Discrete Multivariate Analysis

Yvonne M. M. Bishop, Stephen E. Fienberg, and Paul W. Holland
with the collaboration of
Richard J. Light and Frederick Mosteller

Discrete Multivariate Analysis
Theory and Practice

Springer

Yvonne M. Bishop
Washington, DC 20015-2956
Ymbishop@verizon.net

Stephen E. Fienberg
Department of statistics
Carnegie-Mellon University
Pittsburgh, PA 15213
Fienberg@stat.cmu.edu

Paul W. Holland
Educational Testing Service
Princeton, NJ 08541
pholland@ets.org

ISBN 978-0-387-72805-6

Library of Congress Control Number: 2007928365

Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1

springer.com

Preface

The analysis of discrete multivariate data, especially in the form of cross-classifications, has occupied a prominent place in the statistical literature since the days of Karl Pearson and Sir R. A. Fisher. Although Maurice Bartlett's pioneering paper on testing for absence of second-order interaction in $2 \times 2 \times 2$ tables was published in 1935, the widespread development and use of methods for the analysis of multidimensional cross-classified data had to await the general availability of high-speed computers. As a result, in the last ten years statistical journals, as well as those in the biological, social, and medical sciences, have devoted increasing space to papers dealing with the analysis of discrete multivariate data. Many statisticians have contributed to this progress, as a glance at the reference list will quickly reveal. We point, especially, to the sustained and outstanding contributions of Joseph Berkson, M. W. Birch, I. J. Good, Leo A. Goodman, James E. Grizzle, Marvin Kastenbaum, Gary G. Koch, Solomon Kullback, H. O. Lancaster, Nathan Mantel, and R. L. Plackett.

The one person most responsible for our interest in and continued work on the analysis of cross-classified data is Frederick Mosteller. It is not an overstatement to say that without his encouragement and support in all phases of our effort, this book would not exist. Our interest in the analysis of cross-classified data goes back to 1964 and the problems which arose during and after Mosteller's work on the National Halothane study. This work led directly to the doctoral dissertations of two of us (Bishop and Fienberg), as well as to a number of published papers. But Fred's contributions to this book are more than just encouragement; he has read and copiously commented on nearly every chapter, and while we take complete responsibility for the final manuscript, if it has any virtues they are likely to be due to him.

Richard Light enthusiastically participated in the planning of this book, and offered comments on several chapters. He prepared the earlier drafts of Chapter 11, Measures of Association and Agreement, and he made the major effort on the final version of this chapter.

We owe a great debt to many of our colleagues and students who have commented on parts of our manuscript, made valuable suggestions on aspects of our research, and generally stimulated our interest in the subject. Those to whom we are indebted include Raj Bahadur, Darrell Bock, Tar Chen, William Cochran, Joel Cohen, Arthur Dempster, O. Dudley Duncan, Hillel Einhorn, Robert Fay, John Gilbert, Anne Goldman, Shelby Haberman, David Hoaglin, Nathan Keyfitz, William Kruskal, Kinley Larntz, Siu-Kai Lee, Lincoln Moses, I. R. Savage, Thomas Schoener, Michael Sutherland, John Tukey, David Wallace, James Warram, Sanford Weisberg, Janet Wittes, and Jane Worcester.

For the production of the manuscript we are indebted to Holly Grano, Kathi Hirst, Carol Lambert, and Mary Jane Schleupner.

The National Science Foundation has provided substantial support for our research and writing under grant GS-32327X1 to Harvard University. We have also received extensive support from other research grants. These include: research grants CA-06516 from the National Cancer Institute and RR-05526 from the Division of Research Facilities and Resources, National Institutes of Health to the Children's Cancer Research Foundation; National Science Foundation research grants GP-16071, GS-1905, and a grant from the Statistics Branch, Office of Naval Research, to the Department of Statistics, University of Chicago, as well as a grant from the Alfred P. Sloan Foundation to the Department of Theoretical Biology, University of Chicago; National Science Foundation research grant GJ-1154X to the National Bureau of Economic Research, Inc., and a faculty research grant from the Social Science Research Council to Paul W. Holland.

Earlier versions of material in several chapters appeared in *The Annals of Statistics*, *Biometrics*, *Biometrika*, and *The Journal of the American Statistical Association.*

Brookline, Massachusetts Y.M.M.B.
New Brighton, Minnesota S.E.F.
Hingham, Massachusetts P.W.H.

February 1974

CONTENTS

1 Introduction

1.1 The Need

The scientist searching for structure in large systems of data finds inspiration in his own discipline, support from modern computing, and guidance from statistical models. Because large sets of data are likely to be complicated, and because so many approaches suggest themselves, a codification of techniques of analysis, regarded as attractive paths rather than as straitjackets, offers the scientist valuable directions to try. In statistical specialities such as regression and the analysis of variance, codifications are widely available and sometimes keyed to special disciplines. In discrete multivariate statistics, however, the excellent guides already available, for example Cox [1970], Fleiss [1973], Good [1965], Lancaster [1969], and Maxwell [1961], stop short of giving a systematic treatment of large contingency tables, and especially tables that have troublesome irregularities. This book offers such a treatment.

1.2 Why a Book?

The literature on discrete multivariate analysis, although extensive, unfortunately is widely scattered. This book brings that literature together in an organized way. Although we do report a few new results here, that is not our primary purpose. Our purpose is to organize the materials needed by both theoretical and practical workers so that key ideas stand out. By presenting parametric models, sampling schemes, basic theory, practical examples, and advice on computation, this book serves as a ready reference for various users.

To bring together both the theory and practice of discrete multivariate analysis, a good deal of space is required. We need to relate various techniques of analysis, many of which are quite close to one another in both concept and result, so that the practitioner can know when one method is essentially the same as another, and when it is not. We need to provide basic theory, both for understanding and to lay a basis for new variations in the analysis when conditions do not match the ones presented here.

When we deal with several variables simultaneously, the practical examples we analyze tend to be large—larger than those ordinarily treated in the standard texts and monographs. An exploratory analysis of a set of data often leads us to perform several separate parallel analyses. Sometimes one analysis suggests another. Furthermore, we are obliged to discuss computing to some extent because these large-scale analyses are likely to require iterative methods, which are best done by high-speed computers. The availability of high-speed computing facilities has encouraged investigators to gather and ready for analysis substantial

sets of data. Applications and examples play a central role in most of the chapters in this book, and they take considerable space because we illustrate calculations, present alternative analyses, and discuss the conclusions the practitioner might draw for various data sets.

These reasons all lead to a treatment of book length.

1.3 Different Users

The applied statistician or quantitative research worker looking for comprehensive analyses of discrete multivariate data will find here a variety of ways to attack both standard and nonstandard sets of data. As a result, he has available a systematic approach to the analysis of multiway contingency tables. Naturally, new difficulties or constraints raise new problems, but the availability of a flexible approach should strengthen the practitioner's hand, just as the ready availability of analysis of variance and regression methods has for other data. He will understand his computer output better and know what kinds of computer analyses to ask for.

By skillful use of one computer program for obtaining estimates, the researcher can solve a wide range of problems. By juxtaposing practical examples from a variety of fields, the researcher can gain insight into his own problem by recognizing similarities to and differences from problems that arise in other fields. We have therefore varied the subject matter of the illustrations as well as the size of the examples. We have found the methods described in this book useful for small as well as large data sets.

On many occasions we have helped other people analyze sets of discrete multivariate data. In such consulting work we have found some of the material in this book helpful in guiding the practitioner to suitable analyses. Of course, several of the examples included here are drawn directly from our consulting experiences.

Parts of several chapters have grown out of material used in different university courses or sets of lectures we have given. Some of these courses and lectures stressed the application of statistical methods and were aimed at biological, medical, or social scientists with less preparation than a one-year course in statistics. Others stressed the statistical theory at various graduate levels of mathematical and statistical sophistication.

For the student we have included some exercise work involving both the manipulation of formulas and the analysis of additional data sets. In the last few years, certain examples have been analyzed repeatedly in the statistical literature, gradually bringing us a better understanding of what various methods accomplish. By making more examples of varied character readily available, we hope this healthy tradition of reanalyzing old problems with new methods will receive a substantial boost.

Finally, of course, we expect the book to provide a reference source for the methods collected in it. Although we do not try to compete with the fine bibliography provided by Lancaster [1969], some of the papers we cite have appeared since the publication of that work.

1.4 Sketch of the Chapters

Although each chapter has its own introduction, we present here a brief description of the contents and purposes of each chapter. We have organized the chapters

into three logical groups of unequal size. The first group introduces the log-linear model, presents the statistical theory underlying its use in the analysis of contingency-table data, and illustrates the application of the theory to a wide variety of substantive problems. The second group of chapters deals with approaches and methods not relying directly on the log-linear model. The final pair of chapters contains basic statistical results and theory used throughout the book.

Section 1 Log-Linear Models, Maximum Likelihood Estimation, and Their Application

Chapter 2

With one exception, the example on the relation of survival of mothers to prenatal care, this is a theoretical chapter. It is meant for the practitioner as well as the theoretician, although they may read it from different points of view. The chapter develops the notation and ideas for the log-linear model used so extensively in this book. It begins with two-by-two (2×2) tables of counts and works up to tables of four or more dimensions. The emphasis is, first, on describing structure rather than sampling, second, on the relation of the log-linear model approach to familiar techniques for testing independence in two-way contingency tables, and third, on the generalization of these ideas to several dimensions. Fourth, the chapter shows the variety of models possible in these higher dimensions.

Chapter 3 (Preparation for most readers: Chapter 2)

Although from its title this chapter sounds like a theoretical one, its main emphasis is on important practical devices for computing estimates required for the analysis of multidimensional tables of counts. These devices include both how to recognize when simple direct estimates for cells are and are not available, and how to carry out iterative fitting when they are not. The chapter explains how to count degrees of freedom, occasionally a trickier problem than many of us are used to. All aspects of the analysis—models, estimates, calculation, degrees of freedom, and interpretation—are illustrated with concrete examples, drawn from history, sociology, political science, public health, and medicine, and given in enough detail that the practical reader should now have the main thrust of the method.

Chapter 4 (Preparation: Chapter 3)

Chapter 4 provides tools and approaches to the selection of models for fitting multidimensional tables. While it includes some theory, this is the key chapter for the reader who must actually deal with applications and choose models to describe data.

First, it summarizes the important large sample results on chi square goodness-of-fit statistics. (These matters are more fully treated in Chapter 14.) Second, it explains how the partitioning of chi square quantities leads to tests of special hypotheses and to the possibility of more refined inferences. Third, although there is no "best" way to select a model, the chapter describes several different approaches to selecting log-linear models for multidimensional tables, using large-sample results and the partitioning method.

Chapter 5 (*Preparation: Chapter 3*)

Cell counts can be zero either because of sampling variability even when observations can occur in a cell, or because constraints make the cell automatically zero (for example, the losing football team does not score more than the winning team). In addition, for some problems, certain cells are not expected to fit smoothly into a simple model for the table of counts, so the counts for these cells, although available, are set aside for special treatment. One danger is that the researcher may not recognize that he is afflicted with an incomplete table.

This profusely illustrated chapter offers standard ways of handling such incomplete tables and is oriented to problems of estimation, model selection, counting of degrees of freedom, and applications.

Chapter 6 (*Preparation: Chapter 5*)

This chapter deals with a special application: If, as sometimes happens, we have several samplings or censuses, we may wish to estimate a total count. For example, we may have several lists of voluntary organizations from the telephone book, newspaper articles, and other sources. Although each list may be incomplete, from the several lists we want to estimate the total number of voluntary organizations (including those on *none* of these lists). This chapter offers ways to solve such multiple-census problems by treating the data sets as incomplete multidimensional tables. The method is one generalization of the capture-recapture method of estimation used in wildlife and other sampling operations.

Chapter 7
(*Preparation: Chapter 3*)

Since Markov chains depend on a relation between the results at one stage and those at later stages, there are formal similarities with contingency tables. Consequently, analysis of Markov chains using the log-linear model is an attractive possibility, treated here along with other methods.

This is a practical chapter, containing illustrations which come from market research and studies of political attitudes, language patterns (including bird songs), number-guessing behavior, and interpersonal relations.

Chapter 8
(*Preparation: Chapter 3 and some of Chapter 5*)

Although square tables are discussed elsewhere, this chapter focuses on the special questions of symmetry and marginal homogeneity. These problems arise in panel studies when the same criteria are used at each point in time and in psychological studies, as when both members of a pair are classified simultaneously according to the same criteria. The chapter gives methods for assessing symmetry and homogeneity, illustrated with practical examples. It also treats multidimensional extensions of the notions of symmetry and marginal homogeneity, relating them to the basic approach of this book using log-linear models.

Chapter 9

For practitioners, this chapter gives numerous suggestions and examples about how to get started in building models for data. The attitude is that of exploratory data analysis rather than confirmatory analysis. When models are fitted, the

problem still remains of how to view the fit. What if it is too good? How bad is too bad? Rough approximations may be just what is desired. The beginner may wish to start reading the book with this chapter.

Section II Related models and methods

Chapter 10

Although this book offers a systematic treatment of contingency tables through the approach of log-linear models and maximum likelihood estimation, the reader may want to know what alternative methods are available. This chapter offers introductions to some of the more widely used of these methods and points the reader to further literature.

Chapter 11

This chapter differs from the others in the book because it deals only with two-way tables, and also because the main thrust of measuring association is to summarize many parameters in one. The basic principles for the choice and use of measures of association depend on the purposes of the user. The chapter also treats a special case of association, referred to as "agreement." Because we view interaction and association as due primarily to many different parameters, this chapter presents a different outlook than does the rest of the book.

Chapter 12

In a contingency table of many dimensions, the number of cells is often high while the average number of observations per cell in many practical problems is small, and so many cells may have zero entries. We wish to estimate the probabilities associated with the cells. Extra information about these probabilities may be available from the general distribution of the counts or from the margins. Bayesian approaches offer methods of estimating these probabilities, but they usually leave to the user the problem of choosing the parameter of the prior distribution, a job he may be ill equipped to do. Theoretical investigations can help the reader choose by showing him some methods that will protect him from bad errors. This chapter reviews the literature and provides new results, together with some applications. The treatment deals with the case where the cells merely form a list, and the case where the cells form a complete two-way table.

Section III Theoretical background

Chapter 13

In working with contingency tables, both the practitioner and the theorist face certain standard sampling distributions repeatedly: the binomial, Poisson, multinomial, hypergeometric, and negative binomial distributions, and some of their generalizations. This chapter offers a ready source of reference for information about these distributions.

Chapter 14

In discrete problems, exact calculations are notoriously difficult, especially when the sample sizes are large. This difficulty makes the results of this chapter especially important, since asymptotic (large-sample) methods are so widely used in this work.

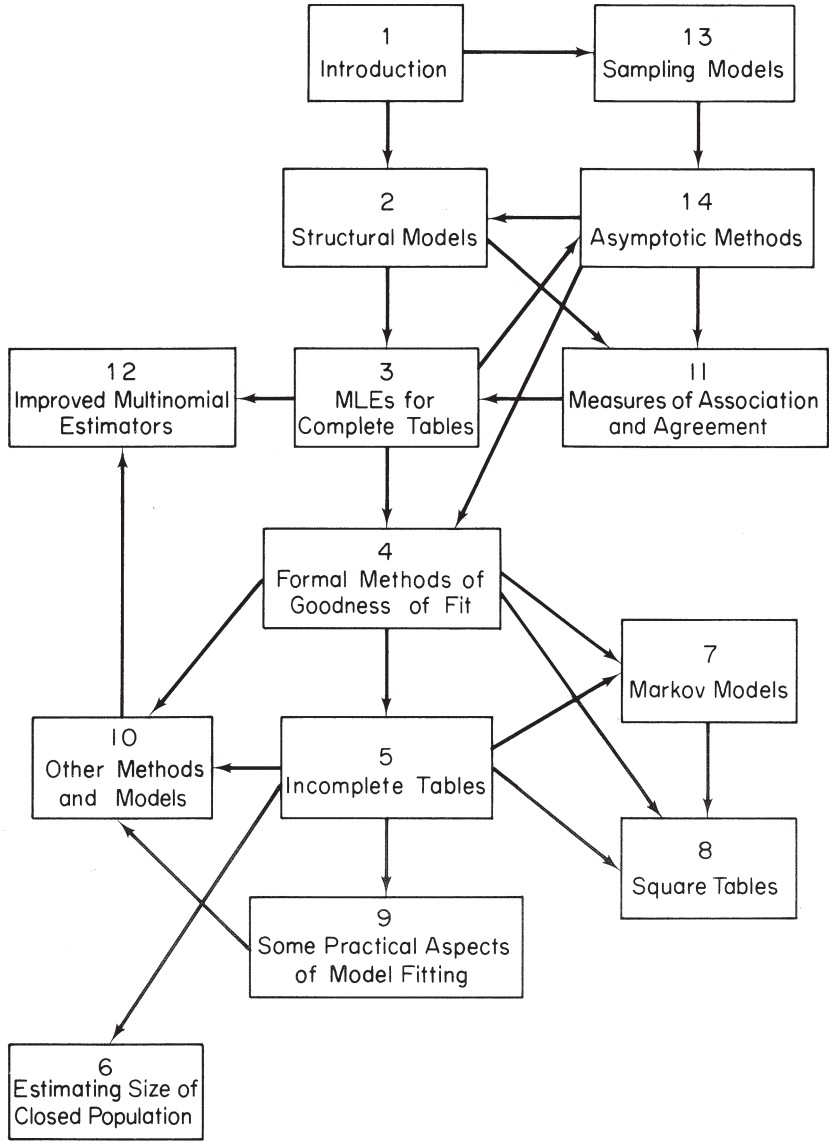
1
Introduction
13
Sampling Models
2
Structural Models
14
Asymptotic Methods
12
Improved Multinomial
Estimators
3
MLEs for
Complete Tables
11
Measures of Association
and Agreement
4
Formal Methods of
Goodness of Fit
7
Markov Models
10
Other Methods
and Models
5
Incomplete Tables
8
Square Tables
9
Some Practical Aspects
of Model Fitting
6
Estimating Size of
Closed Population

Figure 1.6-1.

The mathematics of asymptotic methods useful to the theoretician is scattered throughout the literature. For getting asymptotic distributions, mathematical orders of magnitude play an important role. This chapter provides a convenient short course in the devices and theorems about the O, o notation, for sequences of real numbers and about the analogous O_p, o_p notation, for random variables. The treatment includes vector random variables. The material is illustrated with examples and practice exercises, enabling the student to derive many of the theoretical results in this book. Moreover, the devices discussed are useful in every branch of theoretical statistics.

1.5 Computer Programs

As we have already noted, one general-purpose computer program can be used to carry out most of the calculations described in this book. Many researchers working with multiway table data have prepared such programs to carry out estimation using the method of iterative proportional fitting and to compute various goodness-of-fit statistics. These programs are now available at a large number of computer centers and research installations. We refer those who would like to use programs which are not available at their institutions to the Fortran listings in Bishop [1967, appendix I] and Haberman [1972, 1973b].

1.6 How to Proceed from Here

Readers of this book come to it with different backgrounds and different interests. We have ordered Chapters 2 through 9 so that each chapter builds only on preceding ones, with the exception that most of them use material from the two theoretical background chapters (Chapters 13 and 14). Thus a good basic sequence consists of Chapters 2 through 9. Nevertheless, readers may choose to work with chapters in different orders.

Graduate students in theoretical statistics may choose to begin with a review of the sampling distribution properties and large-sample theory in Chapters 13 and 14, then proceed to Chapters 2, 3, 4, 5, 9, 10, 11, and 12. Chapters 6, 7, and 8 can be handled either after Chapter 5 or at the end of the indicated sequence.

Quantitative biological or social scientists interested in analyzing their own data will most likely profit from a quick reading of Chapters 2 and 3, followed by Chapter 9 (Sections 9.1–9.5) and Chapters 4 and 5. Then they might turn to Chapter 7, 8, or 11, depending on their interests.

Other sequences of chapters come to mind. Figure 1.6-1 gives a schematic representation of alternative orderings for different readers and an indication of how the chapters are linked.

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

$$\{p_i\} = (p_1, p_2, \ldots, p_t) \tag{2.1-1}$$

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

$$\{x_i\} = (x_1, x_2, \ldots, x_t), \tag{2.1-2}$$

where

$$\sum x_i = N.$$

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

$$\{m_i\} = (m_1, m_2, \ldots, m_t), \tag{2.1-3}$$

where

$$E(x_i) = m_i \quad \text{for } i = 1, \ldots, t,$$

$$m_i = Np_i.$$

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 artifactitious 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:

		B	
		1	2
A	1	p_{11}	p_{12}
	2	p_{21}	p_{22}

(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

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

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:

$$p_{i+} = \sum_{k=1}^{2} p_{ik} \qquad i = 1, 2 \tag{2.2-3}$$

$$p_{+j} = \sum_{k=1}^{2} p_{kj} \qquad j = 1, 2. \tag{2.2-4}$$

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):

		B 1	2	Totals
A	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

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

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

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

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

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

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

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

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

		A_2: 1	A_2: 2
A_1: 1		p_{11}	p_{12}
A_1: 2		p_{21}	p_{22}

(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
A_1	1	p_{11}	p_{12}
	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

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

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

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

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

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

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

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

If we specify values for these three contrasts and recall that

$$\sum p_{ij} = 1, \tag{2.2-19}$$

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

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

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

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

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

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

and similarly for the jth level of the second variable:

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

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

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

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

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

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

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

Similarly, the main effects are

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

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

and the interaction term becomes

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

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

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

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

$$u_{2(1)} = \tfrac{1}{4}(\log p_{11} - \log p_{12} + \log p_{21} - \log p_{22})$$
$$= \tfrac{1}{4}\log \alpha_2, \tag{2.2-31}$$

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

$$u_{12(11)} = \tfrac{1}{4}(\log p_{11} - \log p_{12} - \log p_{21} + \log p_{22})$$
$$= \tfrac{1}{4}\log \alpha_3. \tag{2.2-32}$$

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

$$\log p_{11} = u + \tfrac{1}{4}\log \alpha_1 + \tfrac{1}{4}\log \alpha_2 + \tfrac{1}{4}\log \alpha_3, \tag{2.2-33}$$

and hence

$$p_{11} = \lambda \alpha_1' \alpha_2' \alpha_3', \tag{2.2-34}$$

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	A_2: 2	Totals
A_1	1	$\lambda\alpha_1'\alpha_2'\alpha_3'$	$\dfrac{\lambda\alpha_1'}{\alpha_2'\alpha_3'}$	$\lambda\alpha_1'\left(\alpha_2'\alpha_3' + \dfrac{1}{\alpha_2'\alpha_3'}\right)$
	2	$\dfrac{\lambda\alpha_2'}{\alpha_1'\alpha_3'}$	$\dfrac{\lambda\alpha_3'}{\alpha_1'\alpha_2'}$	$\dfrac{\lambda}{\alpha_1'}\left(\dfrac{\alpha_2'}{\alpha_3'} + \dfrac{\alpha_3'}{\alpha_2'}\right)$
Totals		$\lambda\alpha_2'\left(\alpha_1'\alpha_3' + \dfrac{1}{\alpha_1'\alpha_3'}\right)$	$\dfrac{\lambda}{\alpha_2'}\left(\dfrac{\alpha_1'}{\alpha_3'} + \dfrac{\alpha_3'}{\alpha_1'}\right)$	1

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

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

If we set $\alpha_1' = 1$, which is equivalent to setting $u_1 = 0$, the condition (2.2-35) becomes

$$\left(\alpha_2' - \frac{1}{\alpha_2'}\right)\left(\alpha_3' - \frac{1}{\alpha_3'}\right) = 0, \tag{2.2-36}$$

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

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

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

		B		
		1	2	Totals
A	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

$$\begin{aligned} m_{11} + m_{12} &= N_1, \\ m_{21} + m_{22} &= N_2, \\ N_1 + N_2 &= N. \end{aligned}$$

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

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

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

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

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
No Disease	m_{21}	m_{22}	N_2

(2.2-41)

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
No Disease	m_{22}	m_{21}	N_2

(2.2-42)

Now each row yields one of the proportions of interest:

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

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}
No	m_{21}	m_{12}

(2.2-44)

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	$DP_{1(1)}$	$DP_{2(1)}$
No Disease	$(1 - D)P_{1(2)}$	$(1 - D)P_{2(2)}$

(2.2-45)

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

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

and for the negative laboratory results

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

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, $P_{1(i)} = m_{i1}/m_{i+}$ for $i = 1, 2$.

Then the logit for the ith row is defined as

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

From the saturated model

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

* 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

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

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 $w_{1(j)}$ 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

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

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

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

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

$$\begin{aligned} -\tfrac{1}{4} \leqq \beta_i \leqq \tfrac{1}{4}, \\ -\tfrac{1}{4} \leqq \gamma_j \leqq \tfrac{1}{4}, \\ -\tfrac{1}{4} \leqq \varepsilon_{ij} \leqq \tfrac{1}{4}. \end{aligned} \tag{2.2-54}$$

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

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

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.