u_{234}	u_{1234}
6	u_{35}	u_{135}	u_{235}	u_{1235}
7	u_{45}	u_{145}	u_{245}	u_{1245}
8	u_{345}	u_{1345}	u_{2345}	u_{12345}

The sums obtained by adding over variable 1, over variable 2, and over both variables are, respectively,

$$\begin{aligned}
 m_{+jklm} &= \exp(A + C) \sum_i \exp(B), \\
 m_{i+klm} &= \exp(A + B) \sum_j \exp(C), \\
 m_{++klm} &= \exp(A) \sum_i \exp(B) \sum_j \exp(C),
 \end{aligned} \tag{2.5-10}$$

and hence we find

$$m_{ijklm} = \frac{m_{i+klm} m_{+jklm}}{m_{++klm}}. \tag{2.5-11}$$

This is a frequently encountered model. In subsequent chapters we show that the relationship (2.5-11) between the expected elementary cell counts and the sums of counts enables us to fit this model readily. When the array is split into KLM two-way tables relating variables 1 and 2, the margins of each table are members of $\{m_{+jklm}\}$ and $\{m_{i+klm}\}$, and the total in each table is a member of $\{m_{++klm}\}$. Thus expression (2.5-11) describes independence in each table, and we say that variables 1 and 2 are “conditionally independent.”

There are many other models with variables 1 and 2 conditionally independent. Starting at the bottom of column B or C and moving up, any number of terms can be set equal to zero to give a different hierarchical model. When terms in the same row are removed from columns B and C , the term in column A of this row can also be removed. These models all have fewer u -terms than model (2.5-9), but some are more complex in that the expected counts in the elementary cells cannot be derived from sets of sums as in expression (2.5-11). We give rules in Chapter 3 for determining when such relationships exist.

Even within the set of hierarchical models, the hypothesis $u_{12} = 0$ is thus consistent with a variety of structures. Consequently, it is never adequate to describe a hypothesis only in terms of the absence of one u -term. The model underlying the null hypothesis must be stated in full. The verbal interpretation of many high-dimensional models becomes more cumbersome than useful, and the simplest approach is to write out the log-linear model and examine which

terms are included. One of the purposes of such inspection is to determine whether the size of the array can be reduced by summing over some of the variables without distorting the u -terms of interest.

2.5.3 Collapsing arrays

Theorem 2.4-1 deals with collapsing in three dimensions, and states that variable 3 is collapsible with respect to u_{12} if variable 3 is unrelated to either variable 1 or variable 2 or both. Thus in three dimensions at least one two-factor term must be absent for any collapsibility to exist. We now give a general theorem for collapsibility in s dimensions, which indicates for a given model which u -terms remain unchanged in the collapsed table. Following the statement of the theorem, we discuss its implications and consider some examples of its application. We first review the definition of collapsibility.

Definition of collapsibility

We say that the variables we sum over are *collapsible* with respect to specific u -terms when the parameters of the specified u -terms in the original array are identical to those of the same u -terms in the corresponding log-linear model for the reduced array.

THEOREM 2.5-1 *Suppose the variables in an s -dimensional array are divided into three mutually exclusive groups. One group is collapsible with respect to the u -terms involving a second group, but not with respect to the u -terms involving only the third group, if and only if the first two groups are independent of each other (i.e., the u -terms linking them are 0).*

Proof We regard the three groups in the statement of this theorem as being three compound variables. We then apply theorem 2.4-1 to these compound variables, and the result follows. ■

Implications of collapsibility theorems

Independence of two variables implies that the model describing the overall structure has the two-factor term relating the variables and all its higher-order relatives set equal to zero. If a variable is collapsible with respect to specific u -terms, it may be removed by adding over all its categories, or condensed by combining some of its categories, without changing these u -terms.

This definition has two important implications.

1. If all two-factor effects are present, collapsing any variable changes all the u -terms;
2. if any variable is independent of all other variables, it may be removed by summing over its categories without changing any u -terms.

Thus the practice of examining all two-way marginal tables of a complex data base may be very misleading if any of the variables are interrelated. By contrast, the dimensionality of any array may be safely reduced by collapsing over all completely independent variables. The extent to which collapsing or condensing is permissible is determined by the absence of two-factor effects in the structural model, provided the model is hierarchical.

Illustration of collapsing in five dimensions

Suppose we have a five-variable array and wish to know whether we can safely sum over the fifth variable. If $u_{15} = 0$, the hierarchy principle implies that all higher-order relatives are also absent. Theorem 2.5-1 tells us that we can examine the four-dimensional array of sums and obtain valid estimates of all the u -terms involving variable 1, such as u_{12} , u_{123} , and u_{1234} , but we cannot obtain estimates of the terms that do not involve variable 1, such as u_2 , u_{23} , and u_{234} .

Suppose now that we wish to know whether we can safely sum over the fourth and fifth variables. Referring to table 2.5-2, we find that $u_{14} = 0$ and $u_{15} = 0$ imply that all the entries in columns B and D from the third line downward are zero. Theorem 2.5-1 tells us that the terms remaining in columns B and D are unchanged by collapsing over variables 4 and 5, but the other terms in the first two rows are altered.

2.5.4 Irregular tables

In this chapter we have considered the log-linear model as a useful description of data structure and discussed its interpretation and properties. With the exception of a brief discussion of a simple triangular table in Section 2.3.6, we have dealt only with complete tables. Most of the interpretation and properties we discuss are also applicable for incomplete arrays, but difficulties can arise, and elaboration of these is deferred to Chapter 5.

Similarly, we defer examples of rearranging or partitioning a seemingly complete table to form an incomplete array to Chapters 3, 6, and 8. We also defer special interpretations, such as symmetry and marginal homogeneity, to other chapters.

2.6 Exercises

1. If we tried to assess the predictive value $PV+$ of a test from (2.2-44) without knowing the proportion of diseased persons in the population, we would obtain the pseudovalue

$$PPV+ = \frac{m_{11}}{m_{11} + m_{21}}.$$

Show that this is only equal to the true predictive value when $D/(1 - D) = N_1/N_2$.

2. Remembering that $P_{1(1)}$ is sensitivity and $P_{2(2)}$ is specificity, show
 - (i) for a disease prevalence D of 50% and sensitivity equal to specificity, we have $PV+ = P_{1(1)} = P_{2(2)}$;
 - (ii) more generally, the positive predictive value is equal to the sensitivity when $D(1 - P_{1(1)}) = (1 - D)(1 - P_{2(2)})$;
 - (iii) if the positive predictive value is equal to the sensitivity, then the negative predictive value is equal to the specificity.
3. Take a square table with four categories per variable, which is described by the log-linear model with $u_{12} = 0$. Fold the table along the diagonal to obtain a triangular table with expected counts m'_{ij} , where

$$\begin{aligned} m'_{ij} &= m_{ij} + m_{ji} & i \neq j, \\ m'_{ii} &= m_{ii} & i = 1, 2, 3, 4. \end{aligned}$$

Consider the cross product α , defined as

$$\alpha = \frac{m'_{13}m'_{24}}{m'_{14}m'_{23}},$$

and show that a necessary and sufficient condition for $\alpha = 1$ is

$$\frac{m_{1+}}{m_{2+}} = \frac{m_{+1}}{m_{+2}}.$$

Hence show that for an independent two-dimensional structure to exhibit independence when folded, we must have homogeneous margins, i.e., $m_{i+} = m_{+i}$ for all i .

4. Suppose in a $2 \times 3 \times 3$ array the two-dimensional array $\{m_{+jk}\}$ is fixed.

(i) Show that by defining

$$\text{logit}(j, k) = \log \frac{m_{1jk}}{m_{2jk}},$$

the no three-factor effect model can be written

$$\text{logit}(j, k) = w + w_{2(j)} + w_{3(k)}.$$

(ii) How many degrees of freedom are associated with each w -term? Show that the sum for the three terms differs from JK by $(J - 1)(K - 1)$, and compare this difference with the three-dimensional equivalent.

(Answer: The fully saturated three-dimensional model has IJK parameters. The no three-factor effect model differs by $(I - 1)(J - 1)(K - 1)$, which is equal to the logit difference when $I = 2$.)

5. In three dimensions, what structural models permit evaluating each of the three two-factor effects from the corresponding two-dimensional table of sums?

(Answer: Models with three-factor effect and two two-factor effects absent.)

6. In a $2 \times 3 \times 4 \times 5$ array, write down the most highly parametrized model with variable 4 collapsible with respect to u_{123} .

(Answer: Any model with two-factor effect involving variable 4 and higher-order relatives absent will do. We keep most parameters if we choose $u_{14} = 0$.)

7. In four dimensions, if we have a hierarchical model structure and we know $u_{14} = u_{23} = 0$, can we assess any of the other two-factor effects from the corresponding two-way tables of sums?

(Answer: No. If we sum over variable 1 from $\{m_{+jkl}\}$ we can assess u_{24} and u_{34} , but we have w_{23} different from u_{23} . As $w_{23} \neq 0$ we cannot collapse any further.)

2.7 Appendix: The Geometry of a 2×2 Table

In this appendix we discuss structure in a 2×2 table in terms of the geometry of the tetrahedron. In particular, we derive the loci of (i) all points corresponding to tables whose rows and columns are independent, (ii) all points corresponding to tables with a given degree of association as measured by the cross-product ratio, (iii) all points corresponding to tables with a fixed set of marginal totals, and (iv) all points corresponding to tables exhibiting symmetry (or marginal homogeneity).

The geometric ideas discussed here allow us to visualize the properties of the various models discussed in Section 2.2 and are used explicitly in Chapters 11 and

12. The geometric model can also be used to provide a general proof of the convergence of iterative procedures used throughout this book (see Fienberg [1970a]).

2.7.1 The tetrahedron of reference

Suppose we have only two cells with probabilities p_1 and p_2 , where $p_1 + p_2 = 1$. Any value of the $\{p_i\}$ may be represented in two dimensions by a straight line joining the point (0, 1) on the y-axis to the point (1, 0) on the x-axis, as we show in figure 2.7-1a.

When we add another cell, so that our probabilities are p_1 , p_2 , and p_3 , with $p_1 + p_2 + p_3 = 1$, we can represent any set of the $\{p_i\}$ in three dimensions by the surface that joins the points (1, 0, 0), (0, 1, 0), and (0, 0, 1). This surface is a 2-flat (see figure 2.7-1b) and becomes an isosceles triangle when we draw it in two dimensions (see figure 2.7-1c).

Analogously, four cells can be represented in three dimensions by a tetrahedron with vertices (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), and (0, 0, 0, 1). These points, labeled A_1 , A_2 , A_3 , and A_4 in figure 2.7-2, represent extreme 2×2 tables of probabilities with probability 1 in one cell and 0 in the others:

$$A_1 = \begin{array}{|c|c|} \hline 1 & 0 \\ \hline 0 & 0 \\ \hline \end{array} \quad A_2 = \begin{array}{|c|c|} \hline 0 & 1 \\ \hline 0 & 0 \\ \hline \end{array} \quad A_3 = \begin{array}{|c|c|} \hline 0 & 0 \\ \hline 1 & 0 \\ \hline \end{array} \quad A_4 = \begin{array}{|c|c|} \hline 0 & 0 \\ \hline 0 & 1 \\ \hline \end{array}$$

The general point $\mathbf{P} = (p_{11}, p_{12}, p_{21}, p_{22})$ within the tetrahedron corresponds to the general 2×2 table.

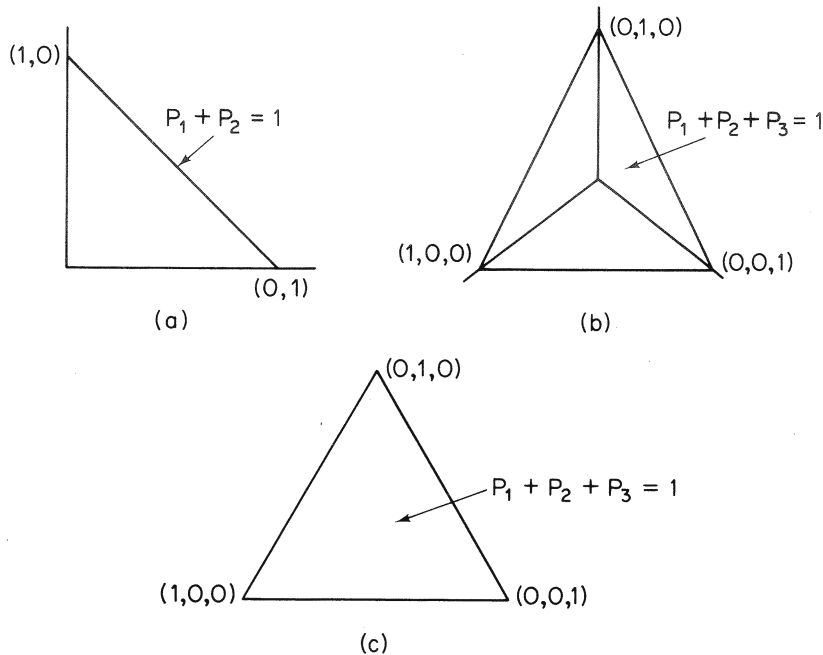


Figure 2.7-1 Triangles of reference for two and three cells. a. Two cells represented by a line in two dimensions. b. Three cells represented by a surface in three dimensions. c. Three cells represented by a surface in two dimensions.

2.7.2 Surface of independence

We can now define a surface in the tetrahedron that gives the locus of all tables exhibiting independence, i.e., those tables for which $\alpha_3 = 1$. (See (2.2-13).)

We take any point \mathbf{T} on the line $\mathbf{A}_1\mathbf{A}_2$, defined by a distance t such that $1 \geq t \geq 0$. By taking weighted averages of \mathbf{A}_1 and \mathbf{A}_2 , we obtain

$$\mathbf{T} = \begin{array}{|c|c|} \hline t & 1 - t \\ \hline 0 & 0 \\ \hline \end{array} \quad (2.7-1)$$

We can similarly choose a point \mathbf{T}' on $\mathbf{A}_3\mathbf{A}_4$ so that

$$\mathbf{T}' = \begin{array}{|c|c|} \hline 0 & 0 \\ \hline t & 1 - t \\ \hline \end{array} \quad (2.7-2)$$

Any point \mathbf{I} on the line \mathbf{TT}' within the tetrahedron corresponds to a second number s such that $1 \geq s \geq 0$, and this point is derived as a weighted average of \mathbf{T} and \mathbf{T}' , so we have

$$\mathbf{I} = \begin{array}{|c|c|} \hline st & s(1 - t) \\ \hline (1 - s)t & (1 - s)(1 - t) \\ \hline \end{array} \quad \begin{array}{l} s \\ 1 - s \end{array} \quad (2.7-3)$$

$t \qquad 1 - t$

The row and column marginal totals are independent, as required. By allowing s and t to take on all possible values between 0 and 1, we can find all points which correspond to tables whose rows and columns are independent. The lines \mathbf{TT}' defined by different values of t ($0 \leq t \leq 1$) lie on this *surface of independence*. Alternatively, we could define the point \mathbf{S} on $\mathbf{A}_1\mathbf{A}_3$ with coordinates $(s, 0, 1 - s, 0)$ and \mathbf{S}' on $\mathbf{A}_2\mathbf{A}_4$ with coordinates $(0, s, 0, 1 - s)$. Any point \mathbf{I}' on \mathbf{SS}' would also have coordinates given by expression (2.7-3). Thus the lines \mathbf{SS}' defined by different values of s ($0 \leq s \leq 1$) also lie on the surface of independence. This surface is completely determined by either family of lines (see figure 2.7-3). The surface of independence is a section of a hyperbolic paraboloid, and its saddle point is at the center of the tetrahedron, $\mathbf{C} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. The hyperbolic paraboloid is a *doubly ruled* surface, since its surface contains two families of straight lines or “rulings.” The tables corresponding to points on any one of the lines \mathbf{TT}' have the same column margins (totals), while the tables corresponding to points on any one of the lines \mathbf{SS}' have the same row margins.

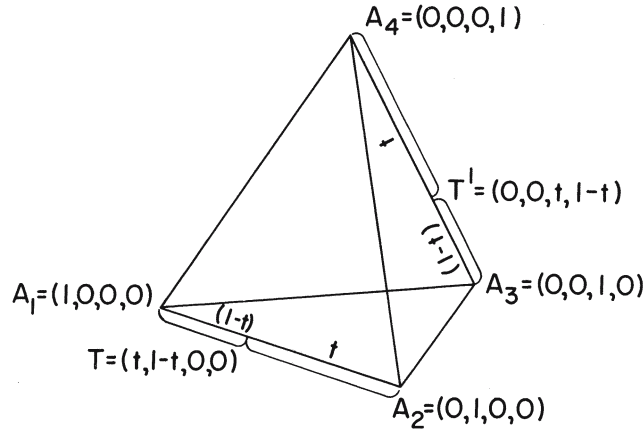


Figure 2.7-2 Tetrahedron of reference. Reproduced from Fienberg and Gilbert [1970].

2.7.3 The surface of constant association

To derive the surface with constant cross-product ratio α_3 , we choose the point \mathbf{T} as before on the line joining \mathbf{A}_1 to \mathbf{A}_2 . On the line joining \mathbf{A}_3 to \mathbf{A}_4 we choose another point \mathbf{T}^* , where

$$\mathbf{T}^* = \begin{array}{|c|c|} \hline 0 & 0 \\ \hline t^* & 1 - t^* \\ \hline \end{array} \quad (2.7-4)$$

and

$$\frac{t}{1-t} = \alpha_3 \frac{t^*}{1-t^*}. \quad (2.7-5)$$

Any point \mathbf{I}^* on the line \mathbf{TT}^* is again a weighted average of \mathbf{T} and \mathbf{T}^* , where the weights are s and $1-s$. Thus we have

$$\mathbf{I}^* = \begin{array}{|c|c|} \hline st & s(1-t) \\ \hline (1-s)t^* & (1-s)(1-t^*) \\ \hline \end{array} \begin{array}{l} s \\ 1-s \end{array} \quad (2.7-6)$$

Expression (2.7-6) gives the cross-product ratio α_3 for any s such that $1 \geq s \geq 0$. We note that the column totals are $t^* + s(t - t^*)$ and $1 - t^* - s(t - t^*)$ and so vary with s . Thus the lines \mathbf{TT}^* which generate the surface of constant association are not the loci of points corresponding to tables with constant column totals as were the lines \mathbf{TT}' . The surface of constant α_3 intersects the surface of independence along the two edges $\mathbf{A}_1\mathbf{A}_2$ and $\mathbf{A}_3\mathbf{A}_4$ of the tetrahedron (see figure 2.7-4). We can similarly generate another surface of constant α_3 by choosing points \mathbf{S} on $\mathbf{A}_1\mathbf{A}_3$ and \mathbf{S}^* on $\mathbf{A}_2\mathbf{A}_4$. The surfaces of constant α_3 are sections of hyperboloids of one sheet and are again doubly ruled surfaces.

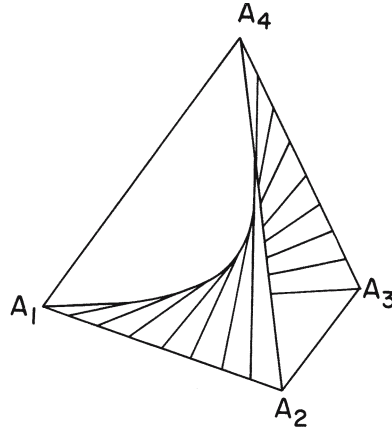


Figure 2.7-3 Surface of independence defined by family of lines TT' . Reproduced from Fienberg and Gilbert [1970].

We previously showed that α_3 was a function of u_{12} , and that by rearranging the cells of the fourfold table we could derive α_1 as a function of u_1 and α_2 as a function of u_2 . It follows that by choosing points on different pairs of edges of the tetrahedron we can generate surfaces for constant values of α_1 and α_2 and so for constant values of u_1 and u_2 . The lines TT^* and SS^* are the intersections of the surface of constant α_3 with surfaces of constant α_1 and α_2 , respectively.

2.7.4 Line of constant margins

A table with fixed margins t and $1 - t$ for rows and s and $1 - s$ for columns can be written as the point \mathbf{P}_a , where

$$\mathbf{P}_a = \begin{array}{cc|c} \hline st + a & (1 - s)t - a & t \\ \hline s(1 - t) - a & (1 - s)(1 - t) + a & 1 - t \\ \hline s & 1 - s & \\ \hline \end{array} \quad (2.7-7)$$

for any a that gives nonnegative probabilities. To determine the locus of \mathbf{P}_a we find the limiting values of a that give zero cell entries:

$$\begin{aligned} a_1 &= -st, \\ a_2 &= (1 - s)t, \\ a_3 &= s(1 - t), \\ a_4 &= -(1 - s)(1 - t), \end{aligned} \quad (2.7-8)$$

and obtain the corresponding points $\mathbf{P}^{(1)}$, $\mathbf{P}^{(2)}$, $\mathbf{P}^{(3)}$, and $\mathbf{P}^{(4)}$, for each of which different conditions must be satisfied by s and t . The coordinates and restrictions

on the $\mathbf{P}^{(i)}$ are

$$\begin{aligned}
 \mathbf{P}^{(1)} &= (0, t, s, 1 - s - t) & s + t &\leq 1, \\
 \mathbf{P}^{(2)} &= (t, 0, s - t, 1 - s) & s &\geq t, \\
 \mathbf{P}^{(3)} &= (s, t - s, 0, 1 - t) & t &\geq s, \\
 \mathbf{P}^{(4)} &= (s + t - 1, 1 - s, 1 - t, 0) & s + t &\geq 1.
 \end{aligned} \tag{2.7-9}$$

If s and t are such that the conditions for $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$ are satisfied, $\mathbf{P}^{(1)}$ is a point on the surface $\mathbf{A}_2\mathbf{A}_3\mathbf{A}_4$ of the tetrahedron, and $\mathbf{P}^{(2)}$ is a point on the surface $\mathbf{A}_1\mathbf{A}_3\mathbf{A}_4$. We choose a point \mathbf{W} on the line $\mathbf{P}^{(1)}\mathbf{P}^{(2)}$ corresponding to a distance w , where $1 \geq w \geq 0$, i.e.,

$$\mathbf{W} = \begin{array}{|c|c|c|} \hline wt & (1-w)t & t \\ \hline s-wt & 1-s-t+wt & 1-t \\ \hline \end{array} \tag{2.7-10}$$

Then we may easily verify that $\mathbf{P}^{(3)}$ and $\mathbf{P}^{(4)}$ are also on the line defined by \mathbf{W} for varying values of w . \mathbf{W} lies on the line of constant margins, and we can confirm that it goes through the surface of independence by putting $w = s$.

Fienberg and Gilbert [1970] show that the line of constant margins is orthogonal to $\mathbf{A}_1\mathbf{A}_4$ and $\mathbf{A}_2\mathbf{A}_3$ and is parallel to the line connecting the mid-points of $\mathbf{A}_1\mathbf{A}_4$ and $\mathbf{A}_2\mathbf{A}_3$. Hence the line of constant margins is not perpendicular to the surface of independence unless the marginal totals all equal $1/2$.

When we have homogeneous margins we put $s = t$ and derive the locus of the point \mathbf{W}^* (for varying w and t), where

$$\mathbf{W}^* = \begin{array}{|c|c|c|} \hline wt & (1-w)t & t \\ \hline (1-w)t & 1-2t+wt & 1-t \\ \hline \end{array}$$

$t \qquad 1-t$
 \mathbf{A}_4

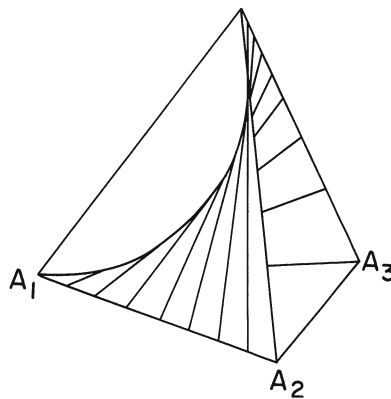


Figure 2.7-4 Surface of constant α ($\alpha = 3$) defined by family of lines \mathbf{TT}^* . Reproduced from Fienberg and Gilbert [1970].

All tables exhibiting marginal homogeneity correspond to points on the intersection of the tetrahedron with the plane through $\mathbf{A}_1\mathbf{A}_4$ and the midpoint of $\mathbf{A}_2\mathbf{A}_3$.

For further details on geometric interpretations, we refer the reader to Fienberg [1968] and Fienberg and Gilbert [1970].

Discrete Multivariate Analysis

Theory and Practice

Bishop, Y.M.; Fienberg, S.E.; Holland, P.W.

2007, IX, 559 p., Softcover

ISBN: 978-0-387-72805-6