Categorical Data Analysis

TRADING SOFTWARE

FOR SALE & EXCHANGE

www.trading-software-collection.com

***Subscribe** for FREE download more stuff.*

Mirrors:

www.forex-warez.com
www.traders-software.com

Contacts

andreybbrv@gmail.com
andreybbrv@hotmail.com
andreybbrv@yandex.ru
Skype: andreybbrv
ICQ: 70966433

Categorical Data Analysis

Second Edition

ALAN AGRESTI

University of Florida
Gainesville, Florida

A JOHN WILEY & SONS, INC., PUBLICATION

This book is printed on acid-free paper. ♾

Published simultaneously in Canada.

For ordering and customer service, call 1-800-CALL-WILEY.

Library of Congress Cataloging-in-Publication Data Is Available

ISBN 0-471-36093-7

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To Jacki

Contents

Preface **xiii**

1. Introduction: Distributions and Inference for Categorical Data **1**

1.1 Categorical Response Data, 1
1.2 Distributions for Categorical Data, 5
1.3 Statistical Inference for Categorical Data, 9
1.4 Statistical Inference for Binomial Parameters, 14
1.5 Statistical Inference for Multinomial Parameters, 21
Notes, 26
Problems, 28

2. Describing Contingency Tables **36**

2.1 Probability Structure for Contingency Tables, 36
2.2 Comparing Two Proportions, 43
2.3 Partial Association in Stratified 2×2 Tables, 47
2.4 Extensions for $I \times J$ Tables, 54
Notes, 59
Problems, 60

3. Inference for Contingency Tables **70**

3.1 Confidence Intervals for Association Parameters, 70
3.2 Testing Independence in Two-Way Contingency Tables, 78
3.3 Following-Up Chi-Squared Tests, 80
3.4 Two-Way Tables with Ordered Classifications, 86
3.5 Small-Sample Tests of Independence, 91

3.6 Small-Sample Confidence Intervals for 2×2 Tables,* 98
3.7 Extensions for Multiway Tables and Nontabulated Responses, 101
Notes, 102
Problems, 104

4. Introduction to Generalized Linear Models 115

4.1 Generalized Linear Model, 116
4.2 Generalized Linear Models for Binary Data, 120
4.3 Generalized Linear Models for Counts, 125
4.4 Moments and Likelihood for Generalized Linear Models,* 132
4.5 Inference for Generalized Linear Models, 139
4.6 Fitting Generalized Linear Models, 143
4.7 Quasi-likelihood and Generalized Linear Models,* 149
4.8 Generalized Additive Models,* 153
Notes, 155
Problems, 156

5. Logistic Regression 165

5.1 Interpreting Parameters in Logistic Regression, 166
5.2 Inference for Logistic Regression, 172
5.3 Logit Models with Categorical Predictors, 177
5.4 Multiple Logistic Regression, 182
5.5 Fitting Logistic Regression Models, 192
Notes, 196
Problems, 197

6. Building and Applying Logistic Regression Models 211

6.1 Strategies in Model Selection, 211
6.2 Logistic Regression Diagnostics, 219
6.3 Inference About Conditional Associations in $2 \times 2 \times K$ Tables, 230
6.4 Using Models to Improve Inferential Power, 236
6.5 Sample Size and Power Considerations,* 240
6.6 Probit and Complementary Log-Log Models,* 245

*Sections marked with an asterisk are less important for an overview.

6.7 Conditional Logistic Regression and Exact Distributions,* 250

Notes, 257

Problems, 259

7. Logit Models for Multinomial Responses 267

7.1 Nominal Responses: Baseline-Category Logit Models, 267

7.2 Ordinal Responses: Cumulative Logit Models, 274

7.3 Ordinal Responses: Cumulative Link Models, 282

7.4 Alternative Models for Ordinal Responses,* 286

7.5 Testing Conditional Independence in $I \times J \times K$ Tables,* 293

7.6 Discrete-Choice Multinomial Logit Models,* 298

Notes, 302

Problems, 302

8. Loglinear Models for Contingency Tables 314

8.1 Loglinear Models for Two-Way Tables, 314

8.2 Loglinear Models for Independence and Interaction in Three-Way Tables, 318

8.3 Inference for Loglinear Models, 324

8.4 Loglinear Models for Higher Dimensions, 326

8.5 The Loglinear–Logit Model Connection, 330

8.6 Loglinear Model Fitting: Likelihood Equations and Asymptotic Distributions,* 333

8.7 Loglinear Model Fitting: Iterative Methods and their Application,* 342

Notes, 346

Problems, 347

9. Building and Extending Loglinear/Logit Models 357

9.1 Association Graphs and Collapsibility, 357

9.2 Model Selection and Comparison, 360

9.3 Diagnostics for Checking Models, 366

9.4 Modeling Ordinal Associations, 367

9.5 Association Models,* 373

9.6 Association Models, Correlation Models, and Correspondence Analysis,* 379

9.7 Poisson Regression for Rates, 385
9.8 Empty Cells and Sparseness in Modeling Contingency Tables, 391
Notes, 398
Problems, 400

10. Models for Matched Pairs **409**

10.1 Comparing Dependent Proportions, 410
10.2 Conditional Logistic Regression for Binary Matched Pairs, 414
10.3 Marginal Models for Square Contingency Tables, 420
10.4 Symmetry, Quasi-symmetry, and Quasi-independence, 423
10.5 Measuring Agreement Between Observers, 431
10.6 Bradley–Terry Model for Paired Preferences, 436
10.7 Marginal Models and Quasi-symmetry Models for Matched Sets,* 439
Notes, 442
Problems, 444

11. Analyzing Repeated Categorical Response Data **455**

11.1 Comparing Marginal Distributions: Multiple Responses, 456
11.2 Marginal Modeling: Maximum Likelihood Approach, 459
11.3 Marginal Modeling: Generalized Estimating Equations Approach, 466
11.4 Quasi-likelihood and Its GEE Multivariate Extension: Details,* 470
11.5 Markov Chains: Transitional Modeling, 476
Notes, 481
Problems, 482

12. Random Effects: Generalized Linear Mixed Models for Categorical Responses **491**

12.1 Random Effects Modeling of Clustered Categorical Data, 492
12.2 Binary Responses: Logistic-Normal Model, 496
12.3 Examples of Random Effects Models for Binary Data, 502
12.4 Random Effects Models for Multinomial Data, 513

12.5 Multivariate Random Effects Models for Binary Data, 516
12.6 GLMM Fitting, Inference, and Prediction, 520
Notes, 526
Problems, 527

13. Other Mixture Models for Categorical Data* 538

13.1 Latent Class Models, 538
13.2 Nonparametric Random Effects Models, 545
13.3 Beta-Binomial Models, 553
13.4 Negative Binomial Regression, 559
13.5 Poisson Regression with Random Effects, 563
Notes, 565
Problems, 566

14. Asymptotic Theory for Parametric Models 576

14.1 Delta Method, 577
14.2 Asymptotic Distributions of Estimators of Model Parameters and Cell Probabilities, 582
14.3 Asymptotic Distributions of Residuals and Goodness-of-Fit Statistics, 587
14.4 Asymptotic Distributions for Logit/Loglinear Models, 592
Notes, 594
Problems, 595

15. Alternative Estimation Theory for Parametric Models 600

15.1 Weighted Least Squares for Categorical Data, 600
15.2 Bayesian Inference for Categorical Data, 604
15.3 Other Methods of Estimation, 611
Notes, 615
Problems, 616

16. Historical Tour of Categorical Data Analysis* 619

16.1 Pearson–Yule Association Controversy, 619
16.2 R. A. Fisher's Contributions, 622

16.3 Logistic Regression, 624
16.4 Multiway Contingency Tables and Loglinear Models, 625
16.5 Recent (and Future?) Developments, 629

Appendix A. Using Computer Software to Analyze Categorical Data 632

A.1 Software for Categorical Data Analysis, 632
A.2 Examples of SAS Code by Chapter, 634

Appendix B. Chi-Squared Distribution Values 654

References 655

Examples Index 689

Author Index 693

Subject Index 701

Preface

The explosion in the development of methods for analyzing categorical data that began in the 1960s has continued apace in recent years. This book provides an overview of these methods, as well as older, now standard, methods. It gives special emphasis to generalized linear modeling techniques, which extend linear model methods for continuous variables, and their extensions for multivariate responses.

Today, because of this development and the ubiquity of categorical data in applications, most statistics and biostatistics departments offer courses on categorical data analysis. This book can be used as a text for such courses. The material in Chapters 1–7 forms the heart of most courses. Chapters 1–3 cover distributions for categorical responses and traditional methods for two-way contingency tables. Chapters 4–7 introduce logistic regression and related logit models for binary and multicategory response variables. Chapters 8 and 9 cover loglinear models for contingency tables. Over time, this model class seems to have lost importance, and this edition reduces somewhat its discussion of them and expands its focus on logistic regression.

In the past decade, the major area of new research has been the development of methods for repeated measurement and other forms of clustered categorical data. Chapters 10–13 present these methods, including marginal models and generalized linear mixed models with random effects. Chapters 14 and 15 present theoretical foundations as well as alternatives to the maximum likelihood paradigm that this text adopts. Chapter 16 is devoted to a historical overview of the development of the methods. It examines contributions of noted statisticians, such as Pearson and Fisher, whose pioneering efforts—and sometimes vocal debates—broke the ground for this evolution.

Every chapter of the first edition has been extensively rewritten, and some substantial additions and changes have occurred. The major differences are:

- A new Chapter 1 that introduces distributions and methods of inference for categorical data.
- A unified presentation of models as special cases of generalized linear models, starting in Chapter 4 and then throughout the text.

- Greater emphasis on logistic regression for binary response variables and extensions for multicategory responses, with Chapters 4–7 introducing models and Chapters 10–13 extending them for clustered data.
- Three new chapters on methods for clustered, correlated categorical data, increasingly important in applications.
- A new chapter on the historical development of the methods.
- More discussion of "exact" small-sample procedures and of conditional logistic regression.

In this text, I interpret *categorical data analysis* to refer to methods for categorical response variables. For most methods, explanatory variables can be qualitative or quantitative, as in ordinary regression. Thus, the focus is intended to be more general than contingency table analysis, although for simplicity of data presentation, most examples use contingency tables. These examples are often simplistic, but should help readers focus on understanding the methods themselves and make it easier for them to replicate results with their favorite software.

Special features of the text include:

- More than 100 analyses of "real" data sets.
- More than 600 exercises at the end of the chapters, some directed towards theory and methods and some towards applications and data analysis.
- An appendix that shows, by chapter, the use of SAS for performing analyses presented in this book.
- Notes at the end of each chapter that provide references for recent research and many topics not covered in the text.

Appendix A summarizes statistical software needed to use the methods described in this text. It shows how to use SAS for analyses included in the text and refers to a web site (www.stat.ufl.edu/~aa/cda/cda.html) that contains (1) information on the use of other software (such as R, S-plus, SPSS, and Stata), (2) data sets for examples in the form of complete SAS programs for conducting the analyses, (3) short answers for many of the odd-numbered exercises, (4) corrections of errors in early printings of the book, and (5) extra exercises. I recommend that readers refer to this appendix or specialized manuals while reading the text, as an aid to implementing the methods.

I intend this book to be accessible to the diverse mix of students who take graduate-level courses in categorical data analysis. But I have also written it with practicing statisticians and biostatisticians in mind. I hope it enables them to catch up with recent advances and learn about methods that sometimes receive inadequate attention in the traditional statistics curriculum.

The development of new methods has influenced—and been influenced by—the increasing availability of data sets with categorical responses in the social, behavioral, and biomedical sciences, as well as in public health, human genetics, ecology, education, marketing, and industrial quality control. And so, although this book is directed mainly to statisticians and biostatisticians, I also aim for it to be helpful to methodologists in these fields.

Readers should possess a background that includes regression and analysis of variance models, as well as maximum likelihood methods of statistical theory. Those not having much theory background should be able to follow most methodological discussions. Sections and subsections marked with an asterisk are less important for an overview. Readers with mainly applied interests can skip most of Chapter 4 on the theory of generalized linear models and proceed to other chapters. However, the book has distinctly higher technical level and is more thorough and complete than my lower-level text, *An Introduction to Categorical Data Analysis* (Wiley, 1996).

I thank those who commented on parts of the manuscript or provided help of some type. Special thanks to Bernhard Klingenberg, who read several chapters carefully and made many helpful suggestions, Yongyi Min, who constructed many of the figures and helped with some software, and Brian Caffo, who helped with some examples. Many thanks to Rosyln Stone and Brian Marx for each reviewing half the manuscript and Brian Caffo, I-Ming Liu, and Yongyi Min for giving insightful comments on several chapters. Thanks to Constantine Gatsonis and his students for using a draft in a course at Brown University and providing suggestions. Others who provided comments on chapters or help of some type include Patricia Altham, Wicher Bergsma, Jane Brockmann, Brent Coull, Al DeMaris, Regina Dittrich, Jianping Dong, Herwig Friedl, Ralitza Gueorguieva, James Hobert, Walter Katzenbeisser, Harry Khamis, Svend Kreiner, Joseph Lang, Jason Liao, Mojtaba Ganjali, Jane Pendergast, Michael Radelet, Kenneth Small, Maura Stokes, Tom Ten Have, and Rongling Wu. I thank my co-authors on various projects, especially Brent Coull, Joseph Lang, James Booth, James Hobert, Brian Caffo, and Ranjini Natarajan, for permission to use material from those articles. Thanks to the many who reviewed material or suggested examples for the first edition, mentioned in the Preface of that edition. Thanks also to Wiley Executive Editor Steve Quigley for his steadfast encouragement and facilitation of this project. Finally, thanks to my wife Jacki Levine for continuing support of all kinds, despite the many days this work has taken from our time together.

ALAN AGRESTI

Gainesville, Florida
November 2001

Categorical Data Analysis, Second Edition. Alan Agresti

ISBN: 0-471-36093-7

CHAPTER 1

Introduction: Distributions and Inference for Categorical Data

From helping to assess the value of new medical treatments to evaluating the factors that affect our opinions and behaviors, analysts today are finding myriad uses for categorical data methods. In this book we introduce these methods and the theory behind them.

Statistical methods for categorical responses were late in gaining the level of sophistication achieved early in the twentieth century by methods for continuous responses. Despite influential work around 1900 by the British statistician Karl Pearson, relatively little development of models for categorical responses occurred until the 1960s. In this book we describe the early fundamental work that still has importance today but place primary emphasis on more recent modeling approaches. Before outlining the topics covered, we describe the major types of categorical data.

1.1 CATEGORICAL RESPONSE DATA

A *categorical variable* has a measurement scale consisting of a set of categories. For instance, political philosophy is often measured as liberal, moderate, or conservative. Diagnoses regarding breast cancer based on a mammogram use the categories normal, benign, probably benign, suspicious, and malignant.

The development of methods for categorical variables was stimulated by research studies in the social and biomedical sciences. Categorical scales are pervasive in the social sciences for measuring attitudes and opinions. Categorical scales in biomedical sciences measure outcomes such as whether a medical treatment is successful.

Although categorical data are common in the social and biomedical sciences, they are by no means restricted to those areas. They frequently

occur in the behavioral sciences (e.g., type of mental illness, with the categories schizophrenia, depression, neurosis), epidemiology and public health (e.g., contraceptive method at last intercourse, with the categories none, condom, pill, IUD, other), genetics (type of allele inherited by an offspring), zoology (e.g., alligators' primary food preference, with the categories fish, invertebrate, reptile), education (e.g., student responses to an exam question, with the categories correct and incorrect), and marketing (e.g., consumer preference among leading brands of a product, with the categories brand A, brand B, and brand C). They even occur in highly quantitative fields such as engineering sciences and industrial quality control. Examples are the classification of items according to whether they conform to certain standards, and subjective evaluation of some characteristic: how soft to the touch a certain fabric is, how good a particular food product tastes, or how easy to perform a worker finds a certain task to be.

Categorical variables are of many types. In this section we provide ways of classifying them and other variables.

1.1.1 Response–Explanatory Variable Distinction

Most statistical analyses distinguish between *response* (or *dependent*) *variables* and *explanatory* (or *independent*) *variables*. For instance, regression models describe how the mean of a response variable, such as the selling price of a house, changes according to the values of explanatory variables, such as square footage and location. In this book we focus on methods for categorical response variables. As in ordinary regression, explanatory variables can be of any type.

1.1.2 Nominal–Ordinal Scale Distinction

Categorical variables have two primary types of scales. Variables having categories without a natural ordering are called *nominal*. Examples are religious affiliation (with the categories Catholic, Protestant, Jewish, Muslim, other), mode of transportation to work (automobile, bicycle, bus, subway, walk), favorite type of music (classical, country, folk, jazz, rock), and choice of residence (apartment, condominium, house, other). For nominal variables, the order of listing the categories is irrelevant. The statistical analysis does not depend on that ordering.

Many categorical variables *do* have ordered categories. Such variables are called *ordinal*. Examples are size of automobile (subcompact, compact, midsize, large), social class (upper, middle, lower), political philosophy (liberal, moderate, conservative), and patient condition (good, fair, serious, critical). Ordinal variables have ordered categories, but distances between categories are unknown. Although a person categorized as moderate is more liberal than a person categorized as conservative, no numerical value describes *how much more* liberal that person is. Methods for ordinal variables utilize the category ordering.

An *interval variable* is one that *does* have numerical distances between any two values. For example, blood pressure level, functional life length of television set, length of prison term, and annual income are interval variables. (An internal variable is sometimes called a *ratio variable* if ratios of values are also valid.)

The way that a variable is measured determines its classification. For example, "education" is only nominal when measured as public school or private school; it is ordinal when measured by highest degree attained, using the categories none, high school, bachelor's, master's, and doctorate; it is interval when measured by number of years of education, using the integers $0, 1, 2, \ldots$.

A variable's measurement scale determines which statistical methods are appropriate. In the measurement hierarchy, interval variables are highest, ordinal variables are next, and nominal variables are lowest. Statistical methods for variables of one type can also be used with variables at higher levels but not at lower levels. For instance, statistical methods for nominal variables can be used with ordinal variables by ignoring the ordering of categories. Methods for ordinal variables cannot, however, be used with nominal variables, since their categories have no meaningful ordering. It is usually best to apply methods appropriate for the actual scale.

Since this book deals with categorical responses, we discuss the analysis of nominal and ordinal variables. The methods also apply to interval variables having a small number of distinct values (e.g., number of times married) or for which the values are grouped into ordered categories (e.g., education measured as < 10 years, 10–12 years, > 12 years).

1.1.3 Continuous–Discrete Variable Distinction

Variables are classified as *continuous* or *discrete*, according to the number of values they can take. Actual measurement of all variables occurs in a discrete manner, due to precision limitations in measuring instruments. The continuous–discrete classification, in practice, distinguishes between variables that take lots of values and variables that take few values. For instance, statisticians often treat discrete interval variables having a large number of values (such as test scores) as continuous, using them in methods for continuous responses.

This book deals with certain types of discretely measured responses: (1) nominal variables, (2) ordinal variables, (3) discrete interval variables having relatively few values, and (4) continuous variables grouped into a small number of categories.

1.1.4 Quantitative–Qualitative Variable Distinction

Nominal variables are *qualitative*—distinct categories differ in quality, not in quantity. Interval variables are *quantitative*—distinct levels have differing amounts of the characteristic of interest. The position of ordinal variables in

the quantitative–qualitative classification is fuzzy. Analysts often treat them as qualitative, using methods for nominal variables. But in many respects, ordinal variables more closely resemble interval variables than they resemble nominal variables. They possess important quantitative features: Each category has a *greater* or *smaller* magnitude of the characteristic than another category; and although not possible to measure, an underlying continuous variable is usually present. The political philosophy classification (liberal, moderate, conservative) crudely measures an inherently continuous characteristic.

Analysts often utilize the quantitative nature of ordinal variables by assigning numerical scores to categories or assuming an underlying continuous distribution. This requires good judgment and guidance from researchers who use the scale, but it provides benefits in the variety of methods available for data analysis.

1.1.5 Organization of This Book

The models for categorical response variables discussed in this book resemble regression models for continuous response variables; however, they assume binomial, multinomial, or Poisson response distributions instead of normality. Two types of models receive special attention, logistic regression and loglinear models. Ordinary *logistic regression models*, also called *logit models*, apply with *binary* (i.e., two-category) responses and assume a binomial distribution. Generalizations of logistic regression apply with multicategory responses and assume a multinomial distribution. *Loglinear models* apply with count data and assume a Poisson distribution. Certain equivalences exist between logistic regression and loglinear models.

The book has four main units. In the first, Chapters 1 through 3, we summarize descriptive and inferential methods for univariate and bivariate categorical data. These chapters cover discrete distributions, methods of inference, and analyses for measures of association. They summarize the non-model-based methods developed prior to about 1960.

In the second and primary unit, Chapters 4 through 9, we introduce models for categorical responses. In Chapter 4 we describe a class of *generalized linear models* having models of this text as special cases. We focus on models for binary and count response variables. Chapters 5 and 6 cover the most important model for binary responses, logistic regression. In Chapter 7 we present generalizations of that model for nominal and ordinal multicategory response variables. In Chapter 8 we introduce the modeling of multivariate categorical response data and show how to represent association and interaction patterns by loglinear models for counts in the table that cross-classifies those responses. In Chapter 9 we discuss model building with loglinear and related logistic models and present some related models.

In the third unit, Chapters 10 through 13, we discuss models for handling repeated measurement and other forms of clustering. In Chapter 10 we

present models for a categorical response with matched pairs; these apply, for instance, with a categorical response measured for the same subjects at two times. Chapter 11 covers models for more general types of repeated categorical data, such as longitudinal data from several times with explanatory variables. In Chapter 12 we present a broad class of models, *generalized linear mixed models*, that use random effects to account for dependence with such data. In Chapter 13 further extensions and applications of the models from Chapters 10 through 12 are described.

The fourth and final unit is more theoretical. In Chapter 14 we develop asymptotic theory for categorical data models. This theory is the basis for large-sample behavior of model parameter estimators and goodness-of-fit statistics. Maximum likelihood estimation receives primary attention here and throughout the book, but Chapter 15 covers alternative methods of estimation, such as the Bayesian paradigm. Chapter 16 stands alone from the others, being a historical overview of the development of categorical data methods.

Most categorical data methods require extensive computations, and statistical software is necessary for their effective use. In Appendix A we discuss software that can perform the analyses in this book and show the use of SAS for text examples. See the Web site *www.stat.ufl.edu/~aa/cda/cda.html* to download sample programs and data sets and find information about other software.

Chapter 1 provides background material. In Section 1.2 we review the key distributions for categorical data: the binomial, multinomial, and Poisson. In Section 1.3 we review the primary mechanisms for statistical inference, using maximum likelihood. In Sections 1.4 and 1.5 we illustrate these by presenting significance tests and confidence intervals for binomial and multinomial parameters.

1.2 DISTRIBUTIONS FOR CATEGORICAL DATA

Inferential data analyses require assumptions about the random mechanism that generated the data. For regression models with continuous responses, the normal distribution plays the central role. In this section we review the three key distributions for categorical responses: *binomial*, *multinomial*, and *Poisson*.

1.2.1 Binomial Distribution

Many applications refer to a fixed number n of binary observations. Let $y_1, y_2, \ldots, y_n$ denote responses for n independent and identical trials such that $P(Y_i = 1) = \pi$ and $P(Y_i = 0) = 1 - \pi$. We use the generic labels "success" and "failure" for outcomes 1 and 0. *Identical trials* means that the probability of success π is the same for each trial. *Independent trials* means

that the $\{Y_i\}$ are independent random variables. These are often called *Bernoulli trials*. The total number of successes, $Y = \sum_{i=1}^{n} Y_i$, has the *binomial distribution* with index n and parameter π, denoted by bin(n,π).

The probability mass function for the possible outcomes y for Y is

$$p(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \qquad y = 0, 1, 2, \ldots, n, \tag{1.1}$$

where the binomial coefficient $\binom{n}{y} = n!/[y!\,(n-y)!]$. Since $E(Y_i) = E(Y_i^2)$ $= 1 \times \pi + 0 \times (1-\pi) = \pi$,

$$E(Y_i) = \pi \quad \text{and} \quad \text{var}(Y_i) = \pi(1-\pi).$$

The binomial distribution for $Y = \sum_i Y_i$ has mean and variance

$$\mu = E(Y) = n\pi \quad \text{and} \quad \sigma^2 = \text{var}(Y) = n\pi(1-\pi).$$

The skewness is described by $E(Y-\mu)^3/\sigma^3 = (1-2\pi)/\sqrt{n\pi(1-\pi)}$. The distribution converges to normality as n increases, for fixed π.

There is no guarantee that successive binary observations are independent or identical. Thus, occasionally, we will utilize other distributions. One such case is sampling binary outcomes without replacement from a finite population, such as observations on gender for 10 students sampled from a class of size 20. The *hypergeometric distribution*, studied in Section 3.5.1, is then relevant. In Section 1.2.4 we mention another case that violates these binomial assumptions.

1.2.2 Multinomial Distribution

Some trials have more than two possible outcomes. Suppose that each of n independent, identical trials can have outcome in any of c categories. Let $y_{ij} = 1$ if trial i has outcome in category j and $y_{ij} = 0$ otherwise. Then $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ic})$ represents a multinomial trial, with $\sum_j y_{ij} = 1$; for instance, $(0, 0, 1, 0)$ denotes outcome in category 3 of four possible categories. Note that y_{ic} is redundant, being linearly dependent on the others. Let $n_j = \sum_i y_{ij}$ denote the number of trials having outcome in category j. The counts $(n_1, n_2, \ldots, n_c)$ have the *multinomial distribution.*

Let $\pi_j = P(Y_{ij} = 1)$ denote the probability of outcome in category j for each trial. The multinomial probability mass function is

$$p(n_1, n_2, \ldots, n_{c-1}) = \left(\frac{n!}{n_1!\,n_2!\cdots n_c!}\right) \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c}. \tag{1.2}$$

Since $\sum_j n_j = n$, this is $(c-1)$-dimensional, with $n_c = n - (n_1 + \cdots + n_{c-1})$. The binomial distribution is the special case with $c = 2$.

For the multinomial distribution,

$$E(n_j) = n\pi_j, \qquad \text{var}(n_j) = n\pi_j(1 - \pi_j), \qquad \text{cov}(n_j, n_k) = -n\pi_j\pi_k. \tag{1.3}$$

We derive the covariance in Section 14.1.4. The marginal distribution of each n_j is binomial.

1.2.3 Poisson Distribution

Sometimes, count data do not result from a fixed number of trials. For instance, if y = number of deaths due to automobile accidents on motorways in Italy during this coming week, there is no fixed upper limit n for y (as you are aware if you have driven in Italy). Since y must be a nonnegative integer, its distribution should place its mass on that range. The simplest such distribution is the *Poisson*. Its probabilities depend on a single parameter, the mean μ. The Poisson probability mass function (Poisson 1837, p. 206) is

$$p(y) = \frac{e^{-\mu}\mu^y}{y!}, \qquad y = 0, 1, 2, \ldots . \tag{1.4}$$

It satisfies $E(Y) = \text{var}(Y) = \mu$. It is unimodal with mode equal to the integer part of μ. Its skewness is described by $E(Y-\mu)^3/\sigma^3 = 1/\sqrt{\mu}$. The distribution approaches normality as μ increases.

The Poisson distribution is used for counts of events that occur randomly over time or space, when outcomes in disjoint periods or regions are independent. It also applies as an approximation for the binomial when n is large and π is small, with $\mu = n\pi$. So if each of the 50 million people driving in Italy next week is an independent trial with probability 0.000002 of dying in a fatal accident that week, the number of deaths Y is a bin(50000000, 0.000002) variate, or approximately Poisson with $\mu = n\pi = 50{,}000{,}000(0.000002) = 100$.

A key feature of the Poisson distribution is that its variance equals its mean. Sample counts vary more when their mean is higher. When the mean number of weekly fatal accidents equals 100, greater variability occurs in the weekly counts than when the mean equals 10.

1.2.4 Overdispersion

In practice, count observations often exhibit variability exceeding that predicted by the binomial or Poisson. This phenomenon is called *overdispersion.* We assumed above that each person has the same probability of dying in a fatal accident in the next week. More realistically, these probabilities vary,

due to factors such as amount of time spent driving, whether the person wears a seat belt, and geographical location. Such variation causes fatality counts to display more variation than predicted by the Poisson model.

Suppose that Y is a random variable with variance $\text{var}(Y|\mu)$ for given μ, but μ itself varies because of unmeasured factors such as those just described. Let $\theta = E(\mu)$. Then unconditionally,

$$E(Y) = E\big[E(Y|\mu)\big], \qquad \text{var}(Y) = E\big[\text{var}(Y|\mu)\big] + \text{var}\big[E(Y|\mu)\big].$$

When Y is conditionally Poisson (given μ), for instance, then $E(Y) = E(\mu) = \theta$ and $\text{var}(Y) = E(\mu) + \text{var}(\mu) = \theta + \text{var}(\mu) > \theta$.

Assuming a Poisson distribution for a count variable is often too simplistic, because of factors that cause overdispersion. The *negative binomial* is a related distribution for count data that permits the variance to exceed the mean. We introduce it in Section 4.3.4.

Analyses assuming binomial (or multinomial) distributions are also sometimes invalid because of overdispersion. This might happen because the true distribution is a mixture of different binomial distributions, with the parameter varying because of unmeasured variables. To illustrate, suppose that an experiment exposes pregnant mice to a toxin and then after a week observes the number of fetuses in each mouse's litter that show signs of malformation. Let n_i denote the number of fetuses in the litter for mouse i. The mice also vary according to other factors that may not be measured, such as their weight, overall health, and genetic makeup. Extra variation then occurs because of the variability from litter to litter in the probability π of malformation. The distribution of the number of fetuses per litter showing malformations might cluster near 0 and near n_i, showing more dispersion than expected for binomial sampling with a single value of π. Overdispersion could also occur when π varies among fetuses in a litter according to some distribution (Problem 1.12). In Chapters 4, 12, and 13 we introduce methods for data that are overdispersed relative to binomial and Poisson assumptions.

1.2.5 Connection between Poisson and Multinomial Distributions

In Italy this next week, let y_1 = number of people who die in automobile accidents, y_2 = number who die in airplane accidents, and y_3 = number who die in railway accidents. A Poisson model for (Y_1, Y_2, Y_3) treats these as independent Poisson random variables, with parameters (μ_1, μ_2, μ_3). The joint probability mass function for $\{Y_i\}$ is the product of the three mass functions of form (1.4). The total $n = \sum Y_i$ also has a Poisson distribution, with parameter $\sum \mu_i$.

With Poisson sampling the total count n is random rather than fixed. If we assume a Poisson model but condition on n, $\{Y_i\}$ no longer have Poisson distributions, since each Y_i cannot exceed n. Given n, $\{Y_i\}$ are also no longer independent, since the value of one affects the possible range for the others.

For c independent Poisson variates, with $E(Y_i) = \mu_i$, let's derive their conditional distribution given that $\Sigma Y_i = n$. The conditional probability of a set of counts $\{n_i\}$ satisfying this condition is

$$\begin{aligned} P\Big[(Y_1 = n_1, Y_2 = n_2, \ldots, Y_c = n_c) \,\Big|\, \sum Y_j = n\Big] \\ = \frac{P(Y_1 = n_1, Y_2 = n_2, \ldots, Y_c = n_c)}{P\big(\Sigma Y_j = n\big)} \\ = \frac{\Pi_i\big[\exp(-\mu_i)\,\mu_i^{n_i}/n_i!\big]}{\exp\big(-\Sigma\mu_j\big)\big(\Sigma\mu_j\big)^n/n!} = \frac{n!}{\Pi_i n_i!}\prod_i \pi_i^{n_i}, \end{aligned} \tag{1.5}$$

where $\{\pi_i = \mu_i/(\Sigma\mu_j)\}$. This is the multinomial $(n, \{\pi_i\})$ distribution, characterized by the sample size n and the probabilities $\{\pi_i\}$.

Many categorical data analyses assume a multinomial distribution. Such analyses usually have the same parameter estimates as those of analyses assuming a Poisson distribution, because of the similarity in the likelihood functions.

1.3 STATISTICAL INFERENCE FOR CATEGORICAL DATA

The choice of distribution for the response variable is but one step of data analysis. In practice, that distribution has unknown parameter values. In this section we review methods of using sample data to make inferences about the parameters. Sections 1.4 and 1.5 cover binomial and multinomial parameters.

1.3.1 Likelihood Functions and Maximum Likelihood Estimation

In this book we use *maximum likelihood* for parameter estimation. Under weak regularity conditions, such as the parameter space having fixed dimension with true value falling in its interior, maximum likelihood estimators have desirable properties: They have large-sample normal distributions; they are asymptotically consistent, converging to the parameter as n increases; and they are asymptotically efficient, producing large-sample standard errors no greater than those from other estimation methods.

Given the data, for a chosen probability distribution the *likelihood function* is the probability of those data, treated as a function of the unknown parameter. The maximum likelihood (ML) estimate is the parameter value that maximizes this function. This is the parameter value under which the data observed have the highest probability of occurrence. The parameter value that maximizes the likelihood function also maximizes the log of that function. It is simpler to maximize the log likelihood since it is a sum rather than a product of terms.

We denote a parameter for a generic problem by β and its ML estimate by $\hat{\beta}$. The likelihood function is $\ell(\beta)$ and the log-likelihood function is $L(\beta) = \log[\ell(\beta)]$. For many models, $L(\beta)$ has concave shape and $\hat{\beta}$ is the point at which the derivative equals 0. The ML estimate is then the solution of the likelihood equation, $\partial L(\beta)/\partial\beta = 0$. Often, β is multidimensional, denoted by $\boldsymbol{\beta}$, and $\hat{\boldsymbol{\beta}}$ is the solution of a set of likelihood equations.

Let SE denote the standard error of $\hat{\beta}$, and let $\text{cov}(\hat{\boldsymbol{\beta}})$ denote the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$. Under regularity conditions (Rao 1973, p. 364), $\text{cov}(\hat{\boldsymbol{\beta}})$ is the inverse of the *information matrix*. The (j, k) element of the information matrix is

$$-E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_j \partial\beta_k}\right). \tag{1.6}$$

The standard errors are the square roots of the diagonal elements for the inverse information matrix. The greater the curvature of the log likelihood, the smaller the standard errors. This is reasonable, since large curvature implies that the log likelihood drops quickly as $\boldsymbol{\beta}$ moves away from $\hat{\boldsymbol{\beta}}$; hence, the data would have been much more likely to occur if $\boldsymbol{\beta}$ took a value near $\hat{\boldsymbol{\beta}}$ rather than a value far from $\hat{\boldsymbol{\beta}}$.

1.3.2 Likelihood Function and ML Estimate for Binomial Parameter

The part of a likelihood function involving the parameters is called the *kernel*. Since the maximization of the likelihood is with respect to the parameters, the rest is irrelevant.

To illustrate, consider the binomial distribution (1.1). The binomial coefficient $\binom{n}{y}$ has no influence on where the maximum occurs with respect to π. Thus, we ignore it and treat the kernel as the likelihood function. The binomial log likelihood is then

$$L(\pi) = \log\left[\pi^y(1-\pi)^{n-y}\right] = y\log(\pi) + (n-y)\log(1-\pi). \tag{1.7}$$

Differentiating with respect to π yields

$$\partial L(\pi)/\partial\pi = y/\pi - (n-y)/(1-\pi) = (y - n\pi)/\pi(1-\pi). \tag{1.8}$$

Equating this to 0 gives the likelihood equation, which has solution $\hat{\pi} = y/n$, the sample proportion of successes for the n trials.

Calculating $\partial^2 L(\pi)/\partial\pi^2$, taking the expectation, and combining terms, we get

$$-E\left[\partial^2 L(\pi)/\partial\pi^2\right] = E\left[y/\pi^2 + (n-y)/(1-\pi)^2\right] = n/[\pi(1-\pi)]. \tag{1.9}$$

Thus, the asymptotic variance of $\hat{\pi}$ is $\pi(1-\pi)/n$. This is no surprise. Since $E(Y) = n\pi$ and $\text{var}(Y) = n\pi(1-\pi)$, the distribution of $\hat{\pi} = Y/n$ has mean and standard error

$$E(\hat{\pi}) = \pi, \qquad \sigma(\hat{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}}\,.$$

1.3.3 Wald–Likelihood Ratio–Score Test Triad

Three standard ways exist to use the likelihood function to perform large-sample inference. We introduce these for a significance test of a null hypothesis H_0: $\beta = \beta_0$ and then discuss their relation to interval estimation. They all exploit the large-sample normality of ML estimators.

With nonnull standard error SE of $\hat{\beta}$, the test statistic

$$z = \left(\hat{\beta} - \beta_0\right)/\text{SE}$$

has an approximate standard normal distribution when $\beta = \beta_0$. One refers z to the standard normal table to obtain one- or two-sided P-values. Equivalently, for the two-sided alternative, z^2 has a chi-squared null distribution with 1 degree of freedom (df); the P-value is then the right-tailed chi-squared probability above the observed value. This type of statistic, using the nonnull standard error, is called a *Wald statistic* (Wald 1943).

The multivariate extension for the Wald test of H_0: $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ has test statistic

$$W = \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)'\left[\text{cov}\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right).$$

(The prime on a vector or matrix denotes the transpose.) The nonnull covariance is based on the curvature (1.6) of the log likelihood at $\hat{\boldsymbol{\beta}}$. The asymptotic multivariate normal distribution for $\hat{\boldsymbol{\beta}}$ implies an asymptotic chi-squared distribution for W. The df equal the rank of $\text{cov}(\hat{\boldsymbol{\beta}})$, which is the number of nonredundant parameters in $\boldsymbol{\beta}$.

A second general-purpose method uses the likelihood function through the ratio of two maximizations: (1) the maximum over the possible parameter values under H_0, and (2) the maximum over the larger set of parameter values permitting H_0 or an alternative H_a to be true. Let ℓ_0 denote the maximized value of the likelihood function under H_0, and let ℓ_1 denote the maximized value generally (i.e., under $H_0 \cup H_a$). For instance, for parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)'$ and H_0: $\boldsymbol{\beta}_0 = \mathbf{0}$, ℓ_1 is the likelihood function calculated at the $\boldsymbol{\beta}$ value for which the data would have been most likely; ℓ_0 is the likelihood function calculated at the $\boldsymbol{\beta}_1$ value for which the data would have been most likely, when $\boldsymbol{\beta}_0 = \mathbf{0}$. Then ℓ_1 is always at least as large as ℓ_0, since ℓ_0 results from maximizing over a restricted set of the parameter values.

The ratio $\Lambda = \ell_0/\ell_1$ of the maximized likelihoods cannot exceed 1. Wilks (1935, 1938) showed that $-2\log\Lambda$ has a limiting null chi-squared distribution, as $n \to \infty$. The df equal the difference in the dimensions of the parameter spaces under $H_0 \cup H_a$ and under H_0. The *likelihood-ratio test statistic* equals

$$-2\log\Lambda = -2\log\left(\ell_0/\ell_1\right) = -2(L_0 - L_1),$$

where L_0 and L_1 denote the maximized log-likelihood functions.

The third method uses the *score statistic*, due to R. A. Fisher and C. R. Rao. The score test is based on the slope and expected curvature of the log-likelihood function $L(\beta)$ at the null value β_0. It utilizes the size of the *score function*

$$u(\beta) = \partial L(\beta)/\partial\beta,$$

evaluated at β_0. The value $u(\beta_0)$ tends to be larger in absolute value when $\hat{\beta}$ is farther from β_0. Denote $-E[\partial^2 L(\beta)/\partial\beta^2]$ (i.e., the information) evaluated at β_0 by $\iota(\beta_0)$. The score statistic is the ratio of $u(\beta_0)$ to its null SE, which is $[\iota(\beta_0)]^{1/2}$. This has an approximate standard normal null distribution. The chi-squared form of the score statistic is

$$\frac{\left[u(\beta_0)\right]^2}{\iota(\beta_0)} = \frac{\left[\partial L(\beta)/\partial\beta_0\right]^2}{-E\left[\partial^2 L(\beta)/\partial\beta_0^2\right]},$$

where the partial derivative notation reflects derivatives with respect to β that are evaluated at β_0. In the multiparameter case, the score statistic is a quadratic form based on the vector of partial derivatives of the log likelihood with respect to $\boldsymbol{\beta}$ and the inverse information matrix, both evaluated at the H_0 estimates (i.e., assuming that $\boldsymbol{\beta} = \boldsymbol{\beta}_0$).

Figure 1.1 is a generic plot of a log-likelihood $L(\beta)$ for the univariate case. It illustrates the three tests of H_0: $\beta = 0$. The Wald test uses the behavior of $L(\beta)$ at the ML estimate $\hat{\beta}$, having chi-squared form $(\hat{\beta}/\text{SE})^2$. The SE of $\hat{\beta}$ depends on the curvature of $L(\beta)$ at $\hat{\beta}$. The score test is based on the slope and curvature of $L(\beta)$ at $\beta = 0$. The likelihood-ratio test combines information about $L(\beta)$ at both $\hat{\beta}$ and $\beta_0 = 0$. It compares the log-likelihood values L_1 at $\hat{\beta}$ and L_0 at $\beta_0 = 0$ using the chi-squared statistic $-2(L_0 - L_1)$. In Figure 1.1, this statistic is twice the vertical distance between values of $L(\beta)$ at $\hat{\beta}$ and at 0. In a sense, this statistic uses the most information of the three types of test statistic and is the most versatile.

As $n \to \infty$, the Wald, likelihood-ratio, and score tests have certain asymptotic equivalences (Cox and Hinkley 1974, Sec. 9.3). For small to moderate sample sizes, the likelihood-ratio test is usually more reliable than the Wald test.

TRADING SOFTWARE

FOR SALE & EXCHANGE

www.trading-software-collection.com

***Subscribe** for FREE download more stuff.*

Mirrors:

www.forex-warez.com
www.traders-software.com

Contacts

andreybbrv@gmail.com
andreybbrv@hotmail.com
andreybbrv@yandex.ru
Skype: andreybbrv
ICQ: 70966433

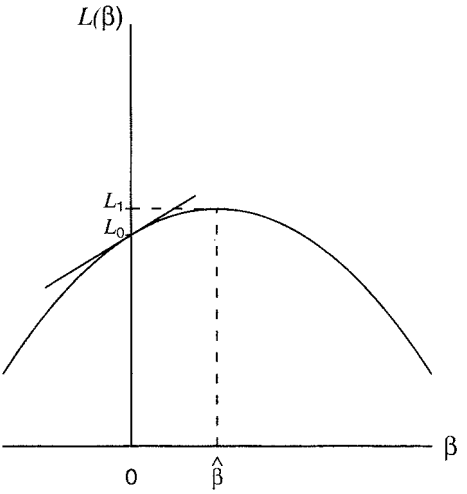


FIGURE 1.1 Log-likelihood function and information used in three tests of H_0: $\beta = 0$.

1.3.4 Constructing Confidence Intervals

In practice, it is more informative to construct confidence intervals for parameters than to test hypotheses about their values. For any of the three test methods, a confidence interval results from inverting the test. For instance, a 95% confidence interval for β is the set of β_0 for which the test of H_0: $\beta = \beta_0$ has a P-value exceeding 0.05.

Let z_a denote the z-score from the standard normal distribution having right-tailed probability a; this is the $100(1 - a)$ percentile of that distribution. Let $\chi^2_{df}(a)$ denote the $100(1 - a)$ percentile of the chi-squared distribution with degrees of freedom df. $100(1 - \alpha)\%$ confidence intervals based on asymptotic normality use $z_{\alpha/2}$, for instance $z_{0.025} = 1.96$ for 95% confidence. The Wald confidence interval is the set of β_0 for which $|\hat{\beta} - \beta_0|/\text{SE} < z_{\alpha/2}$. This gives the interval $\hat{\beta} \pm z_{\alpha/2}(\text{SE})$. The likelihood-ratio-based confidence interval is the set of β_0 for which $-2[L(\beta_0) - L(\hat{\beta})] < \chi^2_1(\alpha)$. [Recall that $\chi^2_1(\alpha) = z^2_{\alpha/2}$.]

When $\hat{\beta}$ has a normal distribution, the log-likelihood function has a parabolic shape (i.e., a second-degree polynomial). For small samples with categorical data, $\hat{\beta}$ may be far from normality and the log-likelihood function can be far from a symmetric, parabolic-shaped curve. This can also happen with moderate to large samples when a model contains many parameters. In such cases, inference based on asymptotic normality of $\hat{\beta}$ may have inadequate performance. A marked divergence in results of Wald and likelihood-ratio inference indicates that the distribution of $\hat{\beta}$ may not be close to normality. The example in Section 1.4.3 illustrates this with quite different confidence intervals for different methods. In many such cases, inference can

instead utilize an exact small-sample distribution or "higher-order" asymptotic methods that improve on simple normality (e.g., Pierce and Peters 1992).

The Wald confidence interval is most common in practice because it is simple to construct using ML estimates and standard errors reported by statistical software. The likelihood-ratio-based interval is becoming more widely available in software and is preferable for categorical data with small to moderate n. For the best known statistical model, regression for a normal response, the three types of inference necessarily provide identical results.

1.4 STATISTICAL INFERENCE FOR BINOMIAL PARAMETERS

In this section we illustrate inference methods for categorical data by presenting tests and confidence intervals for the binomial parameter π, based on y successes in n independent trials. In Section 1.3.2 we obtained the likelihood function and ML estimator $\hat{\pi} = y/n$ of π.

1.4.1 Tests about a Binomial Parameter

Consider H_0: $\pi = \pi_0$. Since H_0 has a single parameter, we use the normal rather than chi-squared forms of Wald and score test statistics. They permit tests against one-sided as well as two-sided alternatives. The Wald statistic is

$$z_W = \frac{\hat{\pi} - \pi_0}{\text{SE}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1 - \hat{\pi})/n}} . \tag{1.10}$$

Evaluating the binomial score (1.8) and information (1.9) at π_0 yields

$$u(\pi_0) = \frac{y}{\pi_0} - \frac{n - y}{1 - \pi_0}, \qquad \iota(\pi_0) = \frac{n}{\pi_0(1 - \pi_0)} .$$

The normal form of the score statistic simplifies to

$$z_S = \frac{u(\pi_0)}{\left[\iota(\pi_0)\right]^{1/2}} = \frac{y - n\pi_0}{\sqrt{n\pi_0(1 - \pi_0)}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}} . \tag{1.11}$$

Whereas the Wald statistic z_W uses the standard error evaluated at $\hat{\pi}$, the score statistic z_S uses it evaluated at π_0. The score statistic is preferable, as it uses the actual null SE rather than an estimate. Its null sampling distribution is closer to standard normal than that of the Wald statistic.

The binomial log-likelihood function (1.7) equals $L_0 = y\log\pi_0 + (n - y)\log(1 - \pi_0)$ under H_0 and $L_1 = y\log\hat{\pi} + (n - y)\log(1 - \hat{\pi})$ more

generally. The likelihood-ratio test statistic simplifies to

$$-2(L_0 - L_1) = 2\left(y \log\frac{\hat{\pi}}{\pi_0} + (n-y)\log\frac{1-\hat{\pi}}{1-\pi_0}\right).$$

Expressed as

$$-2(L_0 - L_1) = 2\left(y \log\frac{y}{n\pi_0} + (n-y)\log\frac{n-y}{n-n\pi_0}\right),$$

it compares observed success and failure counts to fitted (i.e., null) counts by

$$2\sum \text{observed} \log\frac{\text{observed}}{\text{fitted}}. \tag{1.12}$$

We'll see that this formula also holds for tests about Poisson and multinomial parameters. Since no unknown parameters occur under H_0 and one occurs under H_a, (1.12) has an asymptotic chi-squared distribution with df $= 1$.

1.4.2 Confidence Intervals for a Binomial Parameter

A significance test merely indicates whether a particular π value (such as $\pi = 0.5$) is plausible. We learn more by using a confidence interval to determine the range of plausible values.

Inverting the Wald test statistic gives the interval of π_0 values for which $|z_W| < z_{\alpha/2}$, or

$$\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}. \tag{1.13}$$

Historically, this was one of the first confidence intervals used for any parameter (Laplace 1812, p. 283). Unfortunately, it performs poorly unless n is very large (e.g., Brown et al. 2001). The actual coverage probability usually falls below the nominal confidence coefficient, much below when π is near 0 or 1. A simple adjustment that adds $\frac{1}{2}z^2_{\alpha/2}$ observations of each type to the sample before using this formula performs much better (Problem 1.24).

The score confidence interval contains π_0 values for which $|z_S| < z_{\alpha/2}$. Its endpoints are the π_0 solutions to the equations

$$(\hat{\pi} - \pi_0)/\sqrt{\pi_0(1-\pi_0)/n} = \pm z_{\alpha/2}.$$

These are quadratic in π_0. First discussed by E. B. Wilson (1927), this interval is

$$\hat{\pi}\left(\frac{n}{n+z_{\alpha/2}^2}\right)+\frac{1}{2}\left(\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}\right) \pm z_{\alpha/2}\sqrt{\frac{1}{n+z_{\alpha/2}^2}\left[\hat{\pi}(1-\hat{\pi})\left(\frac{n}{n+z_{\alpha/2}^2}\right)+\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{z_{\alpha/2}^2}{n+z_{\alpha/2}^2}\right)\right]}\,.$$

The midpoint $\tilde{\pi}$ of the interval is a weighted average of $\hat{\pi}$ and $\frac{1}{2}$, where the weight $n/(n+z_{\alpha/2}^2)$ given $\hat{\pi}$ increases as n increases. Combining terms, this midpoint equals $\tilde{\pi}=(y+z_{\alpha/2}^2/2)/(n+z_{\alpha/2}^2)$. This is the sample proportion for an adjusted sample that adds $z_{\alpha/2}^2$ observations, half of each type. The square of the coefficient of $z_{\alpha/2}$ in this formula is a weighted average of the variance of a sample proportion when $\pi=\hat{\pi}$ and the variance of a sample proportion when $\pi=\frac{1}{2}$, using the adjusted sample size $n+z_{\alpha/2}^2$ in place of n. This interval has much better performance than the Wald interval.

The likelihood-ratio-based confidence interval is more complex computationally, but simple in principle. It is the set of π_0 for which the likelihood-ratio test has a P-value exceeding α. Equivalently, it is the set of π_0 for which double the log likelihood drops by less than $\chi_1^2(\alpha)$ from its value at the ML estimate $\hat{\pi}=y/n$.

1.4.3 Proportion of Vegetarians Example

To collect data in an introductory statistics course, recently I gave the students a questionnaire. One question asked each student whether he or she was a vegetarian. Of $n=25$ students, $y=0$ answered "yes." They were not a random sample of a particular population, but we use these data to illustrate 95% confidence intervals for a binomial parameter π.

Since $y=0$, $\hat{\pi}=0/25=0$. Using the Wald approach, the 95% confidence interval for π is

$$0\pm 1.96\sqrt{(0.0\times 1.0)/25}\,,\quad \text{or}\quad (0,0)\,.$$

When the observation falls at the boundary of the sample space, often Wald methods do not provide sensible answers.

By contrast, the 95% score interval equals (0.0, 0.133). This is a more believable inference. For H_0: $\pi=0.5$, for instance, the score test statistic is $z_S=(0-0.5)/\sqrt{(0.5\times 0.5)/25}=-5.0$, so 0.5 does not fall in the interval. By contrast, for H_0: $\pi=0.10$, $z_S=(0-0.10)/\sqrt{(0.10\times 0.90)/25}=-1.67$, so 0.10 falls in the interval.

When $y = 0$ and $n = 25$, the kernel of the likelihood function is $\ell(\pi) = \pi^0(1 - \pi)^{25} = (1 - \pi)^{25}$. The log likelihood (1.7) is $L(\pi) = 25\log(1 - \pi)$. Note that $L(\hat{\pi}) = L(0) = 0$. The 95% likelihood-ratio confidence interval is the set of π_0 for which the likelihood-ratio statistic

$$-2(L_0 - L_1) = -2[L(\pi_0) - L(\hat{\pi})]$$
$$= -50\log(1 - \pi_0) \leq \chi_1^2(0.05) = 3.84.$$

The upper bound is $1 - \exp(-3.84/50) = 0.074$, and the confidence interval equals (0.0, 0.074). [In this book, we use the natural logarithm throughout, so its inverse is the exponential function $\exp(x) = e^x$.] Figure 1.2 shows the likelihood and log-likelihood functions and the corresponding confidence region for π.

The three large-sample methods yield quite different results. When π is near 0, the sampling distribution of $\hat{\pi}$ is highly skewed to the right for small n. It is worth considering alternative methods not requiring asymptotic approximations.

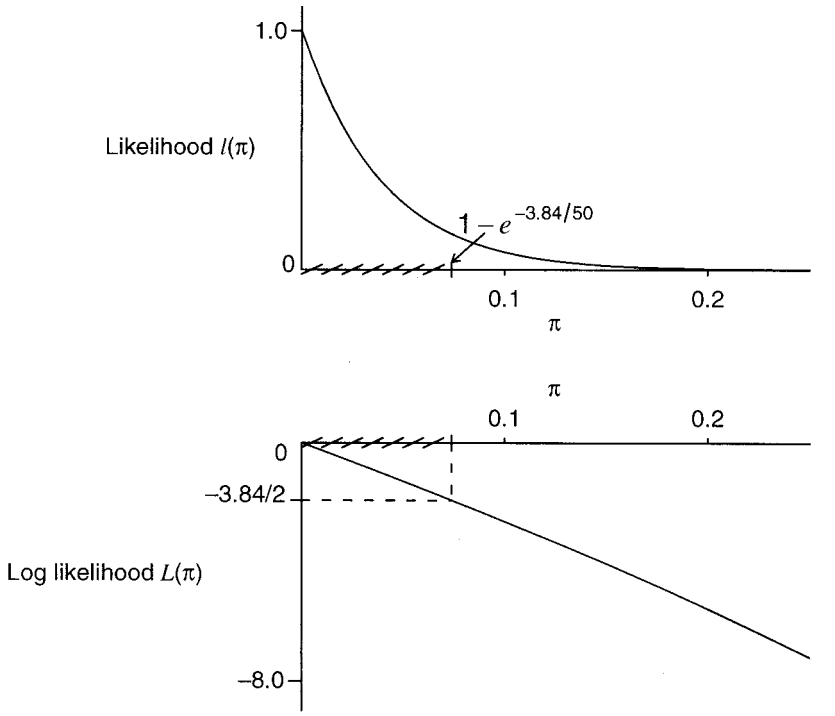


FIGURE 1.2 Binomial likelihood and log likelihood when $y = 0$ in $n = 25$ trials, and confidence interval for π.

1.4.4 Exact Small-Sample Inference*[1]

With modern computational power, it is not necessary to rely on large-sample approximations for the distribution of statistics such as $\hat{\pi}$. Tests and confidence intervals can use the binomial distribution directly rather than its normal approximation. Such inferences occur naturally for small samples, but apply for any n.

We illustrate by testing H_0: $\pi = 0.5$ against H_a: $\pi \neq 0.5$ for the survey results on vegetarianism, $y = 0$ with $n = 25$. We noted that the score statistic equals $z = -5.0$. The exact P-value for this statistic, based on the null bin(25, 0.5) distribution, is

$$P(|z| \geq 5.0) = P(Y = 0 \text{ or } Y = 25) = 0.5^{25} + 0.5^{25} = 0.00000006.$$

$100(1 - \alpha)\%$ confidence intervals consist of all π_0 for which P-values exceed α in exact binomial tests. The best known interval (Clopper and Pearson 1934) uses the *tail method* for forming confidence intervals. It requires each one-sided P-value to exceed $\alpha/2$. The lower and upper endpoints are the solutions in π_0 to the equations

$$\sum_{k=y}^{n} \binom{n}{k} \pi_0^k (1 - \pi_0)^{n-k} = \alpha/2 \quad \text{and} \quad \sum_{k=0}^{y} \binom{n}{k} \pi_0^k (1 - \pi_0)^{n-k} = \alpha/2,$$

except that the lower bound is 0 when $y = 0$ and the upper bound is 1 when $y = n$. When $y = 1, 2, \ldots, n - 1$, from connections between binomial sums and the incomplete beta function and related cumulative distribution functions (cdf's) of beta and F distributions, the confidence interval equals

$$\left[1 + \frac{n - y + 1}{y F_{2y, 2(n-y+1)}(1 - \alpha/2)}\right]^{-1} < \pi < \left[1 + \frac{n - y}{(y + 1) F_{2(y+1), 2(n-y)}(\alpha/2)}\right]^{-1},$$

where $F_{a,b}(c)$ denotes the $1 - c$ quantile from the F distribution with degrees of freedom a and b. When $y = 0$ with $n = 25$, the Clopper–Pearson 95% confidence interval for π is (0.0, 0.137).

In principle this approach seems ideal. However, there is a serious complication. Because of discreteness, the actual coverage probability for any π is at least as large as the nominal confidence level (Casella and Berger 2001, p. 434; Neyman 1935) and it can be much greater. Similarly, for a test of H_0: $\pi = \pi_0$ at a fixed desired size α such as 0.05, it is not usually possible to achieve that size. There is a finite number of possible samples, and hence a finite number of possible P-values, of which 0.05 may not be one. In testing H_0 with fixed π_0, one can pick a particular α that can occur as a P-value.

[1]Sections marked with an asterisk are less important for an overview.

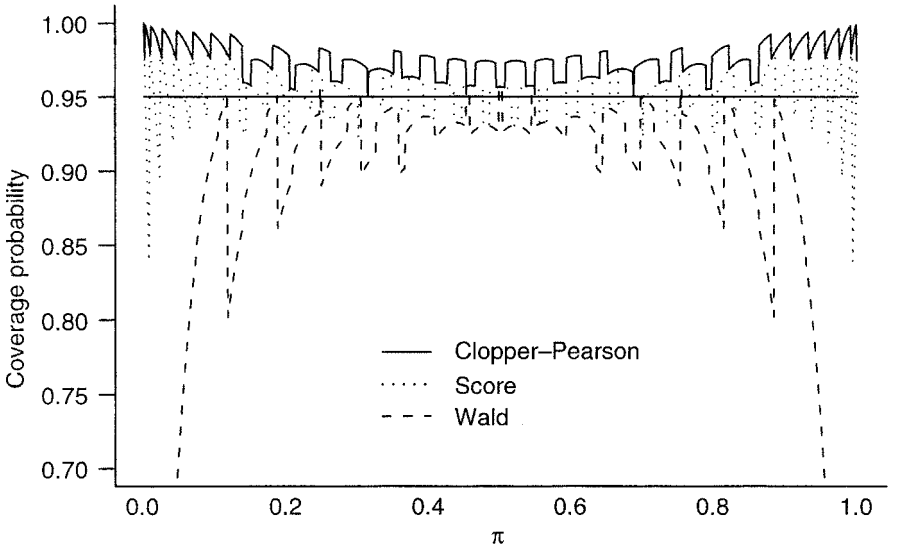


FIGURE 1.3 Plot of coverage probabilities for nominal 95% confidence intervals for binomial parameter π when $n = 25$.

For interval estimation, however, this is not an option. This is because constructing the interval corresponds to inverting an entire range of π_0 values in H_0: $\pi = \pi_0$, and each distinct π_0 value can have its own set of possible P-values; that is, there is not a single null parameter value π_0 as in one test.

For any fixed parameter value, the actual coverage probability can be much larger than the nominal confidence level. When $n = 25$, Figure 1.3 plots the coverage probabilities as a function of π for the Clopper–Pearson method, the score method, and the Wald method. At a fixed π value with a given method, the coverage probability is the sum of the binomial probabilities of all those samples for which the resulting interval contains that π. There are 26 possible samples and 26 corresponding confidence intervals, so the coverage probability is a sum of somewhere between 0 and 26 binomial probabilities. As π moves from 0 to 1, this coverage probability jumps up or down whenever π moves into or out of one of these intervals. Figure 1.3 shows that coverage probabilities are too low for the Wald method, whereas the Clopper–Pearson method errs in the opposite direction. The score method behaves well, except for some π values close to 0 or 1. Its coverage probabilities tend to be near the nominal level, not being consistently conservative or liberal. This is a good method unless π is very close to 0 or 1 (Problem 1.23).

In discrete problems using small-sample distributions, shorter confidence intervals usually result from inverting a single two-sided test rather than two

one-sided tests. The interval is then the set of parameter values for which the P-value of a two-sided test exceeds α. For the binomial parameter, see Blaker (2000), Blyth and Still (1983), and Sterne (1954) for methods. For observed outcome y_o, with Blaker's approach the P-value is the minimum of the two one-tailed binomial probabilities $P(Y \geq y_o)$ and $P(Y \leq y_o)$ plus an attainable probability in the other tail that is as close as possible to, but not greater than, that one-tailed probability. The interval is computationally more complex, although available in software (Blaker gave S-Plus functions). The result is still conservative, but less so than the Clopper–Pearson interval. For the vegetarianism example, the 95% confidence interval using the Blaker exact method is (0.0, 0.128) compared to the Clopper–Pearson interval of (0.0, 0.137).

1.4.5 Inference Based on the Mid-*P*-Value*

To adjust for discreteness in small-sample distributions, one can base inference on the *mid-P-value* (Lancaster 1961). For a test statistic T with observed value t_o and one-sided H_a such that large T contradicts H_0,

$$\text{mid-}P\text{-value} = \tfrac{1}{2}P(T = t_o) + P(T > t_o) ,$$

with probabilities calculated from the null distribution. Thus, the mid-P-value is less than the ordinary P-value by half the probability of the observed result. Compared to the ordinary P-value, the mid-P-value behaves more like the P-value for a test statistic having a continuous distribution. The sum of its two one-sided P-values equals 1.0. Although discrete, under H_0 its null distribution is more like the uniform distribution that occurs in the continuous case. For instance, it has a null expected value of 0.5, whereas this expected value exceeds 0.5 for the ordinary P-value for a discrete test statistic.

Unlike an exact test with ordinary P-value, a test using the mid-P-value does not guarantee that the probability of type I error is no greater than a nominal value (Problem 1.19). However, it usually performs well, typically being a bit conservative. It is less conservative than the ordinary exact test. Similarly, one can form less conservative confidence intervals by inverting tests using the exact distribution with the mid-P-value (e.g., the 95% confidence interval is the set of parameter values for which the mid-P-value exceeds 0.05).

For testing H_0: $\pi = 0.5$ against H_a: $\pi \neq 0.5$ in the example about the proportion of vegetarians, with $y = 0$ for $n = 25$, the result observed is the most extreme possible. Thus the mid-P-value is half the ordinary P-value, or 0.00000003. Using the Clopper–Pearson inversion of the exact binomial test but with the mid-P-value yields a 95% confidence interval of (0.000, 0.113) for π, compared to (0.000, 0.137) for the ordinary Clopper–Pearson interval.

The mid-P-value seems a sensible compromise between having overly conservative inference and using irrelevant randomization to eliminate prob-

lems from discreteness. We recommend it both for tests and confidence intervals with highly discrete distributions.

1.5 STATISTICAL INFERENCE FOR MULTINOMIAL PARAMETERS

We now present inference for multinomial parameters $\{\pi_j\}$. Of n observations, n_j occur in category j, $j = 1, \ldots, c$.

1.5.1 Estimation of Multinomial Parameters

First, we obtain ML estimates of $\{\pi_j\}$. As a function of $\{\pi_j\}$, the multinomial probability mass function (1.2) is proportional to the kernel

$$\prod_j \pi_j^{n_j} \quad \text{where} \quad \text{all } \pi_j \geq 0 \quad \text{and} \quad \sum_j \pi_j = 1. \tag{1.14}$$

The ML estimates are the $\{\pi_j\}$ that maximize (1.14).

The multinomial log-likelihood function is

$$L(\boldsymbol{\pi}) = \sum_j n_j \log \pi_j.$$

To eliminate redundancies, we treat L as a function of $(\pi_1, \ldots, \pi_{c-1})$, since $\pi_c = 1 - (\pi_1 + \cdots + \pi_{c-1})$. Thus, $\partial \pi_c / \partial \pi_j = -1$, $j = 1, \ldots, c-1$.

Since

$$\frac{\partial \log \pi_c}{\partial \pi_j} = \frac{1}{\pi_c} \frac{\partial \pi_c}{\partial \pi_j} = -\frac{1}{\pi_c},$$

differentiating $L(\boldsymbol{\pi})$ with respect to π_j gives the likelihood equation

$$\frac{\partial L(\boldsymbol{\pi})}{\partial \pi_j} = \frac{n_j}{\pi_j} - \frac{n_c}{\pi_c} = 0\,.$$

The ML solution satisfies $\hat{\pi}_j / \hat{\pi}_c = n_j / n_c$. Now

$$\sum_j \hat{\pi}_j = 1 = \frac{\hat{\pi}_c \left(\sum_j n_j \right)}{n_c} = \frac{\hat{\pi}_c n}{n_c},$$

so $\hat{\pi}_c = n_c / n$ and then $\hat{\pi}_j = n_j / n$. From general results presented later in the book (Section 8.6), this solution does maximize the likelihood. Thus, the ML estimates of $\{\pi_j\}$ are the sample proportions.

1.5.2 Pearson Statistic for Testing a Specified Multinomial

In 1900 the eminent British statistician Karl Pearson introduced a hypothesis test that was one of the first inferential methods. It had a revolutionary impact on categorical data analysis, which had focused on describing associations. Pearson's test evaluates whether multinomial parameters equal certain specified values. His original motivation in developing this test was to analyze whether possible outcomes on a particular Monte Carlo roulette wheel were equally likely (Stigler 1986).

Consider H_0: $\pi_j = \pi_{j0}$, $j = 1, \ldots, c$, where $\sum_j \pi_{j0} = 1$. When H_0 is true, the expected values of $\{n_j\}$, called *expected frequencies*, are $\mu_j = n\pi_{j0}$, $j = 1, \ldots, c$. Pearson proposed the test statistic

$$X^2 = \sum_j \frac{(n_j - \mu_j)^2}{\mu_j}. \tag{1.15}$$

Greater differences $\{n_j - \mu_j\}$ produce greater X^2 values, for fixed n. Let X_o^2 denote the observed value of X^2. The P-value is the null value of $P(X^2 \geq X_o^2)$. This equals the sum of the null multinomial probabilities of all count arrays (having a sum of n) with $X^2 \geq X_o^2$.

For large samples, X^2 has approximately a chi-squared distribution with df $= c - 1$. The P-value is approximated by $P(\chi^2_{c-1} \geq X_o^2)$, where χ^2_{c-1} denotes a chi-squared random variable with df $= c - 1$. Statistic (1.15) is called the *Pearson chi-squared statistic*.

1.5.3 Example: Testing Mendel's Theories

Among its many applications, Pearson's test was used in genetics to test Mendel's theories of natural inheritance. Mendel crossed pea plants of pure yellow strain with plants of pure green strain. He predicted that second-generation hybrid seeds would be 75% yellow and 25% green, yellow being the dominant strain. One experiment produced $n = 8023$ seeds, of which $n_1 = 6022$ were yellow and $n_2 = 2001$ were green. The expected frequencies for H_0: $\pi_{10} = 0.75$, $\pi_{20} = 0.25$ are $\mu_1 = 8023(0.75) = 6017.25$ and $\mu_2 = 2005.75$. The Pearson statistic $X^2 = 0.015$ (df = 1) has a P-value of $P = 0.90$. This does not contradict Mendel's hypothesis.

Mendel performed several experiments of this type. In 1936, R. A. Fisher summarized Mendel's results. He used the reproductive property of chi-squared: If $X_1^2, \ldots, X_k^2$ are independent chi-squared statistics with degrees of freedom $\nu_1, \ldots, \nu_k$, then $\sum_i X_i^2$ has a chi-squared distribution with df $= \sum_i \nu_i$. Fisher obtained a summary chi-squared statistic equal to 42, with df = 84. A chi-squared distribution with df = 84 has mean 84 and standard deviation $(2 \times 84)^{1/2} = 13.0$, and the right-tailed probability above 42 is $P = 0.99996$. In other words, the chi-squared statistic was so small that the fit seemed *too* good.

Fisher commented: "The general level of agreement between Mendel's expectations and his reported results shows that it is closer than would be expected in the best of several thousand repetitions . . . . I have no doubt that Mendel was deceived by a gardening assistant, who knew only too well what his principal expected from each trial made." In a letter written at the time (see Box 1978, p. 297), he stated: "Now, when data have been faked, I know very well how generally people underestimate the frequency of wide chance deviations, so that the tendency is always to make them agree too well with expectations." In summary, goodness-of-fit tests can reveal not only when a fit is inadequate, but also when it is better than random fluctuations would have us expect. [R. A. Fisher's daughter, Joan Fisher Box (1978, pp. 295–300), and Freedman et al. (1978, pp. 420–428, 478) discussed Fisher's analysis of Mendel's data and the accompanying controversy. Despite possible difficulties with Mendel's data, subsequent work led to general acceptance of his theories.]

1.5.4 Chi-Squared Theoretical Justification*

We now outline why Pearson's statistic has a limiting chi-squared distribution. For a multinomial sample $(n_1, \ldots, n_c)$ of size n, the marginal distribution of n_j is the bin(n,π_j) distribution. For large n, by the normal approximation to the binomial, n_j (and $\hat{\pi}_j = n_j/n$) have approximate normal distributions. More generally, by the central limit theorem, the sample proportions $\hat{\boldsymbol{\pi}} = (n_1/n, \ldots, n_{c-1}/n)'$ have an approximate multivariate normal distribution (Section 14.1.4). Let $\boldsymbol{\Sigma}_0$ denote the null covariance matrix of $\sqrt{n}\,\hat{\boldsymbol{\pi}}$, and let $\boldsymbol{\pi}_0 = (\pi_{10}, \ldots, \pi_{c-1,0})'$. Under H_0, since $\sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)$ converges to a $N(\mathbf{0}, \boldsymbol{\Sigma}_0)$ distribution, the quadratic form

$$n(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0)'\boldsymbol{\Sigma}_0^{-1}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0) \tag{1.16}$$

has distribution converging to chi-squared with df $= c - 1$.

In Section 14.1.4 we show that the covariance matrix of $\sqrt{n}\,\hat{\boldsymbol{\pi}}$ has elements

$$\sigma_{jk} = \begin{cases} -\pi_j\pi_k & \text{if } j \neq k \\ \pi_j(1 - \pi_j) & \text{if } j = k \end{cases}.$$

The matrix $\boldsymbol{\Sigma}_0^{-1}$ has (j,k)th element $1/\pi_{c0}$ when $j \neq k$ and $(1/\pi_{j0} + 1/\pi_{c0})$ when $j = k$. (You can verify this by showing that $\boldsymbol{\Sigma}_0\,\boldsymbol{\Sigma}_0^{-1}$ equals the identity matrix.) With this substitution, direct calculation (with appropriate combining of terms) shows that (1.16) simplifies to X^2. In Section 14.3 we provide a formal proof in a more general setting.

This argument is similar to Pearson's in 1900. R. A. Fisher (1922) gave a simpler justification, the gist of which follows: Suppose that $(n_1, \ldots, n_c)$ are independent Poisson random variables with means $(\mu_1, \ldots, \mu_c)$. For large

$\{\mu_j\}$, the standardized values $\{z_j = (n_j - \mu_j)/\sqrt{\mu_j}\}$ have approximate standard normal distributions. Thus, $\sum_j z_j^2 = X^2$ has an approximate chi-squared distribution with c degrees of freedom. Adding the single linear constraint $\sum_j (n_j - \mu_j) = 0$, thus converting the Poisson distributions to a multinomial, we lose a degree of freedom.

When $c = 2$, Pearson's X^2 simplifies to the square of the normal score statistic (1.11). For Mendel's data, $\hat{\pi}_1 = 6022/8023$, $\pi_{10} = 0.75$, $n = 8023$, and $z_S = 0.123$, for which $X^2 = (0.123)^2 = 0.015$. In fact, for general c the Pearson test is the score test about multinomial parameters.

1.5.5 Likelihood-Ratio Chi-Squared

An alternative test for multinomial parameters uses the likelihood-ratio test. The kernel of the multinomial likelihood is (1.14). Under H_0 the likelihood is maximized when $\hat{\pi}_j = \pi_{j0}$. In the general case, it is maximized when $\hat{\pi}_j = n_j/n$. The ratio of the likelihoods equals

$$\Lambda = \frac{\prod_j (\pi_{j0})^{n_j}}{\prod_j (n_j/n)^{n_j}}.$$

Thus, the likelihood-ratio statistic, denoted by G^2, is

$$G^2 = -2\log\Lambda = 2\sum n_j \log(n_j/n\pi_{j0}). \tag{1.17}$$

This statistic, which has form (1.12), is called the *likelihood-ratio chi-squared statistic*. The larger the value of G^2, the greater the evidence against H_0.

In the general case, the parameter space consists of $\{\pi_j\}$ subject to $\sum_j \pi_j = 1$, so the dimensionality is $c - 1$. Under H_0, the $\{\pi_j\}$ are specified completely, so the dimension is 0. The difference in these dimensions equals $(c - 1)$. For large n, G^2 has a chi-squared null distribution with df $= c - 1$.

When H_0 holds, the Pearson X^2 and the likelihood ratio G^2 both have asymptotic chi-squared distributions with df $= c - 1$. In fact, they are asymptotically equivalent in that case; specifically, $X^2 - G^2$ converges in probability to zero (Section 14.3.4). When H_0 is false, they tend to grow proportionally to n; they need not take similar values, however, even for very large n.

For fixed c, as n increases the distribution of X^2 usually converges to chi-squared more quickly than that of G^2. The chi-squared approximation is usually poor for G^2 when $n/c < 5$. When c is large, it can be decent for X^2 for n/c as small as 1 if the table does not contain both very small and moderately large expected frequencies. We provide further guidelines in Section 9.8.4. Alternatively, one can use the multinomial probabilities to generate exact distributions of these test statistics (Good et al. 1970).

1.5.6 Testing with Estimated Expected Frequencies

Pearson's X^2 (1.15) compares a sample distribution to a hypothetical one $\{\pi_{j0}\}$. In some applications, $\{\pi_{j0} = \pi_{j0}(\boldsymbol{\theta})\}$ are functions of a smaller set of unknown parameters $\boldsymbol{\theta}$. ML estimates $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ determine ML estimates $\{\pi_{j0}(\hat{\boldsymbol{\theta}})\}$ of $\{\pi_{j0}\}$ and hence ML estimates $\{\hat{\mu}_j = n\pi_{j0}(\hat{\boldsymbol{\theta}})\}$ of expected frequencies in X^2. Replacing $\{\mu_j\}$ by estimates $\{\hat{\mu}_j\}$ affects the distribution of X^2. When $\dim(\boldsymbol{\theta}) = p$, the true df $= (c - 1) - p$ (Section 14.3.3). Pearson failed to realize this (Section 16.2).

We now show a goodness-to-fit test with estimated expected frequencies. A sample of 156 dairy calves born in Okeechobee County, Florida, were classified according to whether they caught pneumonia within 60 days of birth. Calves that got a pneumonia infection were also classified according to whether they got a secondary infection within 2 weeks after the first infection cleared up. Table 1.1 shows the data. Calves that did not get a primary infection could not get a secondary infection, so no observations can fall in the category for "no" primary infection and "yes" secondary infection. That combination is called a *structural zero*.

A goal of this study was to test whether the probability of primary infection was the same as the conditional probability of secondary infection, given that the calf got the primary infection. In other words, if π_{ab} denotes the probability that a calf is classified in row a and column b of this table, the null hypothesis is

$$H_0\colon \pi_{11} + \pi_{12} = \pi_{11}/(\pi_{11} + \pi_{12})$$

or $\pi_{11} = (\pi_{11} + \pi_{12})^2$. Let $\pi = \pi_{11} + \pi_{12}$ denote the probability of primary infection. The null hypothesis states that the probabilities satisfy the structure that Table 1.2 shows; that is, probabilities in a trinomial for the categories (yes–yes, yes–no, no–no) for primary–secondary infection equal $(\pi^2, \pi(1 - \pi), 1 - \pi)$.

Let n_{ab} denote the number of observations in category (a, b). The ML estimate of π is the value maximizing the kernel of the multinomial likelihood

$$(\pi^2)^{n_{11}}(\pi - \pi^2)^{n_{12}}(1 - \pi)^{n_{22}}.$$

TABLE 1.1 Primary and Secondary Pneumonia Infections in Calves

	Secondary Infection[a]	
Primary Infection	Yes	No
Yes	30 (38.1)	63 (39.0)
No	0 (—)	63 (78.9)

Source: Data courtesy of Thang Tran and G. A. Donovan, College of Veterinary Medicine, University of Florida.

[a] Values in parentheses are estimated expected frequencies.

TABLE 1.2 Probability Structure for Hypothesis

Primary Infection	Secondary Infection		
	Yes	No	Total
Yes	π^2	$\pi(1-\pi)$	π
No	—	$1-\pi$	$1-\pi$

The log likelihood is

$$L(\pi) = n_{11}\log\pi^2 + n_{12}\log(\pi - \pi^2) + n_{22}\log(1-\pi).$$

Differentiation with respect to π gives the likelihood equation

$$\frac{2n_{11}}{\pi} + \frac{n_{12}}{\pi} - \frac{n_{12}}{1-\pi} - \frac{n_{22}}{1-\pi} = 0.$$

The solution is

$$\hat{\pi} = (2n_{11} + n_{12})/(2n_{11} + 2n_{12} + n_{22}).$$

For Table 1.1, $\hat{\pi} = 0.494$. Since $n = 156$, the estimated expected frequencies are $\hat{\mu}_{11} = n\hat{\pi}^2 = 38.1$, $\hat{\mu}_{12} = n(\hat{\pi} - \hat{\pi}^2) = 39.0$, and $\hat{\mu}_{22} = n(1-\hat{\pi}) = 78.9$. Table 1.1 shows them. Pearson's statistic is $X^2 = 19.7$. Since the $c = 3$ possible responses have $p = 1$ parameter (π) determining the expected frequencies, df $= (3-1) - 1 = 1$. There is strong evidence against H_0 ($P = 0.00001$). Inspection of Table 1.1 reveals that many more calves got a primary infection but not a secondary infection than H_0 predicts. The researchers concluded that the primary infection had an immunizing effect that reduced the likelihood of a secondary infection.

NOTES

Section 1.1: Categorical Response Data

1.1. Stevens (1951) defined (nominal, ordinal, interval) scales of measurement. Other scales result from mixtures of these types. For instance, *partially ordered* scales occur when subjects respond to questions having categories ordered except for don't know or undecided categories.

Section 1.3: Statistical Inference for Categorical Data

1.2. The score method does not use $\hat{\beta}$. Thus, when β is a model parameter, one can usually compute the score statistic for testing H_0: $\beta = \beta_0$ without fitting the model. This is advantageous when fitting several models in an exploratory analysis and model fitting is computationally intensive. An advantage of the score and likelihood-ratio methods is that

they apply even when $|\hat{\beta}| = \infty$. In that case, one cannot compute the Wald statistic. Another disadvantage of the Wald method is that its results depend on the parameterization; inference based on $\hat{\beta}$ and its SE is not equivalent to inference based on a nonlinear function of it, such as $\log \hat{\beta}$ and its SE.

Section 1.4: Statistical Inference for Binomial Parameters

1.3. Among others, Agresti and Coull (1998), Blyth and Still (1983), Brown et al. (2001), Ghosh (1979), and Newcombe (1998a) showed the superiority of the score interval to the Wald interval for π. Of the "exact" methods, Blaker's (2000) has particularly good properties. It is contained in the Clopper–Pearson interval and has a nestedness property whereby an interval of higher nominal confidence level necessarily contains one of lower level.

1.4. Using continuity corrections with large-sample methods provides approximations to exact small-sample methods. Thus, they tend to behave conservatively. We do not present them, since if one prefers an exact method, with modern computational power it can be used directly rather than approximated.

1.5. In theory, one can eliminate problems with discreteness in tests by performing a supplementary randomization on the boundary of a critical region (see Problem 1.19). In rejecting the null at the boundary with a certain probability, one can obtain a fixed overall type I error probability α even when it is not an achievable P-value. For such randomization, the one-sided $P-$ value is

$$\text{randomized } P\text{-value} = U \times P(T = t_o) + P(T > t_o),$$

where U denotes a uniform $(0, 1)$ random variable (Stevens 1950). In practice, this is not used, as it is absurd to let this random number influence a decision. The mid P-value replaces the arbitrary uniform multiple $U \times P(T = t_o)$ by its expected value.

Section 1.5: Statistical Inference for Multinomial Parameters

1.6. The chi-squared distribution has mean df, variance 2 df, and skewness $(8/\text{df})^{1/2}$. It is approximately normal when df is large. Greenwood and Nikulin (1996), Kendall and Stuart (1979), and Lancaster (1969) presented other properties. Cochran (1952) presented a historical survey of chi-squared tests of fit. See also Cressie and Read (1989), Koch and Bhapkar (1982), Koehler (1998), and Moore (1986b).

PROBLEMS

Applications

1.1 Identify each variable as nominal, ordinal, or interval.

- **a.** UK political party preference (Labour, Conservative, Social Democrat)
- **b.** Anxiety rating (none, mild, moderate, severe, very severe)
- **c.** Patient survival (in number of months)
- **d.** Clinic location (London, Boston, Madison, Rochester, Montreal)

e. Response of tumor to chemotherapy (complete elimination, partial reduction, stable, growth progression)

f. Favorite beverage (water, juice, milk, soft drink, beer, wine)

g. Appraisal of company's inventory level (too low, about right, too high)

1.2 Each of 100 multiple-choice questions on an exam has four possible answers, one of which is correct. For each question, a student guesses by selecting an answer randomly.

a. Specify the distribution of the student′s number of correct answers.

b. Find the mean and standard deviation of that distribution. Would it be surprising if the student made at least 50 correct responses? Why?

c. Specify the distribution of (n_1, n_2, n_3, n_4), where n_j is the number of times the student picked choice j.

d. Find $E(n_j)$, $\text{var}(n_j)$, $\text{cov}(n_j, n_k)$, and $\text{corr}(n_j, n_k)$.

1.3 An experiment studies the number of insects that survive a certain dose of an insecticide, using several batches of insects of size n each. The insects are sensitive to factors that vary among batches during the experiment but were not measured, such as temperature level. Explain why the distribution of the number of insects per batch surviving the experiment might show overdispersion relative to a $\text{bin}(n, \pi)$ distribution.

1.4 In his autobiography *A Sort of Life*, British author Graham Greene described a period of severe mental depression during which he played Russian Roulette. This "game" consists of putting a bullet in one of the six chambers of a pistol, spinning the chambers to select one at random, and then firing the pistol once at one's head.

a. Greene played this game six times and was lucky that none of them resulted in a bullet firing. Find the probability of this outcome.

b. Suppose that he had kept playing this game until the bullet fired. Let Y denote the number of the game on which it fires. Show the probability mass function for Y, and justify.

1.5 Consider the statement, "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if she is married and does not want any more children." For the 1996 General Social Survey, conducted by the National Opinion Research Center (NORC), 842 replied "yes" and 982 replied "no." Let π denote

the population proportion who would reply "yes." Find the P-value for testing H_0: $\pi = 0.5$ using the score test, and construct a 95% confidence interval for π. Interpret the results.

1.6 Refer to the vegetarianism example in Section 1.4.3. For testing H_0: $\pi = 0.5$ against H_a: $\pi \neq 0.5$, show that:

a. The likelihood-ratio statistic equals $2[25\log(25/12.5)] = 34.7$.

b. The chi-squared form of the score statistic equals 25.0.

c. The Wald z or chi-squared statistic is infinite.

1.7 In a crossover trial comparing a new drug to a standard, π denotes the probability that the new one is judged better. It is desired to estimate π and test H_0: $\pi = 0.5$ against H_a: $\pi \neq 0.5$. In 20 independent observations, the new drug is better each time.

a. Find and sketch the likelihood function. Give the ML estimate of π.

b. Conduct a Wald test and construct a 95% Wald confidence interval for π. Are these sensible?

c. Conduct a score test, reporting the P-value. Construct a 95% score confidence interval. Interpret.

d. Conduct a likelihood-ratio test and construct a likelihood-based 95% confidence interval. Interpret.

e. Construct an exact binomial test and 95% confidence interval. Interpret.

f. Suppose that researchers wanted a sufficiently large sample to estimate the probability of preferring the new drug to within 0.05, with confidence 0.95. If the true probability is 0.90, about how large a sample is needed?

1.8 In an experiment on chlorophyll inheritance in maize, for 1103 seedlings of self-fertilized heterozygous green plants, 854 seedlings were green and 249 were yellow. Theory predicts the ratio of green to yellow is 3:1. Test the hypothesis that 3:1 is the true ratio. Report the P-value, and interpret.

1.9 Table 1.3 contains Ladislaus von Bortkiewicz's data on deaths of soldiers in the Prussian army from kicks by army mules (Fisher 1934; Quine and Seneta 1987). The data refer to 10 army corps, each observed for 20 years. In 109 corps-years of exposure, there were no deaths, in 65 corps-years there was one death, and so on. Estimate the mean and test whether probabilities of occurrences in these five categories follow a Poisson distribution (truncated for 4 and above.)

TABLE 1.3 Data for Problem 1.9

Number of Deaths	Number of Corps-Years
0	109
1	65
2	22
3	3
4	1
≥ 5	0

1.10 A sample of 100 women suffer from dysmenorrhea. A new analgesic is claimed to provide greater relief than a standard one. After using each analgesic in a crossover experiment, 40 reported greater relief with the standard analgesic and 60 reported greater relief with the new one. Analyze these data.

Theory and Methods

1.11 Why is it easier to get a precise estimate of the binomial parameter π when it is near 0 or 1 than when it is near $\frac{1}{2}$?

1.12 Suppose that $P(Y_i = 1) = 1 - P(Y_i = 0) = \pi$, $i = 1, \ldots, n$, where $\{Y_i\}$ are independent. Let $Y = \sum_i Y_i$.

a. What are $\text{var}(Y)$ and the distribution of Y?

b. When $\{Y_i\}$ instead have pairwise correlation $\rho > 0$, show that $\text{var}(Y) > n\pi(1 - \pi)$, overdispersion relative to the binomial. [Altham (1978) discussed generalizations of the binomial that allow correlated trials.]

c. Suppose that heterogeneity exists: $P(Y_i = 1|\pi) = \pi$ for all i, but π is a random variable with density function $g(\cdot)$ on $[0, 1]$ having mean ρ and positive variance. Show that $\text{var}(Y) > n\rho(1 - \rho)$. (When π has a beta distribution, Y has the *beta-binomial distribution* of Section 13.3.)

d. Suppose that $P(Y_i = 1|\pi_i) = \pi_i$, $i = 1, \ldots, n$, where $\{\pi_i\}$ are independent from $g(\cdot)$. Explain why Y has a $\text{bin}(n, \rho)$ distribution unconditionally but not conditionally on $\{\pi_i\}$. (*Hint:* In each case, is Y a sum of independent, identical Bernoulli trials?)

1.13 For a sequence of independent Bernoulli trials, Y is the number of successes before the kth failure. Explain why its probability mass

function is the *negative binomial*,

$$p(y) = \frac{(y+k-1)!}{y!(k-1)!}\pi^y(1-\pi)^k, \qquad y = 0, 1, 2, \ldots .$$

[For it, $E(Y) = k\pi/(1-\pi)$ and $\text{var}(Y) = k\pi/(1-\pi)^2$, so $\text{var}(Y) > E(Y)$; the Poisson is the limit as $k \to \infty$ and $\pi \to 0$ with $k\pi = \mu$ fixed.]

1.14 For the multinomial distribution, show that

$$\text{corr}(n_j, n_k) = -\pi_j\pi_k/\sqrt{\pi_j(1-\pi_j)\pi_k(1-\pi_k)}\,.$$

Show that $\text{corr}(n_1, n_2) = -1$ when $c = 2$.

1.15 Show that the moment generating function (mgf) for the binomial distribution is $m(t) = (1 - \pi + \pi e^t)^n$, and use it to obtain the first two moments. Show that the mgf for the Poisson distribution is $m(t) = \exp(\mu[\exp(t) - 1])$, and use it to obtain the first two moments.

1.16 A likelihood-ratio statistic equals t_o. At the ML estimates, show that the data are $\exp(t_o/2)$ times more likely under H_a than under H_0.

1.17 Assume that $y_1, y_2, \ldots, y_n$ are independent from a Poisson distribution.

a. Obtain the likelihood function. Show that the ML estimator $\hat{\mu} = \bar{y}$.

b. Construct a large-sample test statistic for H_0: $\mu = \mu_0$ using **(i)** the Wald method, **(ii)** the score method, and **(iii)** the likelihood-ratio method.

c. Construct a large-sample confidence interval for μ using **(i)** the Wald method, **(ii)** the score method, and **(iii)** the likelihood-ratio method.

1.18 Inference for Poisson parameters can often be based on connections with binomial and multinomial distributions. Show how to test H_0: $\mu_1 = \mu_2$ for two populations based on independent Poisson counts (y_1, y_2), using a corresponding test about a binomial parameter π. [*Hint:* Condition on $n = y_1 + y_2$ and identify $\pi = \mu_1/(\mu_1 + \mu_2)$.] How can one construct a confidence interval for μ_1/μ_2 based on one for π?

1.19 A researcher routinely tests using a nominal $P(\text{type I error}) = 0.05$, rejecting H_0 if the P-value ≤ 0.05. An exact test using test statistic T

has null distribution $P(T = 0) = 0.30$, $P(T = 1) = 0.62$, and $P(T = 2) = 0.08$, where a higher T provides more evidence against the null.

a. With the usual P-value, show that the actual P(type I error) = 0.

b. With the mid-P-value, show that the actual P(type I error) = 0.08.

c. Find P(type I error) in parts (a) and (b) when $P(T = 0) = 0.30$, $P(T = 1) = 0.66$, $P(T = 2) = 0.04$. Note that the test with mid-P-value can be conservative or liberal. The exact test with ordinary P-value cannot be liberal.

d. In part (a), a randomized-decision test generates a uniform random variable U from $[0, 1]$ and rejects H_0 when $T = 2$ and $U \leq \frac{5}{8}$. Show the actual P(type I error) = 0.05. Is this a sensible test?

1.20 For a binomial parameter π, show how the inversion process for constructing a confidence interval works with **(a)** the Wald test, and **(b)** the score test.

1.21 For a flip of a coin, let π denote the probability of a head. An experiment tests H_0: $\pi = 0.5$ against H_a: $\pi \neq 0.5$, using $n = 5$ independent flips.

a. Show that the true null probability of rejecting H_0 at the 0.05 significance level is 0.0 for the exact binomial test and $\frac{1}{16}$ using the large-sample score test.

b. Suppose that truly $\pi = 0.5$. Explain why the probability that the 95% Clopper–Pearson confidence interval contains π equals 1.0. (*Hint:* Is there any possible y for which both one-sided tests of H_0: $\pi = 0.5$ have P-value ≤ 0.025?)

1.22 Consider the Wald confidence interval for a binomial parameter π. Since it is degenerate when $\hat{\pi} = 0$ or 1, argue that for $0 < \pi < 1$ the probability the interval covers π cannot exceed $[1 - \pi^n - (1 - \pi)^n]$; hence, the infimum of the coverage probability over $0 < \pi < 1$ equals 0, regardless of n.

1.23 Consider the 95% binomial score confidence interval for π. When $y = 1$, show that the lower limit is approximately $0.18/n$; in fact, $0 < \pi < 0.18/n$ then falls in an interval only when $y = 0$. Argue that for large n and π just barely below $0.18/n$ or just barely above $1 - 0.18/n$, the actual coverage probability is about $e^{-0.18} = 0.84$. Hence, even as $n \to \infty$, this method is not guaranteed to have coverage probability ≥ 0.95 (Agresti and Coull 1998; Blyth and Still 1983).

1.24 From Section 1.4.2 the midpoint $\tilde{\pi}$ of the score confidence interval for π is the sample proportion for an adjusted data set that adds $z^2_{\alpha/2}/2$

observations of each type to the sample. This motivates an adjusted Wald interval,

$$\tilde{\pi} \pm z_{\alpha/2}\sqrt{\tilde{\pi}(1-\tilde{\pi})/n^*}, \qquad \text{where } n^* = n + z^2_{\alpha/2}.$$

Show that the variance $\tilde{\pi}(1-\tilde{\pi})/n^*$ at the weighted average is at least as large as the weighted average of the variances that appears under the square root sign in the score interval (*Hint:* Use Jensen's inequality). Thus, this interval contains the score interval. [Agresti and Coull (1998) and Brown et al. (2001) showed that it performs much better than the Wald interval. It does not have the score interval's disadvantage (Problem 1.23) of poor coverage near 0 and 1.]

1.25 A binomial sample of size n has $y = 0$ successes.

a. Show that the confidence interval for π based on the likelihood function is $[0.0, 1 - \exp(-z^2_{\alpha/2}/2n)]$. For $\alpha = 0.05$, use the expansion of an exponential function to show that this is approximately $[0, 2/n]$.

b. For the score method, show that the confidence interval is $[0, z^2_{\alpha/2}/(n + z^2_{\alpha/2})]$, or approximately $[0, 4/(n+4)]$ when $\alpha = 0.05$.

c. For the Clopper–Pearson approach, show that the upper bound is $1 - (\alpha/2)^{1/n}$, or approximately $-\log(0.025)/n = 3.69/n$ when $\alpha = 0.05$.

d. For the adaptation of the Clopper–Pearson approach using the mid-*P*-value, show that the upper bound is $1 - \alpha^{1/n}$, or approximately $-\log(0.05)/n = 3/n$ when $\alpha = 0.05$.

1.26 For the geometric distribution $p(y) = \pi^y(1-\pi)$, $y = 0, 1, 2, \ldots$, show that the tail method for constructing a confidence interval [i.e., equating $P(Y \geq y)$ and $P(Y \leq y)$ to $\alpha/2$] yields $[(\alpha/2)^{1/y}, (1-\alpha/2)^{1/(y+1)}]$. Show that all π between 0 and $1 - \alpha/2$ *never* fall above a confidence interval, and hence the actual coverage probability exceeds $1 - \alpha/2$ over this region.

1.27 A statistic T has discrete distribution with cdf $F(t)$. Show that $F(T)$ is *stochastically larger* than uniform over $[0, 1]$; that is, its cdf is everywhere no greater than that of the uniform (Casella and Berger 2001, pp. 77, 434). Explain why an implication is that a *P*-value based on T has null distribution that is stochastically larger than uniform.

1.28 Suppose that $P(T = t_j) = \pi_j$, $j = 1, \ldots$. Show that $E(\text{mid-}P\text{-value}) = 0.5$. [*Hint:* Show that $\sum_j \pi_j(\pi_j/2 + \pi_{j+1} + \cdots) = (\sum_j \pi_j)^2/2$.]

1.29 For a statistic T with cdf $F(t)$ and $p(t) = P(T = t)$, the *mid-distribution function* is $F_{\text{mid}}(t) = F(t) - 0.5p(t)$ (Parzen 1997). Given $T = t_o$, show that the mid-*P*-value equals $1 - F(t_o)$. (It also satisfies $E[F_{\text{mid}}(T)] = 0.5$ and $\text{var}[F_{\text{mid}}(T)] = (1/12)\{1 - E[p^2(T)]\}$.)

1.30 Genotypes AA, Aa, and aa occur with probabilities $[\theta^2, 2\theta(1-\theta), (1-\theta)^2]$. A multinomial sample of size n has frequencies (n_1, n_2, n_3) of these three genotypes.

a. Form the log likelihood. Show that $\hat{\theta} = (2n_1 + n_2)/(2n_1 + 2n_2 + 2n_3)$.

b. Show that $-\partial^2 L(\theta)/\partial\theta^2 = [(2n_1 + n_2)/\theta^2] + [(n_2 + 2n_3)/(1-\theta)^2]$ and that its expectation is $2n/\theta(1-\theta)$. Use this to obtain an asymptotic standard error of $\hat{\theta}$.

c. Explain how to test whether the probabilities truly have this pattern.

1.31 Refer to Section 1.5.6. Using the likelihood function to obtain the information, find the approximate standard error of $\hat{\pi}$.

1.32 Refer to Section 1.5.6. Let a denote the number of calves that got a primary, secondary, and tertiary infection, b the number that received a primary and secondary but not a tertiary infection, c the number that received a primary but not a secondary infection, and d the number that did not receive a primary infection. Let π be the probability of a primary infection. Consider the hypothesis that the probability of infection at time t, given infection at times $1, \ldots, t-1$, is also π, for $t = 2, 3$. Show that $\hat{\pi} = (3a + 2b + c)/(3a + 3b + 2c + d)$.

1.33 Refer to quadratic form (1.16).

a. Verify that the matrix quoted in the text for $\boldsymbol{\Sigma}_0^{-1}$ is the inverse of $\boldsymbol{\Sigma}_0$.

b. Show that (1.16) simplifies to Pearson's statistic (1.15).

c. For the z_S statistic (1.11), show that $z_S^2 = X^2$ for $c = 2$.

1.34 For testing H_0: $\pi_j = \pi_{j0}$, $j = 1, \ldots, c$, using sample multinomial proportions $\{\hat{\pi}_j\}$, the likelihood-ratio statistic (1.17) is

$$G^2 = -2n \sum_j \hat{\pi}_j \log\left(\pi_{j0}/\hat{\pi}_j\right).$$

Show that $G^2 \geq 0$, with equality if and only if $\hat{\pi}_j = \pi_{j0}$ for all j. (*Hint:* Apply Jensen's inequality to $E(-2n \log X)$, where X equals $\pi_{j0}/\hat{\pi}_j$ with probability $\hat{\pi}_j$.)

1.35 The chi-squared mgf with df $= \nu$ is $m(t) = (1 - 2t)^{-\nu/2}$, for $|t| < \frac{1}{2}$. Use it to prove the reproductive property of the chi-squared distribution.

1.36 For the multinomial $(n,\{\pi_j\})$ distribution with $c > 2$, confidence limits for π_j are the solutions of

$$\left(\hat{\pi}_j - \pi_j\right)^2 = \left(z_{\alpha/2c}\right)^2 \pi_j(1 - \pi_j)/n, \qquad j = 1, \ldots, c.$$

a. Using the Bonferroni inequality, argue that these c intervals simultaneously contain all $\{\pi_j\}$ (for large samples) with probability at least $1 - \alpha$.

b. Show that the standard deviation of $\hat{\pi}_j - \hat{\pi}_k$ is $[\pi_j + \pi_k - (\pi_j - \pi_k)^2]/n$. For large n, explain why the probability is at least $1 - \alpha$ that the Wald confidence intervals

$$\left(\hat{\pi}_j - \hat{\pi}_k\right) \pm z_{\alpha/2a}\left\{\left[\hat{\pi}_j + \hat{\pi}_k - \left(\hat{\pi}_j - \hat{\pi}_k\right)^2\right]/n\right\}^{1/2}$$

simultaneously contain the $a = c(c - 1)/2$ differences $\{\pi_j - \pi_k\}$ (see Fitzpatrick and Scott 1987; Goodman 1965).

Categorical Data Analysis, Second Edition. Alan Agresti

ISBN: 0-471-36093-7

CHAPTER 2

Describing Contingency Tables

In this chapter we introduce tables that display relationships between categorical variables. We also define parameters that summarize their association. Parameters in Section 2.2 are used to compare groups on the proportions of responses in the outcome categories. The *odds ratio* has special importance, appearing as a parameter in models discussed later. In Section 2.3 we extend the scope by controlling for a third variable. The association can change dramatically under a control. The chapter's primary focus is binary variables, which have only two categories, but in Section 2.4 we present parameters for nominal and ordinal multicategory variables. First, in Section 2.1, we introduce basic terminology and notation.

2.1 PROBABILITY STRUCTURE FOR CONTINGENCY TABLES

The joint distribution between two categorical variables determines their relationship. This distribution also determines the marginal and conditional distributions.

2.1.1 Contingency Tables and Their Distributions

Let X and Y denote two categorical response variables, X with I categories and Y with J categories. Classifications of subjects on both variables have IJ possible combinations. The responses (X, Y) of a subject chosen randomly from some population have a probability distribution. A rectangular table having I rows for categories of X and J columns for categories of Y displays this distribution. The *cells* of the table represent the IJ possible outcomes. When the cells contain frequency counts of outcomes for a sample, the table is called a *contingency table*, a term introduced by Karl Pearson (1904). Another name is *cross-classification table*. A contingency table with I rows and J columns is called an $I \times J$ (or I-by-J) table.

TABLE 2.1 Cross-Classification of Aspirin Use and Myocardial Infarction

	Myocardial Infarction		
	Fatal Attack	Nonfatal Attack	No Attack
Placebo	18	171	10,845
Aspirin	5	99	10,933

Source: Preliminary report: Findings from the aspirin component of the ongoing Physicians' Health Study. *New Engl. J. Med.* **318**: 262–264 (1988).

Table 2.1, a 2×3 contingency table, is from a report on the relationship between aspirin use and heart attacks by the Physicians' Health Study Research Group at Harvard Medical School. The Physicians' Health Study was a 5-year randomized study of whether regular aspirin intake reduces mortality from cardiovascular disease. Every other day, physicians participating in the study took either one aspirin tablet or a placebo. The study was *blind*—those in the study did not know whether they were taking aspirin or a placebo. Of the 11,034 physicians taking a placebo, 18 suffered fatal heart attacks over the course of the study, whereas of the 11,037 taking aspirin, 5 had fatal heart attacks.

Let π_{ij} denote the probability that (X, Y) occurs in the cell in row i and column j. The probability distribution $\{\pi_{ij}\}$ is the *joint distribution* of X and Y. The *marginal distributions* are the row and column totals that result from summing the joint probabilities. We denote these by $\{\pi_{i+}\}$ for the row variable and $\{\pi_{+j}\}$ for the column variable, where the subscript "+" denotes the sum over that index; that is,

$$\pi_{i+} = \sum_j \pi_{ij} \quad \text{and} \quad \pi_{+j} = \sum_i \pi_{ij}.$$

These satisfy $\sum_i \pi_{i+} = \sum_j \pi_{+j} = \sum_i \sum_j \pi_{ij} = 1.0$. The marginal distributions provide single-variable information.

In most contingency tables (such as Table 2.1), one variable, say Y, is a response variable and the other (X) is an explanatory variable. When X is fixed rather than random, the notion of a joint distribution for X and Y is no longer meaningful. However, for a fixed category of X, Y has a probability distribution. It is germane to study how this distribution changes as the category of X changes. Given that a subject is classified in row i of X, $\pi_{j|i}$ denotes the probability of classification in column j of Y, $j = 1, \ldots, J$. Note that $\sum_j \pi_{j|i} = 1$. The probabilities $\{\pi_{1|i}, \ldots, \pi_{J|i}\}$ form the *conditional distribution* of Y at category i of X. A principal aim of many studies is to compare conditional distributions of Y at various levels of explanatory variables.

TABLE 2.2 Estimated Conditional Distributions for Breast Cancer Diagnoses

Breast Cancer	Diagnosis of Test		
	Positive	Negative	Total
Yes	0.82	0.18	1.0
No	0.01	0.99	1.0

Source: Data from W. Lawrence et al., *J. Natl. Cancer Inst.* **90**: 1792–1800 (1998).

2.1.2 Sensitivity and Specificity

The results in Table 2.2 are from a recent article about various methods of attempting to diagnose breast cancer. Based on a literature survey, the authors reported these results for the impact of using mammography together with clinical breast examination. Let X = true disease status (i.e., whether a woman truly has breast cancer) and let Y = diagnosis (positive, negative), where a positive outcome predicts that a woman has breast cancer. The probabilities estimated in Table 2.2 are conditional probabilities of Y given X.

With diagnostic tests for a disease, the two correct diagnoses are a positive test outcome when the subject has the disease and a negative test outcome when a subject does not have it. Given that the subject has the disease, the conditional probability that the diagnostic test is positive is called the *sensitivity*; given that the subject does not have the disease, the conditional probability that the test is negative is called the *specificity* (Yerushalmy 1947). Ideally, these are both high.

For a 2×2 table with the format of Table 2.2, sensitivity is $\pi_{1|1}$ and specificity is $\pi_{2|2}$. In Table 2.2, the estimated sensitivity of combined mammography and clinical examination is 0.82. Of women with breast cancer, 82% are diagnosed correctly. The estimated specificity is 0.99. Of women not having breast cancer, 99% were diagnosed correctly.

2.1.3 Independence of Categorical Variables

When both variables are response variables, descriptions of the association can use their joint distribution, the conditional distribution of Y given X, or the conditional distribution of X given Y. The conditional distribution of Y given X relates to the joint distribution by

$$\pi_{j|i} = \pi_{ij}/\pi_{i+} \qquad \text{for all } i \text{ and } j.$$

Two categorical response variables are defined to be *independent* if all joint probabilities equal the product of their marginal probabilities,

$$\pi_{ij} = \pi_{i+}\pi_{+j} \qquad \text{for } i = 1, \ldots, I \quad \text{and} \quad j = 1, \ldots, J. \tag{2.1}$$

TRADING SOFTWARE

FOR SALE & EXCHANGE

www.trading-software-collection.com

***Subscribe** for FREE download more stuff.*

Mirrors:

www.forex-warez.com
www.traders-software.com

Contacts

andreybbrv@gmail.com
andreybbrv@hotmail.com
andreybbrv@yandex.ru
Skype: andreybbrv
ICQ: 70966433

TABLE 2.3 Notation for Joint, Conditional, and Marginal Probabilities

	Column		
Row	1	2	Total
1	π_{11}	π_{12}	π_{1+}
	$(\pi_{1\|1})$	$(\pi_{2\|1})$	(1.0)
2	π_{21}	π_{22}	π_{2+}
	$(\pi_{1\|2})$	$(\pi_{2\|2})$	(1.0)
Total	π_{+1}	π_{+2}	1.0

When X and Y are independent,

$$\pi_{j|i} = \pi_{ij}/\pi_{i+} = (\pi_{i+}\pi_{+j})/\pi_{i+} = \pi_{+j} \qquad \text{for } i = 1, \ldots, I.$$

Each conditional distribution of Y is identical to the marginal distribution of Y. Thus, two variables are independent when $\{\pi_{j|1} = \cdots = \pi_{j|I}$, for $j = 1, \ldots, J\}$; that is, the probability of any given column response is the same in each row. When Y is a response and X is an explanatory variable, this is a more natural way to define independence than (2.1). Independence is then often referred to as *homogeneity* of the conditional distributions.

Table 2.3 displays notation for joint, conditional, and marginal distributions for the 2×2 case. Sample distributions use similar notation, with p or $\hat{\pi}$ in place of π. For instance, $\{p_{ij}\}$ denotes the sample joint distribution. The cell frequencies are denoted $\{n_{ij}\}$, and $n = \sum_i \sum_j n_{ij}$ is the total sample size. Thus,

$$p_{ij} = n_{ij}/n.$$

The sample proportion of times that subjects in row i made response j is

$$p_{j|i} = p_{ij}/p_{i+} = n_{ij}/n_{i+},$$

where $n_{i+} = np_{i+} = \sum_j n_{ij}$.

2.1.4 Poisson, Binomial, and Multinomial Sampling

The probability distributions introduced in Section 1.2 extend to cell counts in contingency tables. For instance, a Poisson sampling model treats cell counts $\{Y_{ij}\}$ as independent Poisson random variables with parameters $\{\mu_{ij}\}$. The joint probability mass function for potential outcomes $\{n_{ij}\}$ is then the product of the Poisson probabilities $P(Y_{ij} = n_{ij})$ for the IJ cells, or

$$\prod_i \prod_j \exp(-\mu_{ij})\, \mu_{ij}^{n_{ij}}/n_{ij}!\,.$$

When the total sample size n is fixed but the row and column totals are not, a *multinomial sampling* model applies. The IJ cells are the possible outcomes. The probability mass function of the cell counts has the multinomial form

$$[n!/(n_{11}! \cdots n_{IJ}!)] \prod_i \prod_j \pi_{ij}^{n_{ij}}.$$

Often, observations on a response Y occur separately at each setting of an explanatory variable X. This case normally treats row totals as fixed, and for simplicity, we use the notation $n_i = n_{i+}$. Suppose that the n_i observations on Y at setting i of X are independent, each with probability distribution $\{\pi_{1|i}, \ldots, \pi_{J|i}\}$. The counts $\{n_{ij},\ j = 1, \ldots, J\}$ satisfying $\sum_j n_{ij} = n_i$ then have the multinomial form

$$\frac{n_i!}{\prod_j n_{ij}!} \prod_j \pi_{j|i}^{n_{ij}}. \tag{2.2}$$

When samples at different settings of X are independent, the joint probability function for the entire data set is the product of the multinomial functions (2.2) from the various settings. This sampling scheme is *independent multinomial sampling*, also called *product multinomial sampling*.

Independent multinomial sampling also results under the following conditions: Suppose that $\{n_{ij}\}$ result from either independent Poisson sampling with means $\{\mu_{ij}\}$ or multinomial sampling over the IJ cells with probabilities $\{\pi_{ij} = \mu_{ij}/n\}$. When X is an explanatory variable, it is sensible to perform statistical inference conditional on the totals $\{n_i = \sum_j n_{ij}\}$ even when their values are not fixed by the sampling design. Conditional on $\{n_i\}$, the cell counts $\{n_{ij}, j = 1, \ldots, J\}$ have the multinomial distribution (2.2) with response probabilities $\{\pi_{j|i} = \mu_{ij}/\mu_{i+},\ j = 1, \ldots, J\}$, and cell counts from different rows are independent. With this conditioning, we treat the row totals as fixed and analyze the data as if they formed separate independent samples.

Sometimes both row and column margins are naturally fixed. The appropriate sampling distribution is then the *hypergeometric*. In Section 3.5.1 we discuss this case, which is less common.

2.1.5 Seat Belt Example

Researchers in the Massachusetts Highway Department plan to study the relationship between seat-belt use (yes, no) and outcome of an automobile crash (fatality, nonfatality) for drivers involved in accidents on the Massachusetts Turnpike. They will summarize results in the format shown in Table 2.4. They plan to catalog all accidents on the turnpike for the next year, classifying each according to these variables. The total sample size is

TABLE 2.4 Seat-Belt Use and Results of Automobile Crashes

	Result of Crash	
Seat-Belt Use	Fatality	Nonfatality
Yes		
No		

then a random variable. They might treat the numbers of observations at the four combinations of seat-belt use and outcome of crash as independent Poisson random variables with unknown means $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}\}$.

Suppose, instead, that the researchers randomly sample 200 police records of crashes on the turnpike in the past year and classify each according to seat-belt use and outcome of crash. For this study, the total sample size n is fixed. They might then treat the four cell counts as a multinomial random variable with $n = 200$ trials and unknown joint probabilities $\{\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}\}$.

Suppose, instead, that police records for accidents involving fatalities were filed separately from the others. The researchers might instead randomly sample 100 records of accidents with a fatality and randomly sample 100 records of accidents with no fatality. This approach fixes the column totals in Table 2.4 at 100. They might then regard each column of Table 2.4 as an independent binomial sample. Yet another approach, the traditional experimental design, takes 200 subjects and randomly assigns 100 of them to wear seat belts; the 200 then all are forced to have an accident. The recorded results would then be independent binomial samples in each row, with fixed row totals of 100 each. (Obviously, traditional designs common in some experimental science may not be ethical for humans. This is especially true in medical studies.)

2.1.6 Types of Studies

Table 2.5 comes from one of the first studies of the link between lung cancer and smoking, by Richard Doll and A. Bradford Hill. In 20 hospitals in London, England, patients admitted with lung cancer in the preceding year were queried about their smoking behavior. For each of the 709 patients admitted, researchers studied the smoking behavior of a noncancer patient at the same hospital of the same gender and within the same 5-year grouping on age. The 709 *cases* in the first column of Table 2.5 are those having lung cancer and the 709 *controls* in the second column are those not having it. A smoker was defined as a person who had smoked at least one cigarette a day for at least a year.

Normally, whether lung cancer occurs is a response variable and smoking behavior is an explanatory variable. In this study, however, the marginal

TABLE 2.5 Cross-Classification of Smoking by Lung Cancer

	Lung Cancer	
Smoker	Cases	Controls
Yes	688	650
No	21	59
Total	709	709

Source: Based on data reported in Table IV, R. Doll and A. B. Hill, *British Med. J.*, Sept. 30, 1950, pp. 739–748.

distribution of lung cancer is fixed by the sampling design, and the outcome measured is whether the subject ever was a smoker. The study, which uses a *retrospective* design to "look into the past," is called a *case–control study*. Such studies are common in health-related applications. Often, the two samples are matched, as in this study. Sometimes the samples of cases and controls are independent rather than matched. For instance, another early case–control study on lung cancer and smoking sampled subjects by sending letters to the estates of physicians who had died of some type of cancer in 1950 or 1951, and observations were cross-classified on type of cancer and the subject's smoking behavior (see, e.g., Cornfield 1956).

One might want to compare smokers with nonsmokers in terms of the proportion who suffered lung cancer. These proportions refer to the conditional distribution of lung cancer, given smoking behavior. Instead, case–control studies provide proportions in the reverse direction, for the conditional distribution of smoking behavior, given lung cancer status. For those in Table 2.5 with lung cancer, the proportion who were smokers was $688/709 = 0.970$, while it was $650/709 = 0.917$ for the controls.

When we know the proportion of the population having lung cancer, we can use Bayes' theorem to compute sample conditional distributions in the direction of main interest (Problem 2.21). Otherwise, using a retrospective sample, we cannot estimate the probability of lung cancer at each category of smoking behavior. For Table 2.5 we do not know the population prevalence of lung cancer, and the patients suffering it were probably sampled at a rate far in excess of their occurrence in the general population.

By contrast, imagine a study that samples subjects from the population of teenagers and then 60 years later measures the rates of lung cancer for the smokers and nonsmokers. Such a sampling design is *prospective*. There are two types of prospective studies. *Clinical trials* randomly allocate subjects to the groups who will be smokers and nonsmokers. In *cohort studies*, subjects make their own choice about whether to smoke, and the study observes in future time who develops lung cancer. Yet another approach, a *cross-sectional design*, samples subjects and classifies them simultaneously on both variables.

Prospective studies usually condition on the totals $\{n_i = \sum_j n_{ij}\}$ for categories of X and regard each row of J counts as an independent multinomial sample on Y. *Retrospective studies* usually treat the totals $\{n_{+j}\}$ for Y as fixed and regard each column of I counts as a multinomial sample on X. In *cross-sectional studies*, the total sample size is fixed but not the row or column totals, and the IJ cell counts are a multinomial sample.

Case–control, cohort, and cross-sectional studies are called *observational studies*. They simply observe who chooses each group and who has the outcome of interest. By contrast, a clinical trial is an *experimental* study, the investigator having the advantage of experimental control over which subjects receive each treatment. Such studies can use the power of randomization to make the groups balance roughly on other variables that may be associated with the response. Observational studies are common but have more potential for biases of various types.

2.2 COMPARING TWO PROPORTIONS

Many studies are designed to compare groups on a binary response variable. Then Y has only two categories, such as (success, failure) for outcome of a medical treatment. With two groups, a 2×2 contingency table displays the results. The rows are the groups and the columns are the categories of Y. This section presents parameters for comparing the groups.

2.2.1 Difference of Proportions

For subjects in row i, $\pi_{1|i}$ is the probability that the response has outcome in category 1 ("success"). With only two possible outcomes, $\pi_{2|i} = 1 - \pi_{1|i}$, and we use the simpler notation π_i for $\pi_{1|i}$. The *difference of proportions* of successes, $\pi_1 - \pi_2$, is a basic comparison of the two rows. Comparison on failures is equivalent to comparison on successes, since

$$(1 - \pi_1) - (1 - \pi_2) = \pi_2 - \pi_1.$$

The difference of proportions falls between -1.0 and $+1.0$. It equals zero when the rows have identical conditional distributions. The response Y is statistically independent of the row classification when $\pi_1 - \pi_2 = 0$.

When both variables are responses, conditional distributions apply in either direction. One can also compare the two columns, such as by the difference between the proportions in row 1. This usually is not equal to the difference $\pi_1 - \pi_2$ comparing the rows.

2.2.2 Relative Risk

A value $\pi_1 - \pi_2$ of fixed size may have greater importance when both π_i are close to 0 or 1 than when they are not. For a study comparing two

treatments on the proportion of subjects who die, the difference between 0.010 and 0.001 may be more noteworthy than the difference between 0.410 and 0.401, even though both are 0.009. In such cases, the ratio of proportions is also informative.

The *relative risk* is defined to be the ratio

$$\pi_1/\pi_2. \tag{2.3}$$

It can be any nonnegative real number. A relative risk of 1.0 corresponds to independence. For the proportions just given, the relative risks are $0.010/0.001 = 10.0$ and $0.410/0.401 = 1.02$. Comparing the rows on the second response category gives a different relative risk, $(1 - \pi_1)/(1 - \pi_2)$.

2.2.3 Odds Ratio

For a probability π of success, the *odds* are defined to be

$$\Omega = \pi/(1 - \pi).$$

The odds are nonnegative, with $\Omega > 1.0$ when a success is more likely than a failure. When $\pi = 0.75$, for instance, then $\Omega = 0.75/0.25 = 3.0$; a success is three times as likely as a failure, and we expect about three successes for every one failure. When $\Omega = \frac{1}{3}$, a failure is three times as likely as a success. Inversely,

$$\pi = \Omega/(\Omega + 1).$$

For instance, when $\Omega = \frac{1}{3}$, then $\pi = 0.25$.

Refer again to a 2×2 table. Within row i, the odds of success instead of failure are $\Omega_i = \pi_i/(1 - \pi_i)$. The ratio of the odds Ω_1 and Ω_2 in the two rows,

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)} \tag{2.4}$$

is called the *odds ratio*.

For joint distributions with cell probabilities $\{\pi_{ij}\}$, the equivalent definition for the odds in row i is $\Omega_i = \pi_{i1}/\pi_{i2}$, $i = 1, 2$. Then the odds ratio is

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}. \tag{2.5}$$

An alternative name for θ is the *cross-product ratio*, since it equals the ratio of the products $\pi_{11}\pi_{22}$ and $\pi_{12}\pi_{21}$ of probabilities from diagonally opposite cells (Yule 1900, 1912).

2.2.4 Properties of the Odds Ratio

The odds ratio can equal any nonnegative number. The condition $\Omega_1 = \Omega_2$ and hence (when all cell probabilities are positive) $\theta = 1$ corresponds to independence of X and Y. When $1 < \theta < \infty$, subjects in row 1 are more likely to have a success than are subjects in row 2; that is, $\pi_1 > \pi_2$. For instance, when $\theta = 4$, the odds of success in row 1 are four times the odds in row 2. This does not mean that the *probability* $\pi_1 = 4\pi_2$; that is the interpretation of a *relative risk* of 4.0. When $0 < \theta < 1$, $\pi_1 < \pi_2$. When one cell has zero probability, θ equals 0 or ∞.

Values of θ farther from 1.0 in a given direction represent stronger association. Two values represent the same association, but in opposite directions, when one is the inverse of the other. For instance, when $\theta = 0.25$, the odds of success in row 1 are 0.25 times the odds in row 2, or equivalently, the odds of success in row 2 are $1/0.25 = 4.0$ times the odds in row 1. When the order of the rows is reversed or the order of the columns is reversed, the new value for θ is the inverse of the original value.

For inference, we shall see it is convenient to use $\log \theta$. Independence corresponds to $\log \theta = 0$. The log odds ratio is symmetric about this value—reversal of rows or of columns results in a change in its sign. Two values for $\log \theta$ that are the same except for sign, such as $\log 4 = 1.39$ and $\log 0.25 = -1.39$, represent the same strength of association.

The odds ratio does not change value when the orientation of the table reverses so that the rows become the columns and the columns become the rows. This is clear from the symmetric form of (2.5). It is unnecessary to identify one classification as the response variable in order to use θ. In fact, although (2.4) defined it in terms of odds using $\pi_i = P(Y = 1 \mid X = i)$, one could just as well define it using reverse conditional probabilities. With a joint distribution, conditional distributions exist in each direction, and

$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{P(Y = 1 \mid X = 1)/P(Y = 2 \mid X = 1)}{P(Y = 1 \mid X = 2)/P(Y = 2 \mid X = 2)}$$
$$= \frac{P(X = 1 \mid Y = 1)/P(X = 2 \mid Y = 1)}{P(X = 1 \mid Y = 2)/P(X = 2 \mid Y = 2)}. \tag{2.6}$$

In fact, the odds ratio is equally valid for prospective, retrospective, or cross-sectional sampling designs. The sample odds ratio estimates the same parameter in each case.

For cell counts $\{n_{ij}\}$, the sample odds ratio is

$$\hat{\theta} = n_{11}n_{22}/n_{12}n_{21}.$$

This does not change when both cell counts within any row are multiplied by a nonzero constant or when both cell counts within any column are multiplied by a nonzero constant. An implication is that the sample odds ratio

estimates the same characteristic (θ) even when the sample is disproportionately large or small from marginal categories of a variable. For a retrospective study of the association between vaccination and catching a certain strain of flu, the sample odds ratio estimates the same characteristic with a random sample of (1) 100 people who got the flu and 100 people who did not, or (2) 40 people who got the flu and 160 people who did not. The sample versions of the difference of proportions and relative risk (2.3) are invariant to multiplication of counts within rows by a constant, but they change with multiplication within columns or with row–column interchange.

2.2.5 Aspirin and Heart Attacks Revisited

We illustrate the three association measures with Table 2.1 on aspirin use and heart attacks. The table differentiates between fatal and nonfatal heart attacks, but we combine these outcomes for now. Of the 11,034 physicians taking placebo, 189 suffered heart attacks, a proportion of $189/11{,}034 = 0.0171$. Of the 11,037 taking aspirin, 104 had heart attacks, a proportion of 0.0094. The sample difference of proportions is $0.0171 - 0.0094 = 0.0077$. The relative risk is $0.0171/0.0094 = 1.82$. The proportion suffering heart attacks of those taking placebo was 1.82 times the proportion suffering heart attacks of those taking aspirin. The sample odds ratio is $(189 \times 10{,}933)/(10{,}845 \times 104) = 1.83$. The odds of heart attack for those taking placebo was 1.83 times the odds for those taking aspirin.

2.2.6 Case–Control Studies and the Odds Ratio

With retrospective sampling designs, such as case–control studies, it is possible to estimate conditional probabilities of form $P(X = i \mid Y = j)$. It is usually not possible to estimate the probability $P(Y = j \mid X = i)$ of an outcome of interest or the difference of proportions or relative risk for that outcome. It is possible to estimate the odds ratio, however, since by (2.6) it is determined by conditional probabilities in *either* direction.

To illustrate, we revisit Table 2.5 on X = smoking behavior and Y = lung cancer. The data were two binomial samples on X at fixed levels of Y. Thus, we can estimate the probability a subject was a smoker, given the outcome on whether the subject had lung cancer; this was $688/709$ for the cases and $650/709$ for the controls. We cannot estimate the probability of lung cancer, given whether one smoked, which is more relevant. Thus, we cannot estimate differences or ratios of probabilities of lung cancer. The difference of proportions and relative risk are limited to comparisons of the probabilities of being a smoker. However, we can compute the odds ratio using the sample analog of (2.6),

$$\frac{(688/709)/(21/709)}{(650/709)/(59/709)} = \frac{688 \times 59}{650 \times 21} = 3.0.$$

Moreover, by (2.6), interpretations can use the direction of interest, even though the study was retrospective: The estimated odds of lung cancer for smokers were 3.0 times the estimated odds for nonsmokers.

2.2.7 Relationship between Odds Ratio and Relative Risk

From definitions (2.3) and (2.4),

$$\text{odds ratio} = \text{relative risk}\left(\frac{1-\pi_2}{1-\pi_1}\right).$$

Their magnitudes are similar whenever the probability π_i of the outcome of interest is close to zero for both groups. We saw this similarity in Section 2.2.5 for the aspirin study, where the heart attack proportion was less than 0.02 for each group. The relative risk was 1.82 and the odds ratio was 1.83.

Because of this similarity, when each π_i is small, the odds ratio provides a rough indication of the relative risk when it is not directly estimable, such as in case–control studies (Cornfield 1951). For instance, for Table 2.5, if the probability of lung cancer is small regardless of smoking behavior, 3.0 is also a rough estimate of the relative risk; that is, smokers had about 3.0 times the relative frequency of lung cancer as nonsmokers.

2.3 PARTIAL ASSOCIATION IN STRATIFIED 2 × 2 TABLES

An important part of most studies, especially observational studies, is the choice of control variables. In studying the effect of X on Y, one should control any covariate that can influence that relationship. This involves using some mechanism to hold the covariate constant. Otherwise, an observed effect of X on Y may actually reflect effects of that covariate on both X and Y. The relationship between X and Y then shows *confounding*. Experimental studies can remove effects of confounding covariates by randomly assigning subjects to different levels of X, but this is not possible with observational studies.

Suppose that a study considers effects of passive smoking, the effects on a nonsmoker of living with a smoker. To analyze whether passive smoking is associated with lung cancer, a cross-sectional study might compare lung cancer rates between nonsmokers whose spouses smoke and nonsmokers whose spouses do not smoke. The study should attempt to control for age, socioeconomic status, or other factors that might relate both to spouse smoking and to developing lung cancer. Otherwise, results will have limited usefulness. Spouses of nonsmokers may tend to be younger than spouses of smokers, and younger people are less likely to have lung cancer. Then a lower proportion of lung cancer cases among spouses of nonsmokers may merely reflect their lower average age.

In this section we discuss the analysis of the association between categorical variables X and Y while controlling for a possibly confounding variable Z. For simplicity, the examples refer to a single control variable. In later chapters we treat more general cases and discuss the use of models to perform statistical control.

2.3.1 Partial Tables

We control for Z by studying the XY relationship at fixed levels of Z. Two-way cross-sectional slices of the three-way contingency table cross classify X and Y at separate categories of Z. These cross sections are called *partial tables*. They display the XY relationship while removing the effect of Z by holding its value constant.

The two-way contingency table obtained by combining the partial tables is called the *XY marginal table*. Each cell count in the marginal table is a sum of counts from the same location in the partial tables. The marginal table, rather than controlling Z, ignores it. The marginal table contains no information about Z. It is simply a two-way table relating X and Y but may reflect the effects of Z on X and Y.

The associations in partial tables are called *conditional associations*, because they refer to the effect of X on Y conditional on fixing Z at some level. Conditional associations in partial tables can be quite different from associations in marginal tables. In fact, it can be misleading to analyze only marginal tables of a multiway contingency table. The following example illustrates.

2.3.2 Death Penalty Example

Table 2.6 is a $2 \times 2 \times 2$ contingency table–two rows, two columns, and two layers–from an article that studied effects of racial characteristics on whether persons convicted of homicide received the death penalty. The 674 subjects classified in Table 2.6 were the defendants in indictments involving cases

TABLE 2.6 Death Penalty Verdict by Defendant's Race and Victims' Race

Victims' Race	Defendant's Race	Death Penalty		Percent Yes
		Yes	No	
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8
Total	White	53	430	11.0
	Black	15	176	7.9

Source: M. L. Radelet and G. L. Pierce, *Florida Law Rev.* **43**: 1–34 (1991). Reprinted with permission from the *Florida Law Review*.

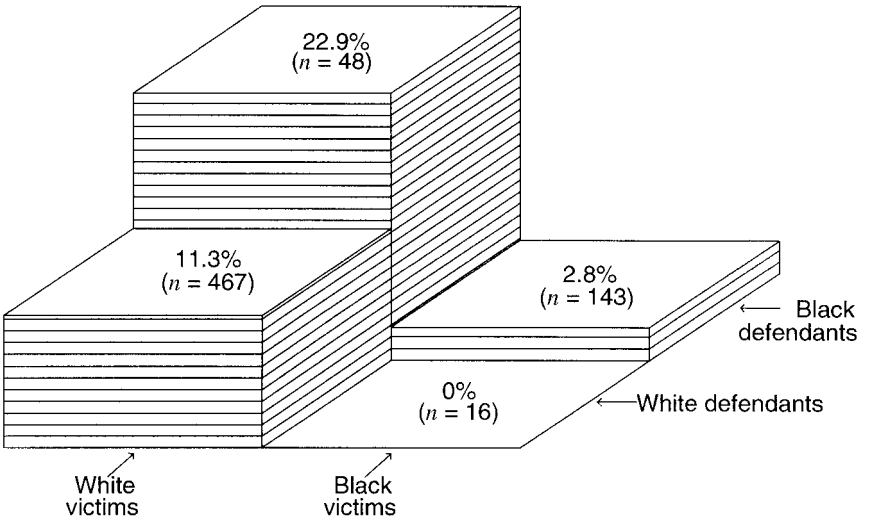


FIGURE 2.1 Percent receiving death penalty.

with multiple murders in Florida between 1976 and 1987. The variables in Table 2.6 are Y = death penalty verdict, having the categories (yes, no), X = race of defendant, and Z = race of victims, each having the categories (white, black). We study the effect of defendant's race on the death penalty verdict, treating victims' race as a control variable. Table 2.6 has a 2×2 partial table relating defendant's race and the death penalty verdict at each category of victims' race.

For each combination of defendant's race and victims' race, Table 2.6 lists and Figure 2.1 displays the percentage of defendants who received the death penalty. These describe the conditional associations. When the victims were white, the death penalty was imposed $22.9\% - 11.3\% = 11.6\%$ more often for black defendants than for white defendants. When the victims were black, the death penalty was imposed 2.8% more often for black defendants than for white defendants. *Controlling* for victims' race by keeping it fixed, the death penalty was imposed more often on black defendants than on white defendants.

The bottom portion of Table 2.6 displays the marginal table. It results from summing the cell counts in Table 2.6 over the two categories of victims' race, thus combining the two partial tables (e.g., $11 + 4 = 15$). Overall, 11.0% of white defendants and 7.9% of black defendants received the death penalty. *Ignoring* victims' race, the death penalty was imposed less often on black defendants than on white defendants. The association reverses direction compared to the partial tables.

Why does the association change so much when we ignore versus control victims' race? This relates to the nature of the association between victims' race and each of the other variables. First, the association between victims'

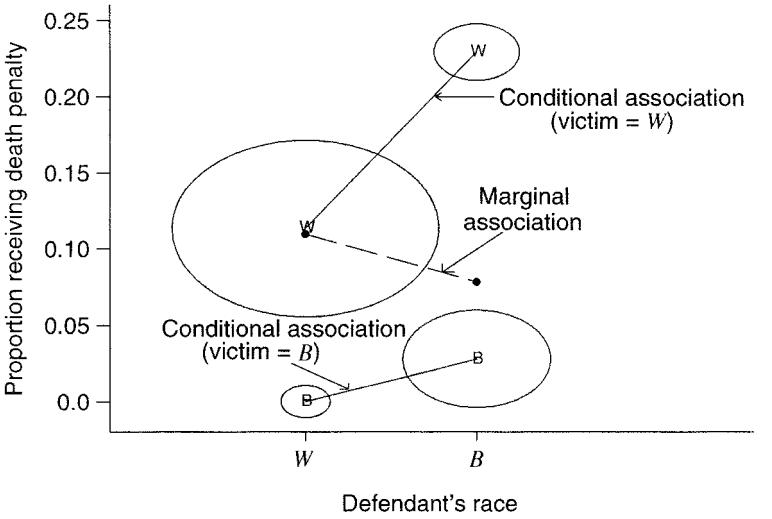


FIGURE 2.2 Proportion receiving death penalty by defendant's race, controlling and ignoring victims' race.

race and defendant's race is extremely strong. The marginal table relating these variables has odds ratio $(467 \times 143)/(48 \times 16) = 87.0$. Second, Table 2.6 shows that, regardless of defendant's race, the death penalty was much more likely when the victims were white than when the victims were black. So whites are tending to kill whites, and killing whites is more likely to result in the death penalty. This suggests that the marginal association should show a greater tendency than the conditional associations for white defendants to receive the death penalty. In fact, Table 2.6 has this pattern.

Figure 2.2 illustrates why the marginal association differs so from the conditional associations. For each defendant's race, the figure plots the proportion receiving the death penalty at each category of victims' race. Each proportion is labeled by a letter symbol giving the category of victims' race. Surrounding each observation is a circle having area proportional to the number of observations at that combination of defendant's race and victims' race. For instance, the W in the largest circle represents a proportion of 0.113 receiving the death penalty for cases with white defendants and white victims. That circle is largest because the number of cases at that combination $(53 + 414 = 467)$ is largest. The next-largest circle relates to cases in which blacks kill blacks.

We control for victims' race by comparing circles having the same victims' race letter at their centers. The line connecting the two W circles has a positive slope, as does the line connecting the two B circles. Controlling for victims' race, this reflects the death penalty being more likely for black defendants than for white defendants. When we add results across victims'

race to get a summary result for the marginal effect of defendant's race on the death penalty verdict, the larger circles, having the greater number of cases, have greater influence. Thus, the summary proportions for each defendant's race, marked on the figure by periods, fall closer to the center of the larger circles than to the center of the smaller circles. A line connecting the summary marginal proportions has negative slope, indicating that overall the death penalty was more likely for white defendants than for black defendants.

The result that a marginal association can have a different direction from each conditional association is called *Simpson's paradox* (Simpson 1951, Yule 1903). It applies to quantitative as well as categorical variables. Statisticians commonly use it to caution against imputing causal effects from an association of X with Y. For instance, when doctors started to observe strong odds ratios between smoking and lung cancer, statisticians such as R. A. Fisher warned that some variable (e.g., a genetic factor) could exist such that the association would disappear under the relevant control. However, other statisticians (such as J. Cornfield) showed that with a very strong XY association, a very strong association must exist between the confounding variable Z and both X and Y in order for the effect to disappear or change under the control (Breslow and Day 1980, Sec. 3.4).

2.3.3 Conditional and Marginal Odds Ratios

Odds ratios can describe marginal and conditional associations. We illustrate for $2 \times 2 \times K$ tables, where K denotes the number of categories of a control variable, Z. Let $\{\mu_{ijk}\}$ denote cell expected frequencies for some sampling model, such as binomial, multinomial, or Poisson sampling.

Within a fixed category k of Z, the odds ratio

$$\theta_{XY(k)} = \frac{\mu_{11k}\,\mu_{22k}}{\mu_{12k}\,\mu_{21k}} \tag{2.7}$$

describes conditional XY association in partial table k. The odds ratios for the K partial tables are called *XY conditional odds ratios*. These can be quite different from marginal odds ratios. The XY marginal table has expected frequencies $\{\mu_{ij+} = \sum_k \mu_{ijk}\}$. The XY marginal odds ratio is

$$\theta_{XY} = \frac{\mu_{11+}\,\mu_{22+}}{\mu_{12+}\,\mu_{21+}}\,.$$

Sample values of $\theta_{XY(k)}$ and θ_{XY} use similar formulas with cell counts substituted for expected frequencies. We illustrate for the association between defendant's race and the death penalty in Table 2.6. In the first partial

table, victims' race is white and

$$\hat{\theta}_{XY(1)} = \frac{53 \times 37}{414 \times 11} = 0.43.$$

The sample odds for white defendants receiving the death penalty were 43% of the sample odds for black defendants. In the second partial table, victims' race is black and the estimated odds ratio equals $\hat{\theta}_{XY(2)} = (0 \times 139)(16 \times 4) = 0.0$, since the death penalty was never given to white defendants with black victims.

Estimation of the marginal odds ratio uses the 2×2 marginal table within Table 2.6, collapsing over victims' race, or $(53 \times 176)/(430 \times 15) = 1.45$. The sample odds of the death penalty were 45% higher for white defendants than for black defendants. Yet within each victims' race category, those odds were smaller for white defendants. This reversal in the association after controlling for victims' race illustrates Simpson's paradox.

2.3.4 Marginal versus Conditional Independence

More generally, X may have I categories and Y may have J categories. An $I \times J \times K$ table describes the relationship between X and Y, controlling for Z. If X and Y are independent in partial table k, then X and Y are called *conditionally independent at level k* of Z. When Y is a response, this means that

$$P(Y = j \,|\, X = i, Z = k) = P(Y = j \,|\, Z = k), \qquad \text{for all } i, j. \tag{2.8}$$

More generally, X and Y are said to be *conditionally independent given Z* when they are conditionally independent at every level of Z, that is, when (2.8) holds for all k. Then, given Z, Y does not depend on X.

Suppose that a single multinomial applies to the entire three-way table, with joint probabilities $\{\pi_{ijk} = P(X = i, Y = j, Z = k)\}$. Then

$$\pi_{ijk} = P(X = i, Z = k)\, P(Y = j \,|\, X = i, Z = k),$$

which under conditional independence of X and Y, given Z, equals

$$= \pi_{i+k} P(Y = j \,|\, Z = k) = \pi_{i+k} P(Y = j, Z = k)/P(Z = k)\,.$$

Thus, conditional independence is then equivalent to

$$\pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k} \qquad \text{for all } i, j, \text{ and } k. \tag{2.9}$$

TABLE 2.7 Expected Frequencies Showing That Conditional Independence Does Not Imply Marginal Independence

		Response	
Clinic	Treatment	Success	Failure
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32
Total	A	20	20
	B	20	40

Conditional independence does not imply marginal independence (Yule 1903). For instance, summing (2.9) over k on both sides yields

$$\pi_{ij+} = \sum_k \left(\pi_{i+k} \pi_{+jk} / \pi_{++k} \right).$$

All three terms in the summation involve k, and this does not simplify to $\pi_{ij+} = \pi_{i++} \pi_{+j+}$, marginal independence.

For $2 \times 2 \times K$ tables, X and Y are conditionally independent when the odds ratio between X and Y equals 1 at each category of Z. The expected frequencies $\{\mu_{ijk}\}$ in Table 2.7 illustrate this relation for $Y =$ response (success, failure), $X =$ drug treatment (A, B), and $Z =$ clinic $(1, 2)$. From (2.7), the conditional XY odds ratios are

$$\theta_{XY(1)} = \frac{18 \times 8}{12 \times 12} = 1.0, \qquad \theta_{XY(2)} = \frac{2 \times 32}{8 \times 8} = 1.0.$$

Given the clinic, response and treatment are conditionally independent. The marginal table combines the tables for the two clinics. Its odds ratio is $\theta_{XY} = (20 \times 40)/(20 \times 20) = 2.0$, so the variables are not marginally independent.

Ignoring the clinic, why are the odds of a success for treatment A twice those for treatment B? The conditional XZ and YZ odds ratios give a clue. The odds ratio between Z and either X or Y, at each fixed category of the other variable, equals 6.0. For instance, the XZ odds ratio at the first category of Y equals $(18 \times 8)/(12 \times 2) = 6.0$. The conditional odds (given response) of receiving treatment A at clinic 1 are six times those at clinic 2, and the conditional odds (given treatment) of success at clinic 1 are six times those at clinic 2. Clinic 1 tends to use treatment A more often, and clinic 1 also tends to have more successes. For instance, if patients at clinic 1 tended to be younger and in better health than those at clinic 2, perhaps they had a better success rate regardless of the treatment received.

It is misleading to study only the marginal table, concluding that successes are more likely with treatment A. Subjects within a particular clinic are likely to be more homogeneous than the overall sample, and response is independent of treatment in each clinic.

2.3.5 Homogeneous Association

A $2 \times 2 \times K$ table has *homogeneous XY association* when

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}.$$

Then the effect of X on Y is the same at each category of Z. Conditional independence of X and Y is the special case in which each $\theta_{XY(k)} = 1.0$.

Under homogeneous XY association, homogeneity also holds for the other associations. For instance, the conditional odds ratio between two categories of X and two categories of Z is identical at each category of Y. For the odds ratio, homogeneous association is a symmetric property. It applies to any pair of variables viewed across the categories of the third. When it occurs, there is said to be *no interaction* between two variables in their effects on the other variable.

When interaction exists, the conditional odds ratio for any pair of variables changes across categories of the third. For X = smoking (yes, no), Y = lung cancer (yes, no), and Z = age (< 45, 45–65, > 65), suppose that $\theta_{XY(1)} = 1.2$, $\theta_{XY(2)} = 3.9$, and $\theta_{XY(3)} = 8.8$. Then smoking has a weak effect on lung cancer for young people, but the effect strengthens considerably with age. Age is called an *effect modifier*; the effect of smoking is modified depending on its value.

For the death penalty data (Table 2.6), $\hat{\theta}_{XY(1)} = 0.43$ and $\hat{\theta}_{XY(2)} = 0.0$. The values are not close, but the second estimate is unstable because of the zero cell count. Adding $\frac{1}{2}$ to each cell count, $\hat{\theta}_{XY(2)} = 0.94$. Because $\hat{\theta}_{XY(2)}$ is unstable and because further variation occurs from sampling variability, these partial tables do not necessarily contradict homogeneous association in a population. In Section 6.3 we show how to analyze whether sample data are consistent with homogeneous association or conditional independence.

2.4 EXTENSIONS FOR $I \times J$ TABLES

For 2×2 tables, a single number such as the odds ratio can summarize the association. For $I \times J$ tables, it is rarely possible to summarize association by a single number without some loss of information. However, a set of odds ratios or another summary index can describe certain features of the association.

2.4.1 Odds Ratios in $I \times J$ Tables

Odds ratios can use each of the $\binom{I}{2} = I(I-1)/2$ pairs of rows in combination with each of the $\binom{J}{2} = J(J-1)/2$ pairs of columns. For rows a and b and columns c and d, the odds ratio $(\pi_{ac}\pi_{bd})/(\pi_{bc}\pi_{ad})$ uses four cells in a rectangular pattern. There are $\binom{I}{2}\binom{J}{2}$ odds ratios of this type. This set of odds ratios contains much redundant information.

Consider the subset of $(I-1)(J-1)$ *local odds ratios*

$$\theta_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}}, \qquad i = 1, \ldots, I-1, \quad j = 1, \ldots, J-1. \quad (2.10)$$

Figure 2.3 shows that local odds ratios use cells in adjacent rows and adjacent columns. These $(I-1)(J-1)$ odds ratios determine all odds ratios formed from pairs of rows and pairs of columns. To illustrate, in Table 2.1, the sample local odds ratio is 2.08 for the first two columns and 1.74 for the

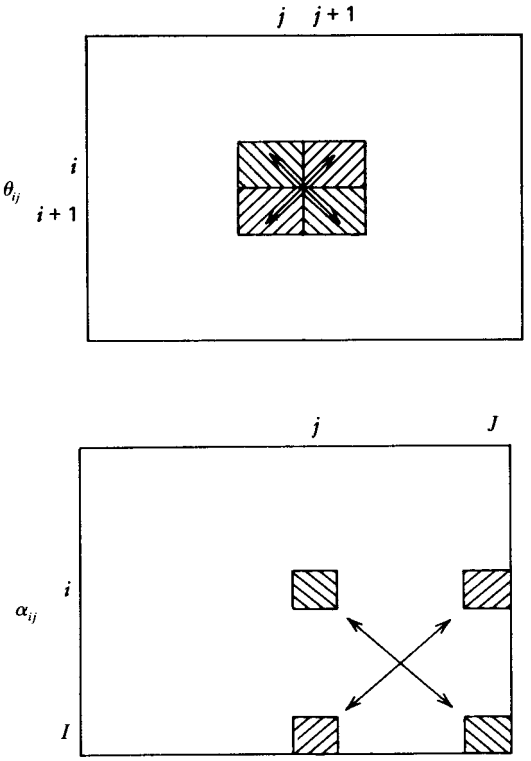


FIGURE 2.3 Odds ratios for $I \times J$ tables.

second and third columns. In each case, the more serious outcome was more prevalent for the placebo group. The product of these two odds ratios is 3.63, which is the odds ratio for the first and third columns.

Construction (2.10) for a minimal set of odds ratios is not unique. Another basic set is

$$\alpha_{ij} = \frac{\pi_{ij}\pi_{IJ}}{\pi_{Ij}\pi_{iJ}}, \qquad i = 1, \ldots, I-1, \quad j = 1, \ldots, J-1. \tag{2.11}$$

This uses the rectangular pattern of cells determined by the cell in row i and column j and the cell in the last row and last column. Figure 2.3 illustrates.

Given the marginal distributions $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$, when $\{\pi_{ij} > 0\}$, conversion of the probabilities into the set of odds ratios (2.10) or (2.11) does not discard information. The cell probabilities determine the odds ratios, and given the marginals, the odds ratios determine the cell probabilities. In this sense, $(I-1)(J-1)$ parameters can describe any association in an $I \times J$ table. Independence is equivalent to all $(I-1)(J-1)$ odds ratios equaling 1.0.

For three-way $I \times J \times K$ tables, sets of odds ratios in the partial tables describe the conditional association. Homogeneous XY association means that any conditional odds ratio formed using two categories of X and two categories of Y is the same at each category of Z.

2.4.2 Summary Measures of Association

An alternative way to describe association uses a single summary index. We discuss this first for nominal variables and then ordinal variables. The most interpretable indices for nominal variables have the same structure as R-squared for interval variables. It and the more general intraclass correlation coefficient and correlation ratio (Kendall and Stuart 1979) describe the proportional reduction in variance from the marginal distribution of the response Y to the conditional distributions of Y given an explanatory variable X.

Let $V(Y)$ denote a measure of variation for the marginal distribution $\{\pi_{+j}\}$ of Y, and let $V(Y|i)$ denote this measure computed for the conditional distribution $\{\pi_{1|i}, \ldots, \pi_{J|i}\}$ of Y at the ith setting of X. A proportional reduction in variation measure has the form

$$\frac{V(Y) - E[V(Y|X)]}{V(Y)}, \tag{2.12}$$

where $E[V(Y|X)]$ is the expectation of the conditional variation taken with respect to the distribution of X. For the marginal distribution $\{\pi_{i+}\}$ of X, $E[V(Y|X)] = \sum_i \pi_{i+} V(Y|i)$.

For a nominal response, Theil (1970) proposed an index using the variation measure $V(Y) = \Sigma \pi_{+j} \log \pi_{+j}$, called the *entropy*. For contingency tables, the proportional reduction in entropy equals

$$U = -\frac{\Sigma_i \Sigma_j \pi_{ij} \log(\pi_{ij}/\pi_{i+}\pi_{+j})}{\Sigma_j \pi_{+j} \log \pi_{+j}}, \tag{2.13}$$

called the *uncertainty coefficient*. This measure is well defined when more than one $\pi_{+j} > 0$. It takes value between 0 and 1: $U = 0$ is equivalent to independence of X and Y; $U = 1$ is equivalent to a lack of conditional variation, in the sense that for each i, $\pi_{j|i} = 1$ for some j.

Various measures of form (2.12) describe association in $I \times J$ tables (e.g., Problems 2.38 and 2.39). A difficulty with them is developing intuition for how large a value constitutes a strong association. What does it mean, for instance, to say that there is a 30% reduction in entropy? Summary measures seem easier to interpret and more useful when both classifications are ordinal, as discussed next.

2.4.3 Ordinal Trends: Concordant and Discordant Pairs

In Table 2.8 the variables are income and job satisfaction, measured for the black males in a national (U.S.) sample. Both classifications are ordinal, job satisfaction with the categories very dissatisfied (VD), little dissatisfied (LD), moderately satisfied (MS), and very satisfied (VS).

When X and Y are ordinal, a monotone trend association is common. As the level of X increases, responses on Y tend to increase toward higher levels, or responses on Y tend to decrease toward lower levels. For instance, perhaps job satisfaction tends to increase as income does. A single parameter can describe this trend. Measures analogous to the correlation describe the degree to which the relationship is monotone. Some measures are based on classifying each pair of subjects as concordant or discordant. A pair is *concordant* if the subject ranked higher on X also ranks higher on Y. The

TABLE 2.8 Cross-Classification of Job Satisfaction by Income

Income (dollars)	Job Satisfaction			
	Very Dissatisfied	Little Dissatisfied	Moderately Satisfied	Very Satisfied
< 15,000	1	3	10	6
15,000–25,000	2	3	10	7
25,000–40,000	1	6	14	12
> 40,000	0	1	9	11

Source: 1996 General Social Survey, National Opinion Research Center.

pair is *discordant* if the subject ranking higher on X ranks lower on Y. The pair is *tied* if the subjects have the same classification on X and/or Y.

We illustrate for Table 2.8. Consider a pair of subjects, one in the cell (< 15, VD) and the other in the cell (15–25, LD). This pair is concordant, since the second subject ranks higher than the first both on income and on job satisfaction. The subject in cell (< 15, VD) forms concordant pairs when matched with each of the three subjects classified (15–25, LD), so these two cells provide $1 \times 3 = 3$ concordant pairs. The subject in the cell (< 15, VD) is also part of a concordant pair when matched with each of the other $(10 + 7 + 6 + 14 + 12 + 1 + 9 + 11)$ subjects ranked higher on both variables. Similarly, the three subjects in the (< 15, LD) cell are part of concordant pairs when matched with the $(10 + 7 + 14 + 12 + 9 + 11)$ subjects ranked higher on both variables.

The total number of concordant pairs, denoted by C, equals

$$\begin{aligned} C = {} & 1(3 + 10 + 7 + 6 + 14 + 12 + 1 + 9 + 11) \\ & + 3(10 + 7 + 14 + 12 + 9 + 11) + 10(7 + 12 + 11) \\ & + 2(6 + 14 + 12 + 1 + 9 + 11) + 3(14 + 12 + 9 + 11) \\ & + 10(12 + 11) + 1(1 + 9 + 11) + 6(9 + 11) + 14(11) = 1331. \end{aligned}$$

The total number of discordant pairs of observations is

$$D = 3(2 + 1 + 0) + 10(2 + 3 + 1 + 6 + 0 + 1) + \cdots + 12(0 + 1 + 9) = 849.$$

In this example, $C > D$, suggesting a tendency for low income to occur with low job satisfaction and high income with high job satisfaction.

Consider two independent observations from a joint probability distribution $\{\pi_{ij}\}$. For that pair, the probabilities of concordance and discordance are

$$\Pi_c = 2\sum_i \sum_j \pi_{ij}\Big(\sum_{h>i} \sum_{k>j} \pi_{hk}\Big), \qquad \Pi_d = 2\sum_i \sum_j \pi_{ij}\Big(\sum_{h>i} \sum_{k<j} \pi_{hk}\Big).$$

Here i and j are fixed in the inner summations, and the factor of 2 occurs because the first observation could be in cell (i, j) and the second in cell (h, k), or vice versa. Several association measures for ordinal variables utilize the difference $\Pi_c - \Pi_d$.

2.4.4 Ordinal Measure of Association: Gamma

Given that a pair is untied on both variables, $\Pi_c/(\Pi_c + \Pi_d)$ is the probability of concordance and $\Pi_d/(\Pi_c + \Pi_d)$ is the probability of discordance. The

difference between these probabilities is

$$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d}, \tag{2.14}$$

called *gamma* (Goodman and Kruskal 1954). The sample version is $\hat{\gamma} = (C - D)/(C + D)$.

Like the correlation, gamma treats the variables symmetrically—it is unnecessary to identify one classification as a response variable. Also like the correlation, gamma has range $-1 \leq \gamma \leq 1$. A reversal in the category orderings of one variable causes a change in the sign of γ. Whereas the absolute value of the correlation is 1 when the relationship between X and Y is perfectly linear, only monotonicity is required for $|\gamma| = 1$, with $\gamma = 1$ if $\Pi_d = 0$ and $\gamma = -1$ if $\Pi_c = 0$. Independence implies that $\gamma = 0$, but the converse is not true. For instance, a U-shaped joint distribution can have $\Pi_c = \Pi_d$ and hence $\gamma = 0$.

2.4.5 Gamma for Job Satisfaction Example

For Table 2.8, $C = 1331$ and $D = 849$. Hence,

$$\hat{\gamma} = (1331 - 849)/(1331 + 849) = 0.221.$$

Only a weak tendency exists for job satisfaction to increase as income increases. Of the untied pairs, the proportion of concordant pairs is 0.221 higher than the proportion of discordant pairs.

NOTES

Section 2.2: Comparing Two Proportions

2.1. Breslow (1996) presented an interesting overview of the development of methods for case–control studies.

2.2. For 2×2 tables, Edwards (1963) showed that functions of the odds ratio are the only statistics that are invariant both to row–column interchange and to multiplication within rows or within columns by a constant. For $I \times J$ tables, Altham (1970) gave related results. Yule (1912, p. 587) had argued that multiplicative invariance is a desirable property for measures of association, especially when proportions sampled in various marginal categories are arbitrary. Goodman (2000) showed five ways of viewing association in a 2×2 table and proposed a general measure that includes all five.

Section 2.3: Partial Association in Stratified 2×2 Tables

2.3. Paik (1985) proposed circle diagrams of type Figure 2.2 to summarize three-way tables. Friendly (2000) discussed graphical presentation of categorical data. For more on Simpson's paradox and when it can happen, see Blyth (1972), Davis (1989), Dong (1998),

Samuels (1993), and Simpson (1951). Good and Mittal (1989) extended it to an *amalgamation paradox*, whereby a marginal measure is greater than the maximum or less than the minimum of the partial table measures.

Section 2.4: Extensions for $I \times J$ Tables

2.4. For continuous variables, samples can be fully ranked (i.e., no ties occur), so $C + D = \binom{n}{2}$ and $\hat{\gamma} = (C - D)/\binom{n}{2}$. This is *Kendall's tau*. Agresti (1984, Chaps. 9 and 10) and Kruskal (1958) surveyed ordinal measures of association. These also apply when one variable is ordinal and the other is binary. When Y is ordinal and X is nominal with $I > 2$, no measure presented in Section 2.4 is very helpful. Ordinal modeling approaches (Section 7.2) use a parameter for each category of X; comparing parameters compares the ordinal response for pairs of categories of X.

PROBLEMS

Applications

2.1 An article in the *New York Times* (Feb. 17, 1999) about the PSA blood test for detecting prostate cancer stated: "The test fails to detect prostate cancer in 1 in 4 men who have the disease (false-negative results), and as many as two-thirds of the men tested receive false-positive results." Let $C(\overline{C})$ denote the event of having (not having) prostate cancer, and let $+$ $(-)$ denote a positive (negative) test result. Which is true: $P(-\,|\,C) = \frac{1}{4}$ or $P(C\,|\,-) = \frac{1}{4}$? $P(\overline{C}\,|\,+) = \frac{2}{3}$ or $P(+\,|\,\overline{C}) = \frac{2}{3}$? Determine the sensitivity and specificity.

2.2 A diagnostic test has sensitivity = specificity = 0.80. Find the odds ratio between true disease status and the diagnostic test result.

2.3 Table 2.9 is based on records of accidents in 1988 compiled by the Department of Highway Safety and Motor Vehicles in Florida. Identify the response variable, and find and interpret the difference of proportions, relative risk, and odds ratio. Why are the relative risk and odds ratio approximately equal?

TABLE 2.9 Data for Problem 2.3

Safety Equipment in Use	Injury	
	Fatal	Nonfatal
None	1601	162,527
Seat belt	510	412,368

Source: Florida Department of Highway Safety and Motor Vehicles.

2.4 Consider the following two studies reported in the *New York Times*.

a. A British study reported (Dec. 3, 1998) that of smokers who get lung cancer, "women were 1.7 times more vulnerable than men to get small-cell lung cancer." Is 1.7 the odds ratio or the relative risk?

b. A National Cancer Institute study about tamoxifen and breast cancer reported (Apr. 7, 1998) that the women taking the drug were 45% less likely to experience invasive breast cancer then were women taking placebo. Find the relative risk for **(i)** those taking the drug compared to those taking placebo, and **(ii)** those taking placebo compared to those taking the drug.

2.5 A study (E. G. Krug et al., *Internat. J. Epidemiol.*, **27**: 214–221, 1998) reported that the number of gun-related deaths per 100,000 people in 1994 was 14.24 in the United States, 4.31 in Canada, 2.65 in Australia, 1.24 in Germany, and 0.41 in England and Wales. Use the relative risk to compare the United States with the other countries. Interpret.

2.6 A newspaper article preceding the 1994 World Cup semifinal match between Italy and Bulgaria stated that "Italy is favored 10–11 to beat Bulgaria, which is rated at 10–3 to reach the final." Suppose that this means that the odds that Italy wins are $\frac{11}{10}$ and the odds that Bulgaria wins are $\frac{3}{10}$. Find the probability that each team wins, and comment.

2.7 In the United States, the estimated annual probability that a woman over the age of 35 dies of lung cancer equals 0.001304 for current smokers and 0.000121 for nonsmokers (M. Pagano and K. Gauvreau, *Principles of Biostatistics*, Duxbury Press, Pacific Grove, CA. 1993, p. 134).

a. Find and interpret the difference of proportions and the relative risk. Which measure is more informative for these data? Why?

b. Find and interpret the odds ratio. Explain why the relative risk and odds ratio take similar values.

2.8 For adults who sailed on the *Titanic* on its fateful voyage, the odds ratio between gender (female, male) and survival (yes, no) was 11.4. (For data, see R. J. M. Dawson, *J. Statist. Ed.* **3**, 1995.)

a. What is wrong with the interpretation, "The probability of survival for females was 11.4 times that for males"? Give the correct interpretation. When would the quoted interpretation be approximately correct?

b. The odds of survival for females equaled 2.9. For each gender, find the proportion who survived.

2.9 In an article about crime in the United States, *Newsweek* (Jan. 10, 1994) quoted FBI statistics for 1992 stating that of blacks slain, 94% were slain by blacks, and of whites slain, 83% were slain by whites. Let Y = race of victim and X = race of murderer. Which conditional distribution do these statistics refer to, $Y|X$, or $X|Y$? What additional information would you need to estimate the probability that the victim was white given that a murderer was white? Find and interpret the odds ratio.

2.10 A research study estimated that under a certain condition, the probability that a subject would be referred for heart catheterization was 0.906 for whites and 0.847 for blacks.

a. A press release about the study stated that the odds of referral for cardiac catheterization for blacks are 60% of the odds for whites. Explain how they obtained 60% (more accurately, 57%).

b. An Associated Press story later described the study and said "Doctors were only 60% as likely to order cardiac catheterization for blacks as for whites." Explain what is wrong with this interpretation. Give the correct percentage for this interpretation. (In stating results to the general public, it is better to use the relative risk than the odds ratio. It is simpler to understand and less likely to be misinterpreted. For details, see *New Engl. J. Med.* **341**: 279–283, 1999.)

2.11 A 20-year cohort study of British male physicians (R. Doll and R. Peto, *British Med. J.* **2**: 1525–1536, 1976) noted that the proportion per year who died from lung cancer was 0.00140 for cigarette smokers and 0.00010 for nonsmokers. The proportion who died from coronary heart disease was 0.00669 for smokers and 0.00413 for nonsmokers.

a. Describe the association of smoking with each of lung cancer and heart disease, using the difference of proportions, relative risk, and odds ratio. Interpret.

b. Which response is more strongly related to cigarette smoking, in terms of the reduction in number of deaths that would occur with elimination of cigarettes? Explain.

2.12 Table 2.10 refers to applicants to graduate school at the University of California at Berkeley, for fall 1973. It presents admissions decisions by gender of applicant for the six largest graduate departments. Denote the three variables by A = whether admitted, G = gender, and D = department. Find the sample AG conditional odds ratios and the marginal odds ratio. Interpret, and explain why they give such different indications of the AG association.

TABLE 2.10 Data for Problem 2.12

	Whether Admitted			
	Male		Female	
Department	Yes	No	Yes	No
A	512	313	89	19
B	353	207	17	8
C	120	205	202	391
D	138	279	131	244
E	53	138	94	299
F	22	351	24	317
Total	1198	1493	557	1278

Source: Data from Freedman et al. (1978, p.14). See also P. Bickel et al., *Science* **187**: 398–403 (1975).

2.13 State three "real-world" variables X, Y, and Z for which you expect a marginal association between X and Y but conditional independence controlling for Z.

2.14 Based on 1987 murder rates in the United States, an Associated Press story reported that the probability that a newborn child has of eventually being a murder victim is 0.0263 for nonwhite males, 0.0049 for white males, 0.0072 for nonwhite females, and 0.0023 for white females.

a. Find the conditional odds ratios between race and whether a murder victim, given the gender. Interpret. Do these variables exhibit homogeneous association?

b. Half the newborns are of each gender, for each race. Find the marginal odds ratio between race and whether a murder victim.

2.15 At each age level, the death rate is higher in South Carolina than in Maine, but overall, the death rate is higher in Maine. Explain how this could be possible. (For data, see H. Wainer, *Chance* **12**: 44, 1999.)

2.16 A study of the death penalty for cases in Kentucky between 1976 and 1991 (T. Keil and G. Vito, *Amer. J. Criminal Justice* **20**: 17–36, 1995) indicated that the defendant received the death penalty in 8% of the 391 cases in which a white killed a white, in 2% of the 108 cases in which a black killed a black, in 12% of the 57 cases in which a black killed a white, and in 0% of the 18 cases in which a white killed a black. Form the three-way contingency table, obtain the conditional odds ratios between the defendant's race and the death penalty verdict, interpret those associations, study whether Simpson's paradox occurs,

and explain why the marginal association is so different from the conditional associations.

2.17 An estimated odds ratio for adult females between the presence of squamous cell carcinoma (yes, no) and smoking behavior (smoker, nonsmoker) equals 11.7 when the smoker category has subjects whose smoking level s is $0 < s < 20$ cigarettes per day; it is 26.1 for smokers with $s \geq 20$ cigarettes per day (R. C. Brownson et al., *Epidemiology* **3**: 61–64, 1992). Show that the estimated odds ratio between carcinoma (yes, no) and the smoking levels ($s \geq 20$, $0 < s < 20$) equals 2.2.

2.18 Table 2.11 refers to a retrospective study of lung cancer and tobacco smoking among patients in several English hospitals. The table compares male lung cancer patients with control patients having other diseases, according to the average number of cigarettes smoked daily over a 10-year period preceding the onset of the disease.

a. Find the sample odds of lung cancer at each smoking level and the five odds ratios that pair each level of smoking with no smoking. As smoking increases, is there a trend? Interpret.

b. If the log odds of lung cancer is linearly related to smoking level, the log odds in row i satisfies $\log(\text{odds}_i) = \alpha + \beta i$. Show that this implies that the local odds ratios are identical.

c. Using these data, can you estimate the probability of lung cancer at each level of smoking? Are the estimated odds ratios in part (a) meaningful? Explain.

d. Show that the disease groups are *stochastically ordered* with respect to their distributions on smoking of cigarettes (see Problem 2.34 and Section 7.3.4). Interpret.

TABLE 2.11 Data for Problem 2.18

Daily Average Number of Cigarettes	Disease Group	
	Lung Cancer Patients	Control Patients
None	7	61
< 5	55	129
5–14	489	570
15–24	475	431
25–49	293	154
50 +	38	12

Source: Reprinted with permission from R. Doll and A. B. Hill, *British Med. J.* **2**: 1271–1286 (1952).

TABLE 2.12 Data for Problem 2.19

	Wife's Rating of Sexual Fun			
Husband's Rating	Never or Occasionally	Fairly Often	Very Often	Almost Always
Never or occasionally	7	7	2	3
Fairly often	2	8	3	7
Very often	1	5	4	9
Almost always	2	8	9	14

Source: Reprinted with permission from Hout et al. (1987).

2.19 Table 2.12 summarizes responses of 91 married couples in Arizona to a question about how often sex is fun. Find and interpret a measure of association between wife's response and husband's response.

2.20 Table 2.13 is from an early study on the death penalty in Florida. Analyze these data and show that Simpson's paradox occurs.

TABLE 2.13 Data for Problem 2.20

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

Source: Reprinted with permission from M. L. Radelet, *Amer. Sociol. Rev.* **46**: 918–927 (1981)

Theory and Methods

2.21 For a diagnostic test of a certain disease, π_1 denotes the probability that the diagnosis is positive given that a subject has the disease, and π_2 denotes the probability that the diagnosis is positive given that a subject does not have it. Let ρ denote the probability that a subject does have the disease.

a. Given that the diagnosis is positive, show that the probability that a subject does have the disease is

$$\pi_1 \rho / \left[\pi_1 \rho + \pi_2 (1 - \rho) \right].$$

b. Suppose that a diagnostic test for HIV+ status has both sensitivity and specificity equal to 0.95, and $\rho = 0.005$. Find the probability that a subject is truly HIV+ , given that the diagnostic test is positive. To better understand this answer, find the joint probabilities relating diagnosis to actual disease status, and discuss their relative sizes.

2.22 Binomial parameters for two groups are graphed, with π_1 on the horizontal axis and π_2 on the vertical axis. Plot the locus of points for a 2×2 table having (**a**) relative risk $= 0.5$, (**b**) odds ratio $= 0.5$, and (**c**) difference of proportions $= -0.5$.

2.23 Let D denote having a certain disease and E denote having exposure to a certain risk factor. The *attributable risk* (AR) is the proportion of disease cases attributable to that exposure (see Benichou 1998).

a. Let $P(\bar{E}) = 1 - P(E)$. Explain why

$$\text{AR} = \left[P(D) - P(D\,|\,\bar{E})\right]/P(D).$$

b. Show that AR relates to the relative risk RR by

$$\text{AR} = \left[P(E)(\text{RR} - 1)\right]/\left[1 + P(E)(\text{RR} - 1)\right].$$

2.24 For a 2×2 table of counts $\{n_{ij}\}$, show that the odds ratio is invariant to (**a**) interchanging rows with columns, and (**b**) multiplication of cell counts within rows or within columns by $c \neq 0$. Show that the difference of proportions and the relative risk do not have these properties.

2.25 For given π_1 and π_2, show that the relative risk cannot be farther than the odds ratio from their independence value of 1.0.

2.26 Explain why for three events E_1, E_2, and E_3 and their complements, it is possible that $P(E_1\,|\,E_2) > P(E_1\,|\,\bar{E}_2)$ even if both $P(E_1\,|\,E_2E_3) < P(E_1\,|\,\bar{E}_2E_3)$ and $P(E_1\,|\,E_2\bar{E}_3) < P(E_1\,|\,\bar{E}_2\bar{E}_3)$. (*Hint:* Use Simpson's paradox for a three-way table.)

2.27 Let $\pi_{ij\,|\,k} = P(X = i, Y = j\,|\,Z = k)$. Explain why XY conditional independence is

$$\pi_{ij|k} = \pi_{i+|k}\pi_{+j|k} \quad \text{for all } i \text{ and } j \text{ and } k.$$

2.28 For a $2 \times 2 \times 2$ table, show that homogeneous association is a symmetric property, by showing that equal XY conditional odds ratios is equivalent to equal YZ conditional odds ratios.

2.29 Smith and Jones are baseball players. Smith has a higher batting average than Jones in each of K years. Is is possible that for the combined data from the K years, Jones has the higher batting average? Explain, using an example to illustrate.

2.30 When X and Y are conditionally dependent at each level of Z yet marginally independent, Z is called a *suppressor variable*. Specify joint probabilities for a $2 \times 2 \times 2$ table to show that this can happen (**a**) when there is homogeneous association, and (**b**) when the association has opposite direction in the partial tables.

2.31 Show that the $\{\alpha_{ij}\}$ in (2.11) determine (**a**) all $\binom{I}{2}\binom{J}{2}$ odds ratios formed from pairs of rows and pairs of columns, (**b**) all $\{\theta_{ij}\}$ in (2.10), and vice versa.

2.32 Refer to Problem 2.31. When all rows and columns have positive probability, show that independence is equivalent to all $\{\alpha_{ij} = 1\}$.

2.33 For $I \times J$ contingency tables, explain why the variables are independent when the $(I-1)(J-1)$ differences $\pi_{j|i} - \pi_{j|I} = 0$, $i = 1, \ldots, I-1$, $j = 1, \ldots, J-1$.

2.34 A $2 \times J$ table has ordinal response. Let $F_{j|i} = \pi_{1|i} + \cdots + \pi_{j|i}$. When $F_{j|2} \leq F_{j|1}$ for $j = 1, \ldots, J$, the conditional distribution in row 2 is *stochastically higher* than the one in row 1. Consider the *cumulative odds ratios*

$$\theta_j = \frac{F_{j|1}/(1 - F_{j|1})}{F_{j|2}/(1 - F_{j|2})}, \qquad j = 1, \ldots, J-1.$$

- **a.** Show that $\log \theta_j \geq 0$ for all j is equivalent to row 2 being stochastically higher than row 1. Explain why row 2 is then more likely than row 1 to have observations at the high end of the ordinal scale.
- **b.** If all local log odds ratios are nonnegative, $\log \theta_j \geq 0$ for $1 \leq j \leq J-1$ (Lehmann 1966). Show by counterexample that the converse is not true.

2.35 Suppose that $\{Y_{ij}\}$ are independent Poisson variates with means $\{\mu_{ij}\}$. Show that $P(Y_{ij} = n_{ij})$ for all i, j, conditional on $\{Y_{i+} = n_i\}$, satisfy independent multinomial sampling [i.e., the product of (2.2) for all i] within the rows.

2.36 For 2×2 tables, Yule (1900, 1912) introduced

$$Q = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{\pi_{11}\pi_{22} + \pi_{12}\pi_{21}},$$

which he labeled Q in honor of the Belgian statistician Quetelet. It is now called *Yule's Q*.

a. Show that for 2×2 tables, Goodman and Kruskal's $\gamma = Q$.

b. Show that Q falls between -1 and 1.

c. State conditions under which $Q = -1$ or $Q = 1$.

d. Show that Q relates to the odds ratio by $Q = (\theta - 1)/(\theta + 1)$, a monotone transformation of θ from the $[0, \infty]$ scale onto the $[-1, +1]$ scale.

2.37 When X and Y are ordinal with counts $\{n_{ij}\}$:

a. Explain why the $\binom{n}{2}$ pairs of observations partition into $C + D + T_X + T_Y - T_{XY}$, where $T_X = \Sigma n_{i+}(n_{i+} - 1)/2$ pairs are tied on X, T_Y pairs are tied on Y, and T_{XY} pairs are tied on X and Y.

b. For each ordered pair of observations (X_a, Y_a) and (X_b, Y_b), let $X_{ab} = \text{sign}(X_a - X_b)$ and $Y_{ab} = \text{sign}(Y_a - Y_b)$. Show that the sample correlation for the $n(n-1)$ distinct (X_{ab}, Y_{ab}) pairs is

$$\tau_b = \frac{C - D}{\left\{\left[\binom{n}{2} - T_X\right]\left[\binom{n}{2} - T_Y\right]\right\}^{1/2}}.$$

This ordinal measure, called *Kendall's tau-b* (Kendall 1945), is less sensitive than gamma to the choice of response categories.

c. Let $d = (C - D)/\left[\binom{n}{2} - T_X\right]$. Explain why d is the difference between the proportions of concordant and discordant pairs out of those pairs untied on X (Somers 1962). (For 2×2 tables, d equals the difference of proportions, and tau-b equals the correlation between X and Y.)

2.38 Goodman and Kruskal (1954) proposed an association measure (tau) for nominal variables based on variation measure

$$V(Y) = \sum \pi_{+j}(1 - \pi_{+j}) = 1 - \sum \pi_{+j}^2.$$

a. Show $V(Y)$ is the probability that two independent observations on Y fall in different categories (called the *Gini concentration index*).

Show that $V(Y) = 0$ when $\pi_{+j} = 1$ for some j and $V(Y)$ takes maximum value of $(J - 1)/J$ when $\pi_{+j} = 1/J$ for all j.

b. For the proportional reduction in variation, show that $E[V(Y|X)] = 1 - \sum_i \sum_j \pi_{ij}^2 / \pi_{i+}$. [The resulting measure (2.12) is called the *concentration coefficient*. Like U, $\tau = 0$ is equivalent to independence. Haberman (1982) presented generalized concentration and uncertainty coefficients.]

2.39 The measure of association *lambda* for nominal variables (Goodman and Kruskal 1954) has $V(Y) = 1 - \max\{\pi_{+j}\}$ and $V(Y|i) = 1 - \max_j\{\pi_{j|i}\}$. Interpret lambda as a proportional reduction in prediction error for predictions which select the response category that is most likely. Show that independence implies $\lambda = 0$ but that the converse is not true.

Categorical Data Analysis, Second Edition. Alan Agresti

ISBN: 0-471-36093-7

CHAPTER 3

Inference for Contingency Tables

In this chapter we introduce inferential methods for contingency tables. Many of these methods also play a vital role in analyses of later chapters for which categorical data need not have contingency table form. The methods assume Poisson, multinomial, or independent binomial sampling.

In Section 3.1 we present confidence intervals for measures of association for 2×2 tables such as the odds ratio. Section 3.2 covers chi-squared tests of the hypothesis of independence between two categorical variables. Like any significance test, these have limited usefulness. In Section 3.3 we show how to follow-up the test using residuals or the partitioning property of chi-squared to extract components that describe the evidence about the association. In Section 3.4 we present more powerful inference applicable with ordered categories. The methods of Sections 3.1 through 3.4 assume large samples. In Sections 3.5 and 3.6 we introduce small-sample methods.

3.1 CONFIDENCE INTERVALS FOR ASSOCIATION PARAMETERS

The accuracy of estimators of association parameters is characterized by standard errors of their sampling distributions. In this section we present large-sample standard errors and confidence intervals.

3.1.1 Interval Estimation of Odds Ratios

The sample odds ratio $\hat{\theta} = n_{11}n_{22}/n_{12}n_{21}$ for a 2×2 table equals 0 or ∞ if any $n_{ij} = 0$, and it is undefined if both entries in a row or column are zero. Since these outcomes have positive probabilities, the expected value and variance of $\hat{\theta}$ and $\log \hat{\theta}$ do not exist. (In fact, this is also true for ML estimators of model parameters presented in later chapters.) In terms of bias and mean-squared error, Gart and Zweiful (1967) and Haldane (1956)

showed that the amended estimators

$$\tilde{\theta} = \frac{(n_{11} + 0.5)(n_{22} + 0.5)}{(n_{12} + 0.5)(n_{21} + 0.5)}$$

and $\log \tilde{\theta}$ behave well (Problem 14.4).

The estimators $\hat{\theta}$ and $\tilde{\theta}$ have the same asymptotic normal distribution around θ. Unless n is quite large, however, their distributions are highly skewed. When $\theta = 1$, for instance, $\hat{\theta}$ cannot be much smaller than θ (since $\hat{\theta} \geq 0$), but it could be much larger with nonnegligible probability. The log transform, having an additive rather than multiplicative structure, converges more rapidly to normality. An estimated standard error for $\log \hat{\theta}$ is

$$\hat{\sigma}(\log \hat{\theta}) = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \right)^{1/2}. \tag{3.1}$$

We derive this formula in Section 3.1.7.

By the large-sample normality of $\log \hat{\theta}$,

$$\log \hat{\theta} \pm z_{\alpha/2}\, \hat{\sigma}(\log \hat{\theta}) \tag{3.2}$$

is a Wald confidence interval for $\log \theta$. Exponentiating (taking antilogs of) its endpoints provides a confidence interval for θ. Woolf (1955) proposed this interval. It works quite well, usually being a bit conservative (i.e., actual coverage probability higher than the nominal level).

When $\hat{\theta} = 0$ or ∞, Woolf's interval does not exist. When $\hat{\theta} = 0$, one should take 0 as the lower limit and when $\hat{\theta} = \infty$, one should take ∞ as the upper limit. The other bound can use the Woolf formula following some adjustment, such as Gart's (1966), which replaces $\{n_{ij}\}$ by $\{n_{ij} + 0.5\}$ in the estimator and standard error. A less ad hoc approach forms the interval by inverting score tests (Cornfield 1956) or likelihood-ratio tests for θ, as we discuss in Section 3.1.8.

3.1.2 Aspirin and Myocardial Infarction Example

We illustrate inference for the odds ratio with Table 3.1 based on a Swedish study of the association between aspirin use and myocardial infarction similar to that described in Section 2.2.5. The study randomly assigned 1360 patients who had already suffered a stroke to an aspirin treatment (one low-dose tablet a day) or to a placebo treatment. Table 3.1 reports the number of deaths due to myocardial infarction during a follow-up period of about 3 years.

The sample odds ratio $\hat{\theta} = 1.56$ is close to $\tilde{\theta} = 1.55$, since no cell count is especially small. The standard error (3.1) of $\log \hat{\theta} = 0.445$ is $\hat{\sigma}(\log \hat{\theta}) = 0.307$.

TABLE 3.1 Swedish Study on Aspirin Use and Myocardial Infarction

	Myocardial Infarction		
	Yes	No	Total
Placebo	28	656	684
Aspirin	18	658	676

Source: Based on results described in *Lancet* **338**: 1345–1349 (1991).

A 95% confidence interval for $\log\theta$ in the population this sample represents is $0.445 \pm 1.96(0.307)$, or $(-0.157, 1.047)$. The corresponding interval for θ is $[\exp(-0.157), \exp(1.047)]$, or $(0.85, 2.85)$. The estimate of the true odds ratio is rather imprecise.

Since the confidence interval for θ contains 1.0, it is plausible that the true odds of death due to myocardial infarction are equal for aspirin and placebo. If there truly is a beneficial effect of aspirin but the odds ratio is not large, it may require a large sample size to show that benefit because of the relatively small number of myocardial infarction cases (Problem 3.21).

3.1.3 Interval Estimation of Difference of Proportions

The difference of proportions and the relative risk compare conditional distributions of a response variable for two groups. For these measures, we treat the samples as independent binomials. For group i, y_i has a binomial distribution with sample size n_i and a probability π_i of a "success" response.

The sample proportion $\hat{\pi}_i = y_i/n_i$ has expectation π_i and variance $\pi_i(1-\pi_i)/n_i$. Since $\hat{\pi}_1$ and $\hat{\pi}_2$ are independent, their difference has

$$E(\hat{\pi}_1 - \hat{\pi}_2) = \pi_1 - \pi_2$$

and standard error

$$\sigma(\hat{\pi}_1 - \hat{\pi}_2) = \left[\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right]^{1/2}. \tag{3.3}$$

The estimate $\hat{\sigma}(\hat{\pi}_1 - \hat{\pi}_2)$ uses formula (3.3) with π_i replaced by $\hat{\pi}_i$. Then

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\alpha/2}\hat{\sigma}(\hat{\pi}_1 - \hat{\pi}_2) \tag{3.4}$$

is a Wald confidence interval for $\pi_1 - \pi_2$. Like the Wald interval (1.13) for a single proportion, it usually has true coverage probability less than the nominal confidence coefficient, especially when π_1 and π_2 are near 0 or 1. More complex but better methods are cited in Section 3.1.8, Note 3.2, and Problem 3.23.

3.1.4 Interval Estimation of Relative Risk

The sample relative risk is $r = \hat{\pi}_1/\hat{\pi}_2$. Like the odds ratio, it converges to normality faster on the log scale. The asymptotic standard error of $\log r$ is

$$\sigma(\log r) = \left(\frac{1-\pi_1}{\pi_1 n_1} + \frac{1-\pi_2}{\pi_2 n_2}\right)^{1/2}. \tag{3.5}$$

The Wald interval exponentiates endpoints of $\log r \pm z_{\alpha/2}\,\hat{\sigma}(\log r)$. It works well but can be somewhat conservative. We discuss an alternative method in Section 3.1.8.

For Table 3.1, the sample proportion of myocardial infarction deaths was 0.0409 for subjects taking placebo and 0.0266 for subjects taking aspirin. The sample relative risk is $0.0409/0.0266 = 1.54$. The 95% confidence interval for the log relative risk of $\log(1.54) \pm 1.96(0.297)$ translates to (0.86, 2.75) for the relative risk. We infer that the death rate for those taking placebo was between 0.86 and 2.75 times that for those taking aspirin. The Wald 95% confidence interval for $\pi_1 - \pi_2$ is $0.014 \pm 1.96(0.0098)$ or $(-0.005, 0.033)$. According to either measure, substantial public health benefits could result from taking aspirin, but no effect or a slight negative effect are also plausible. Results for the larger study described in Section 2.2.5 do show a benefit.

3.1.5 Deriving Standard Errors with the Delta Method*

A simple and useful method exists of deriving standard errors for large-sample inferences. Let T_n denote a statistic that is asymptotically normally distributed about a parameter θ, the subscript n expressing its dependence on sample size. Suppose that an estimator is a function $g(T_n)$ of T_n. Then, under mild conditions, $g(T_n)$ itself has a large-sample normal distribution. The standard error depends on how fast $g(t)$ changes for t near θ.

Specifically, for large n, suppose that T_n is normally distributed about θ with standard error $\sigma/\sqrt{n}$. That is, as $n \to \infty$, the cdf of $\sqrt{n}\,(T_n - \theta)$ converges to the cdf of a normal random variable with mean 0 and variance σ^2. This limiting behavior is an example of *convergence in distribution*, denoted by

$$\sqrt{n}\,(T_n - \theta) \xrightarrow{d} N(0, \sigma^2).$$

Let g be a function that is at least twice differentiable at θ. Using the Taylor series expansion for $g(t)$ in a neighborhood of $t = \theta$, in Section 14.1.2 we show

$$\sqrt{n}\,[g(T_n) - g(\theta)] \approx \sqrt{n}\,(T_n - \theta)g'(\theta)$$

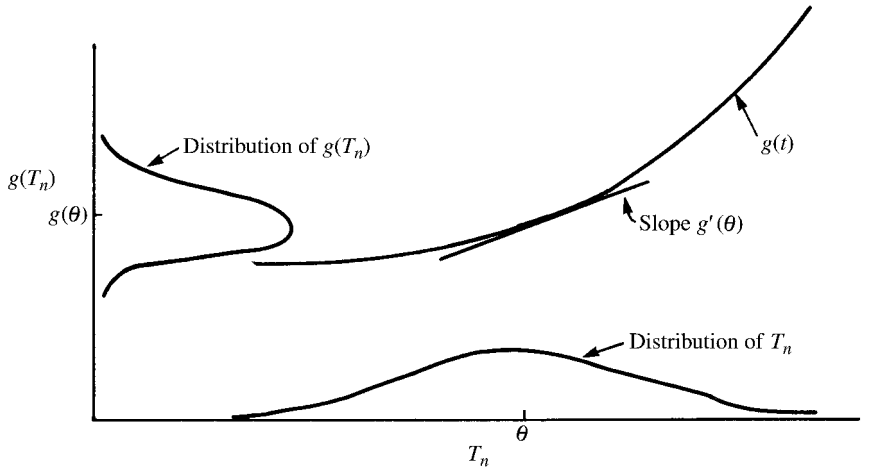


FIGURE 3.1 Depiction of delta method.

for large n, where $g'(\theta) = \partial g/\partial t$ evaluated at $t = \theta$. Recall if a variate $Y \sim N(0, \sigma^2)$, then $cY \sim N(0, c^2\sigma^2)$. Thus,

$$\sqrt{n}\,[g(T_n) - g(\theta)] \xrightarrow{d} N\big(0, [g'(\theta)]^2\sigma^2\big). \tag{3.6}$$

In other words, $g(T_n)$ is approximately normal around $g(\theta)$ with variance $[g'(\theta)]^2\sigma^2/n$.

Figure 3.1 portrays this result. Locally around θ, $g(t)$ is approximately linear, with slope $g'(\theta)$. Then $g(T_n)$ is approximately normal, since linear transformations of normal random variables are themselves normal. The dispersion of $g(T_n)$ values about $g(\theta)$ is about $|g'(\theta)|$ times the dispersion of T_n values about θ. If the slope of g at θ is $\frac{1}{2}$, then g maps a region of T_n values into a region of $g(T_n)$ values only about half as wide.

Result (3.6) is called the *delta method*. Since $g'(\theta)$ and $\sigma^2 = \sigma^2(\theta)$ usually depend on the unknown parameter θ, the asymptotic variance is unknown. Confidence intervals and tests substitute T_n for θ and use the result that $\sqrt{n}\,[g(T_n) - g(\theta)]/\,|g'(T_n)|\,\sigma(T_n)$ is asymptotically standard normal. For instance,

$$g(T_n) \pm 1.96|g'(T_n)|\sigma(T_n)/\sqrt{n}$$

is a large-sample Wald 95% confidence interval for $g(\theta)$.

3.1.6 Delta Method Applied to Sample Logit*

We illustrate the delta method for a function of the ML estimator $T_n = \hat{\pi} = y/n$ of the binomial parameter π, for y successes in n trials. Since $E(Y) = n\pi$ and $\text{var}(Y) = n\pi(1 - \pi)$, $E(\hat{\pi}) = \pi$ and $\text{var}(\hat{\pi}) = \pi(1 - \pi)/n$. Also, $\hat{\pi}$

has a large-sample normal distribution by the central limit theorem. So do many functions of $\hat{\pi}$.

The log odds function of $\hat{\pi}$,

$$g(\hat{\pi}) = \log[\hat{\pi}/(1 - \hat{\pi})],$$

is called the sample *logit*. Evaluated at π, its derivative equals $1/\pi(1 - \pi)$. By the delta method, the asymptotic variance of the sample logit is $\pi(1 - \pi)/n$ (the variance of $\hat{\pi}$) multiplied by the square of $[1/\pi(1 - \pi)]$. That is

$$\sqrt{n}\left(\log\frac{\hat{\pi}}{1-\hat{\pi}} - \log\frac{\pi}{1-\pi}\right) \xrightarrow{d} N\left(0, \frac{1}{\pi(1-\pi)}\right).$$

The asymptotic normality of $\hat{\pi}$ propagates to asymptotic normality of $\log[\hat{\pi}/(1 - \hat{\pi})]$.

The asymptotic variance is the variance of the normal distribution that approximates the true distribution, for large n. It is *not* an approximation for the variance of the true distribution. For $0 < \pi < 1$, the asymptotic variance $[n\pi(1 - \pi)]^{-1}$ of the sample logit is finite. By contrast, the true variance does not exist: Since $\hat{\pi} = 0$ or 1 with positive probability, the logit can equal $-\infty$ or ∞ with positive probability. The probability of an infinite logit converges to zero rapidly as n increases. For large n, the distribution of the sample logit looks essentially normal with mean $\log[\pi/(1 - \pi)]$ and standard deviation $[n\pi(1 - \pi)]^{-1/2}$. Thus, for the logit, the asymptotic variance actually has greater use than the true variance. Incidentally, related to this, the bootstrap is not helpful for approximating standard errors for many discrete measures, because it mimics the true rather than the more relevant asymptotic standard error.

3.1.7 Delta Method for Log Odds Ratio*

Standard errors for the log odds ratio and the log relative risk result from a multiparameter version of the delta method. Suppose that $\{n_i,\ i = 1, \ldots, c\}$ have a multinomial $(n, \{\pi_i\})$ distribution. The sample proportion $\hat{\pi}_i = n_i/n$ has mean and variance

$$E(\hat{\pi}_i) = \pi_i \quad \text{and} \quad \text{var}(\hat{\pi}_i) = \pi_i(1 - \pi_i)/n. \tag{3.7}$$

In Section 14.1.4 we show that for $i \neq j$, $\hat{\pi}_i$ and $\hat{\pi}_j$ have covariance

$$\text{cov}(\hat{\pi}_i, \hat{\pi}_j) = -\pi_i\pi_j/n. \tag{3.8}$$

The sample proportions $(\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_{c-1})$ have a large-sample multivariate normal distribution. For functions of them, the delta method implies the

following result, proved in Section 14.1.4:

Let $g(\boldsymbol{\pi})$ denote a differentiable function of $\{\pi_i\}$, with sample value $g(\hat{\boldsymbol{\pi}})$ for a multinomial sample. Let

$$\phi_i = \frac{\partial g(\boldsymbol{\pi})}{\partial \pi_i}, \qquad i = 1, \ldots, c.$$

Then as $n \to \infty$, the distribution of $\sqrt{n}[g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi})]/\sigma$ converges to standard normal, where

$$\sigma^2 = \sum \pi_i \phi_i^2 - \Big(\sum \pi_i \phi_i\Big)^2. \tag{3.9}$$

The asymptotic variance depends on $\{\pi_i\}$ and the partial derivatives of the measure with respect to $\{\pi_i\}$. In practice, replacing $\{\pi_i\}$ and $\{\phi_i\}$ in (3.9) by their sample values yields an ML estimate $\hat{\sigma}^2$ of σ^2. Then $\hat{\sigma}/\sqrt{n}$ is an estimated standard error for $g(\hat{\boldsymbol{\pi}})$. A large-sample Wald confidence interval for $g(\boldsymbol{\pi})$ is

$$g(\hat{\boldsymbol{\pi}}) \pm z_{\alpha/2}\hat{\sigma}/\sqrt{n}\,.$$

With the substitution of $\hat{\sigma}$ for σ in (3.9), the limiting distribution is still standard normal, but convergence is slower. The equivalence in the large-sample distribution is justified as follows: The sample proportions converge in probability to $\{\pi_i\}$, by the weak law of large numbers. Since $\hat{\sigma}$ is a continuous function of the sample proportions, it converges in probability to σ, and $\sigma/\hat{\sigma}$ converges in probability to 1. Now

$$\sqrt{n}\,\frac{g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi})}{\hat{\sigma}} = \sqrt{n}\,\frac{g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi})}{\sigma}\,\frac{\sigma}{\hat{\sigma}}.$$

The first term on the right-hand side converges in distribution to standard normal, by (3.9), and the second term converges in probability to 1. Thus, their product also has a limiting standard normal distribution.

We now apply the delta method to the log odds ratio, taking $g(\boldsymbol{\pi}) = \log\theta = \log\pi_{11} + \log\pi_{22} - \log\pi_{12} - \log\pi_{21}$. Since

$$\phi_{11} = \partial(\log\theta)/\partial\pi_{11} = 1/\pi_{11}$$

$$\phi_{12} = -1/\pi_{12}, \qquad \phi_{21} = -1/\pi_{21}, \qquad \phi_{22} = 1/\pi_{22},$$

$\sum_i\sum_j \pi_{ij}\phi_{ij} = 0$ and $\sigma^2 = \sum_i\sum_j \pi_{ij}\phi_{ij}^2 = \sum_i\sum_j(1/\pi_{ij})$. The asymptotic standard error of $\log\hat{\theta}$ for a multinomial sample $\{n_{ij}\}$ is

$$\sigma\big(\log\hat{\theta}\big) = \sigma/\sqrt{n} = \left(\sum_i\sum_j 1/n\pi_{ij}\right)^{1/2}.$$

Since $n\hat{\pi}_{ij} = n_{ij}$, the estimated standard error is (3.1).

The delta method also applies directly with θ to obtain $\hat{\sigma}(\hat{\theta})$ and a Wald confidence interval $\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}(\hat{\theta})$. This is not recommended; $\hat{\theta}$ converges more slowly than $\log \hat{\theta}$ to normality, this interval could contain negative values, and it does not give results equivalent to those obtained with the Wald interval using $1/\hat{\theta}$ and its standard error.

3.1.8 Score and Profile Likelihood Confidence Intervals*

Standard errors obtained with the delta method appear in Wald confidence intervals. However, intervals based on inverting Wald tests sometimes work poorly for small to moderate n. Alternative intervals result from inverting likelihood-ratio or score tests. Although computationally more complex, these methods often perform better.

We illustrate first with the score method for the difference of proportions. The score test (Mee 1984; Miettinen and Nurminen 1985) of H_0: $\pi_1 - \pi_2 = \Delta$ has the test statistic

$$z(\Delta) = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - \Delta}{\sqrt{\hat{\pi}_1(\Delta)[1 - \hat{\pi}_1(\Delta)]/n_1 + \hat{\pi}_2(\Delta)[1 - \hat{\pi}_2(\Delta)]/n_2}}$$

where $\hat{\pi}_i(\Delta)$ denotes the ML estimate of π_i subject to the constraint $\pi_1 - \pi_2 = \Delta$. That is, $\hat{\pi}_1(\Delta)$ and $\hat{\pi}_2(\Delta)$ are the values of π_1 and π_2 satisfying $\pi_1 - \pi_2 = \Delta$ that maximize the product of the two binomial probability mass functions. These values do not have closed-form expressions and are determined using numerical methods. The score confidence interval is the set of Δ such that $|z(\Delta)| < z_{\alpha/2}$. Computations for such intervals require iteration (Nurminen 1986).

For the relative risk also, slightly better performance results with an interval using the score method (Bedrick 1987; Gart and Nam 1988; Koopman 1984, Miettinen and Nurminen 1985; Nurminen 1986). Cornfield (1956) and Miettinen and Nurminen (1985) showed the score interval for the odds ratio. We prefer not to use a continuity or finite-sampling correction with these intervals, as then performance is too conservative. The fact that the score intervals are computationally more complex than Wald intervals should not be an impediment to their use in this modern era of computing, as the principle behind them is simple. However, currently they are not available in standard software.

For a confidence interval based on the likelihood-ratio test, we illustrate with the odds ratio. The multinomial likelihood for a 2×2 table is a function of $\{\pi_{11},\pi_{12},\pi_{21}\}$. Equivalently, it can be expressed in terms of $\{\theta,\pi_{1+},\pi_{+1}\}$ (recall Section 2.4.1). Thus, in inverting a likelihood-ratio test of H_0: $\theta = \theta_0$ to check whether θ_0 belongs in the confidence interval, there are two *nuisance parameters*. Their null ML estimates $\hat{\pi}_{1+}(\theta_0)$ and $\hat{\pi}_{+1}(\theta_0)$ that maximize the likelihood under the null vary as θ_0 does.

The *profile log-likelihood function* is $L(\theta_0, \hat{\pi}_{1+}(\theta_0), \hat{\pi}_{+1}(\theta_0))$, viewed as a function of θ_0. For each θ_0 this function gives the maximum of the ordinary log likelihood subject to the constraint $\theta = \theta_0$. Evaluated at $\theta_0 = \hat{\theta}$, this is the maximized log likelihood $L(\hat{\theta}, \hat{\pi}_{1+}, \hat{\pi}_{+1})$, which occurs at the sample proportions $\hat{\pi}_{1+} = n_{1+}/n$ and $\hat{\pi}_{+1} = n_{+1}/n$. The profile likelihood confidence interval for θ is the set of θ_0 for which

$$-2\Big[L(\theta_0, \hat{\pi}_{1+}(\theta_0), \hat{\pi}_{+1}(\theta_0)) - L\Big(\hat{\theta}, \hat{\pi}_{1+}, \hat{\pi}_{+1}\Big)\Big] < \chi_1^2(\alpha) .$$

This contains all θ_0 not rejected in likelihood-ratio tests of nominal size α.

The profile likelihood approach is available with some software (e.g., for SAS, see Table A.2 in Appendix A). A related approach, discussed in Section 6.7.1, uses a *conditional likelihood function* that eliminates the nuisance parameters by conditioning on their sufficient statistics. This is beneficial when there are many nuisance parameters. An advantage of score and likelihood-based intervals is that unlike the Wald, they are not adversely affected when the sample relative risk or odds ratio is 0 or ∞.

In this section we have discussed interval estimation. Significance tests normally refer to a null hypothesis value of 0.0 for the log odds ratio, log relative risk, and difference of proportions. These are special cases of independence applied to 2×2 tables. In the next section we present tests of independence for two-way contingency tables.

3.2 TESTING INDEPENDENCE IN TWO-WAY CONTINGENCY TABLES

For multinomial sampling with probabilities $\{\pi_{ij}\}$ in an $I \times J$ contingency table, the null hypothesis of statistical independence is H_0: $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all i and j. For independent multinomial samples in the I rows, independence corresponds to homogeneity of each outcome probability among the rows. Our discussion refers to a single multinomial sample, but the same tests apply with independent multinomial samples.

3.2.1 Pearson and Likelihood-Ratio Chi-Squared Tests

In Section 1.5.2 we introduced the Pearson X^2 statistic (1.15) for tests about multinomial probabilities. A test of H_0: independence uses X^2 with n_{ij} in place of n_i and with $\mu_{ij} = n\pi_{i+}\pi_{+j}$ in place of μ_i. Here $\mu_{ij} = E(n_{ij})$ under H_0. Usually, $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ are unknown. Their ML estimates are the sample marginal proportions $\hat{\pi}_{i+} = n_{i+}/n$ and $\hat{\pi}_{+j} = n_{+j}/n$, so estimated expected frequencies are $\{\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = n_{i+}n_{+j}/n\}$. Then X^2 equals

$$X^2 = \sum_i \sum_j \frac{\left(n_{ij} - \hat{\mu}_{ij}\right)^2}{\hat{\mu}_{ij}} . \tag{3.10}$$

Pearson (1900, 1904, 1922) claimed that replacing $\{\mu_{ij}\}$ by estimates $\{\hat{\mu}_{ij}\}$ would not affect the distribution of X^2. Since the contingency table has IJ categories, he argued that X^2 is asymptotically chi-squared with df $= IJ - 1$. On the contrary, since $\{\hat{\mu}_{ij}\}$ require estimating $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$, by Section 1.5.6

$$\text{df} = (IJ - 1) - (I - 1) - (J - 1) = (I - 1)(J - 1).$$

The dimensions of $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ reflect the constraints $\sum_i \pi_{i+} = \sum_j \pi_{+j} = 1$. R. A. Fisher (1922) corrected Pearson's error (see Section 16.2). His article introduced the notion of *degrees of freedom*. (Pearson had dealt with an indexed family of chi-squared distributions but had not dealt explicitly with "degrees of freedom.")

The score test produces the X^2 statistic. The likelihood-ratio test produces a different one. For multinomial sampling, the kernel of the likelihood is

$$\prod_i \prod_j \pi_{ij}^{n_{ij}}, \quad \text{where all } \pi_{ij} \geq 0 \quad \text{and} \quad \sum_i \sum_j \pi_{ij} = 1.$$

Under H_0: independence, $\hat{\pi}_{ij} = \hat{\pi}_{i+}\hat{\pi}_{+j} = n_{i+}n_{+j}/n^2$. In the general case, $\hat{\pi}_{ij} = n_{ij}/n$. The ratio of the likelihoods equals

$$\Lambda = \frac{\prod_i \prod_j (n_{i+}n_{+j})^{n_{ij}}}{n^n \prod_i \prod_j n_{ij}^{n_{ij}}}.$$

The likelihood-ratio chi-squared statistic is $-2\log\Lambda$. Denoted by G^2, it equals

$$G^2 = -2\log\Lambda = 2\sum_i \sum_j n_{ij} \log\left(n_{ij}/\hat{\mu}_{ij}\right) \tag{3.11}$$

where $\{\hat{\mu}_{ij} = n_{i+}n_{+j}/n\}$. The larger the values of G^2 and X^2, the more evidence exists against independence.

In the general case, the parameter space consists of $\{\pi_{ij}\}$ subject to the linear restriction $\sum_i \sum_j \pi_{ij} = 1$, so the dimension is $IJ - 1$. Under H_0, $\{\pi_{ij}\}$ are determined by $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$, so the dimension is $(I - 1) + (J - 1)$. The difference in these dimensions equals $(I - 1)(J - 1)$. For large samples, G^2 has a chi-squared null distribution with df $= (I - 1)(J - 1)$. So G^2 and X^2 have the same limiting null chi-squared distribution. In fact, they are then asymptotically equivalent; $X^2 - G^2$ converges in probability to zero (Section 14.3.4). The limiting results for multinomial sampling also hold with other sampling schemes (Roy and Mitra 1956, Watson 1959).

These results apply as n grows, and hence $\{\mu_{ij} = n\pi_{ij}\}$ grow, for a fixed number of cells. As they grow, the multinomial distribution for $\{n_{ij}\}$ is better

approximated by a multivariate normal, and X^2 and G^2 have more nearly chi-squared distributions. The convergence to chi-squared is quicker for X^2 than G^2. The approximation is usually poor for G^2 when $n/IJ < 5$. When I or J is large, it can be decent for X^2 when some expected frequencies are as small as 1 but most exceed 5. In Section 9.8.4 we provide further guidelines. Small-sample methods (Section 3.5) are available whenever it is doubtful whether n is sufficiently large.

3.2.2 Education and Religious Fundamentalism Example

Table 3.2 cross-classifies the degree of fundamentalism of subjects' religious beliefs by their highest degree of education. The table also contains the estimated expected frequencies for H_0: independence. For instance, $\hat{\mu}_{11} = n_{1+}n_{+1}/n = (424 \times 886)/2726 = 137.8$. The chi-squared statistics are $X^2 = 69.2$ and $G^2 = 69.8$, with df $= (3-1)(3-1) = 4$. The P-values are < 0.0001. These statistics provide extremely strong evidence of an association.

3.3 FOLLOWING-UP CHI-SQUARED TESTS

Like any significance test, chi-squared tests of independence have limited usefulness. A small P-value indicates strong evidence of association but provides little information about the nature or strength of the association. Statisticians have long warned about dangers of relying solely on results of chi-squared tests rather than studying the nature of the association (e.g., Berkson 1938; Cochran 1954). In this section we discuss ways to follow up the tests to learn more about the association.

TABLE 3.2 Education and Religious Beliefs

	Religious Beliefs			
Highest Degree	Fundamentalist	Moderate	Liberal	Total
Less than high school	178	138	108	424
	(137.8)[1]	(161.5)	(124.7)	
	(4.5)[2]	(−2.6)	(−1.9)	
High school or junior college	570	648	442	1660
	(539.5)	(632.1)	(488.4)	
	(2.6)	(1.3)	(−4.0)	
Bachelor or graduate	138	252	252	642
	(208.7)	(244.5)	(188.9)	
	(−6.8)	(0.7)	(6.3)	
Total	886	1038	802	2726

Source: 1996 General Social Survey, National Opinion Research Center.

[1]Estimated expected frequencies for testing independence; [2]standardized Pearson residuals.

3.3.1 Pearson and Standardized Residuals

A cell-by-cell comparison of observed and estimated expected frequencies helps show the nature of the dependence. Under H_0, larger differences $(n_{ij} - \hat{\mu}_{ij})$ tend to occur in cells with larger μ_{ij}. For Poisson sampling, for instance, the standard deviation of n_{ij} and hence $(n_{ij} - \mu_{ij})$ is $\sqrt{\mu_{ij}}$; the standard deviation of $(n_{ij} - \hat{\mu}_{ij})$ is less than that of $n_{ij} - \mu_{ij}$ but is proportional to $\sqrt{\mu_{ij}}$. Thus, this raw difference is insufficient. The *Pearson residual*, defined for a cell by

$$e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\hat{\mu}_{ij}^{1/2}}, \tag{3.12}$$

attempts to adjust for this. Pearson residuals relate to the Pearson statistic by $\sum_i \sum_j e_{ij}^2 = X^2$.

Under H_0, $\{e_{ij}\}$ are asymptotically normal with mean 0. However, in Section 14.3.2 we show that their asymptotic variances are less than 1.0, averaging $[(I-1)(J-1)]/$(number of cells). Comparing Pearson residuals to standard normal percentage points provides conservative indications of cells having lack of fit.

A *standardized Pearson residual* that is asymptotically standard normal results from dividing it by its standard error (Haberman 1973a; see also Section 14.3.2). For H_0: independence, this is

$$\frac{n_{ij} - \hat{\mu}_{ij}}{\left[\hat{\mu}_{ij}(1 - p_{i+})(1 - p_{+j})\right]^{1/2}}. \tag{3.13}$$

A standardized Pearson residual that exceeds about 2 or 3 in absolute value indicates lack of fit of H_0 in that cell. Larger values are more relevant when df is larger and it becomes more likely that at least one is large simply by chance.

3.3.2 Education and Religious Fundamentalism Revisited

Table 3.2 also shows standardized Pearson residuals for testing independence. For instance, $n_{11} = 178$ and $\hat{\mu}_{11} = 137.8$. The relevant marginal proportions equal $p_{1+} = 424/2726 = 0.156$ and $p_{+1} = 886/2726 = 0.325$. The standardized Pearson residual (3.13) for this cell equals

$$(178 - 137.8)/\left[(137.8)(1 - 0.156)(1 - 0.325)\right]^{1/2} = 4.5.$$

This cell shows a much greater discrepancy between n_{11} and $\hat{\mu}_{11}$ than expected if the variables were truly independent.

Table 3.2 shows large positive residuals for subjects with less than a high school education and fundamentalist views and for subjects with a bachelor's

or graduate degree and liberal views. This means that significantly more subjects were at these combinations than H_0: independence predicts. Similarly, there were fewer subjects with high levels of education and fundamentalist views and with low levels of education and liberal views than independence predicts.

Odds ratios describe this trend. The 2×2 table constructed from the first and last rows and the first and last columns of Table 3.2 has a sample odds ratio of $(178 \times 252)/(108 \times 138) = 3.0$. For those with a bachelor's or graduate degree, the estimated odds of selecting liberal instead of fundamentalist were 3.0 times the estimated odds for those with less than a high school education.

3.3.3 Partitioning Chi-Squared

Let Z denote a standard normal random variable. Then Z^2 has a chi-squared distribution with df = 1. A chi-squared random variable with df = ν has representation $Z_1^2 + \cdots + Z_\nu^2$, where $Z_1, \ldots, Z_\nu$ are independent standard normal variables. Thus, a chi-squared statistic having df = ν has partitionings into independent chi-squared components—for example, into ν components each having df = 1. Conversely, if X_1^2 and X_2^2 are independent chi-squared random variables having degrees of freedom ν_1 and ν_2, then $X^2 = X_1^2 + X_2^2$ has a chi-squared distribution with df = $\nu_1 + \nu_2$. Another supplement to a chi-squared test partitions its test statistic so that the components represent certain aspects of the effects. A partitioning may show that an association reflects primarily differences between certain categories or groupings of categories.

We begin with a partitioning for the test of independence in $2 \times J$ tables. We partition G^2, which has df = $(J - 1)$, into $J - 1$ components. The jth component is G^2 for a 2×2 table where the first column combines columns 1 through j of the full table and the second column is column $j + 1$. That is, G^2 for testing independence in a $2 \times J$ table equals a statistic that compares the first two columns, plus a statistic that combines the first two columns and compares them to the third column, and so on, up to a statistic that combines the first $J - 1$ columns and compares them to the last column. (In Section 9.2.4 we justify this partitioning.) Each component statistic has df = 1.

It might seem more natural to compute G^2 for the $(J - 1)$ separate 2×2 tables that pair each column with a particular one, say the last. However, these component statistics are not independent and do not sum to G^2 for the full table. (This is beyond our scope at this stage but relates to the contrasts of log probabilities that form the log odds ratios for the two tables not being orthogonal.)

For an $I \times J$ table, independent chi-squared components result from comparing columns 1 and 2 and then combining them and comparing them to column 3, and so on. Each of the $J - 1$ statistics has df = $I - 1$. More refined partitions contain $(I - 1)(J - 1)$ statistics, each having df = 1. One

such partition (Lancaster 1949) applies to the $(I-1)(J-1)$ separate 2×2 tables

$$\begin{array}{c|c} \sum_{a<i}\sum_{b<j} n_{ab} & \sum_{a<i} n_{aj} \\ \hline \sum_{b<j} n_{ib} & n_{ij} \end{array} \tag{3.14}$$

for $i = 2, \ldots, I$ and $j = 2, \ldots, J$. For others, see Gilula and Haberman (1998) and Goodman (1969a, 1971b).

3.3.4 Origin of Schizophrenia Example

Table 3.3 classifies a sample of psychiatrists by their school of psychiatric thought and by their opinion on the origin of schizophrenia. Here $G^2 = 23.04$ with df = 4. To understand this association better, we partition G^2 into four independent components. The partitioning (3.14) applies to the subtables shown in Table 3.4.

The first subtable compares the eclectic and medical schools of psychiatric thought on whether the origin of schizophrenia is biogenic or environmental given that the classification was in one of these two categories. For this subtable, $G^2 = 0.29$, with df = 1. The second subtable compares these two schools on the proportion of times the origin was ascribed to be a combination, rather than biogenic or environmental. This subtable has $G^2 = 1.36$,

TABLE 3.3 Most Influential School of Psychiatric Thought and Ascribed Origin of Schizophrenia

School of Psychiatric Thought	Origin of Schizophrenia		
	Biogenic	Environmental	Combination
Eclectic	90	12	78
Medical	13	1	6
Psychoanalytic	19	13	50

Source: Reprinted with permission, based on data from B. J. Gallagher III, B. J. Jones, and L. P. Barakat, *J. Clin. Psychol.* **43:** 438–443 (1987).

TABLE 3.4 Subtables Used in Partitioning Chi-Squared for Table 3.3[a]

	Bio	Env		Bio + Env	Com		Bio	Env		Bio + Env	Com
Ecl	90	12	Ecl	102	78	Ecl + Med	103	13	Ecl + Med	116	84
Med	13	1	Med	14	6	Psy	19	13	Psy	32	50

[a]Bio, biogenic; Com, combination; Ecl, eclectic; Env, environmental; Psy, psychoanalytic

with df = 1. The sum of these two components equals G^2 for testing independence with the first two rows of Table 3.3. There is little evidence of a difference between the eclectic and medical schools of thought on the ascribed origin of schizophrenia.

Next we combine the eclectic and medical schools and compare them to the psychoanalytic school. The third subtable in Table 3.4 compares them for the (biogenic, environmental) classification, giving $G^2 = 12.95$ with df = 1. The fourth subtable compares them for the (biogenic or environmental, combination) split, giving $G^2 = 8.43$ with df = 1.

The psychoanalytic school seems more likely than the other schools to ascribe the origins of schizophrenia as being a combination. Of those who chose either the biogenetic or environmental origin, members of the psychoanalytic school were somewhat more likely than the other schools to choose the environmental origin. The sum of these four G^2 components equals the value of 23.04 for testing independence in the full table.

3.3.5 Rules for Partitioning

Goodman (1968, 1969a, 1971b) and Lancaster (1949, 1969) gave rules for determining independent components of chi-squared. For forming subtables, among the necessary conditions are the following:

1. The df for the subtables must sum to df for the full table.
2. Each cell count in the full table must be a cell count in one and only one subtable.
3. Each marginal total of the full table must be a marginal total for one and only one subtable.

For a certain partitioning, when the subtable df values sum properly but the G^2 values do not, the components are not independent.

For the G^2 statistic, exact partitionings occur. The Pearson X^2 need not equal the sum of the X^2 values for the subtables. It is valid to use the X^2 statistics for the separate subtables; they simply need not provide an exact algebraic partitioning of X^2 for the full table. When the null hypotheses all hold, X^2 does have an asymptotic equivalence with G^2, however. In addition, when the table has small counts, in large-sample chi-squared tests it is safer to use X^2 to study the subtables.

3.3.6 Limitations of Chi-Squared Tests

Chi-squared tests of independence merely indicate the degree of evidence of association. They are rarely adequate for answering all questions about a data set. Rather than relying solely on results of these tests, investigate the nature of the association: Study residuals, decompose chi-squared into components, and estimate parameters such as odds ratios that describe the strength of association.

The chi-squared tests also have limitations in the types of data to which they apply. For instance, they require large samples. Also, the $\{\hat{\mu}_{ij} = n_{i+}n_{+j}/n\}$ used in X^2 and G^2 depend on the marginal totals but not on the order of listing the rows and columns. Thus, X^2 and G^2 do not change value with arbitrary reorderings of rows or of columns. This implies that they treat both classifications as nominal. When at least one variable is ordinal, test statistics that utilize the ordinality are usually more appropriate. We present such tests in Section 3.4.

3.3.7 Why Consider Independence?

Any idealized structure such as independence is unlikely to hold in any given practical situation. With large samples such as in Table 3.2 it is not surprising to obtain a small P-value. Given this and the limitations just mentioned, why even bother to consider independence as a possible representation for a joint distribution? One reason refers to the benefits of model parsimony. If the independence model approximates the true probabilities well, then unless n is very large, the model-based estimates $\{\hat{\pi}_{ij} = n_{i+}n_{+j}/n^2\}$ of cell probabilities tend to be better than the sample proportions $\{p_{ij} = n_{ij}/n\}$. The independence ML estimates smooth the sample counts, somewhat damping the random sampling fluctuations.

The mean-squared error (MSE) formula

$$\text{MSE} = \text{variance} + (\text{bias})^2$$

explains why the independence estimators can have smaller MSE. Although they may be biased, they have smaller variance because they are based on estimating fewer parameters ($\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ instead of $\{\pi_{ij}\}$). Hence, MSE can be smaller unless n is so large that the bias term dominates the variance.

We illustrate using Table 3.5, which has $\pi_{ij} = \pi_{i+}\pi_{+j}[1 + \delta(i - 2)(j - 2)]$ for $\pi_{i+} = \pi_{+j} = \frac{1}{3}$. Here $-1 < \delta < 1$, with $\delta = 0$ equivalent to independence. Independence approximates the relationship well when δ is close to zero. The total MSE values of the two estimators are

$$\begin{aligned}
\text{MSE}\big(\{p_{ij}\}\big) &= \sum_i \sum_j E(p_{ij} - \pi_{ij})^2 = \sum_i \sum_j \text{var}(p_{ij}) \\
&= \sum_i \sum_j \pi_{ij}(1 - \pi_{ij})/n = \left(1 - \sum_i \sum_j \pi_{ij}^2\right)\Big/n \\
\text{MSE}\big(\{\hat{\pi}_{ij}\}\big) &= \sum_i \sum_j E\big(\hat{\pi}_{ij} - \pi_{ij}\big)^2.
\end{aligned}$$

TABLE 3.5 Cell Probabilities for Comparison of Estimators

$(1 + \delta)/9$	$1/9$	$(1 - \delta)/9$
$1/9$	$1/9$	$1/9$
$(1 - \delta)/9$	$1/9$	$(1 + \delta)/9$

TABLE 3.6 Comparison of Total MSE($\times$10,000)for Sample Proportion and Independence Estimators

	$\delta = 0$		$\delta = 0.1$		$\delta = 0.2$		$\delta = 0.6$		$\delta = 1.0$	
n	p	$\hat{\pi}$	p	$\hat{\pi}$	p	$\hat{\pi}$	p	$\hat{\pi}$	p	$\hat{\pi}$
10	889	489	888	493	887	505	871	634	840	893
50	178	91	178	95	177	110	174	261	168	565
100	89	45	89	50	89	65	87	220	84	529
500	18	9	18	14	18	28	17	186	17	500
∞	0	0	0	5	0	20	0	178	0	494

For Table 3.5,

$$\text{MSE}\big(\{p_{ij}\}\big) = \frac{1}{n}\left\{\frac{8}{9} - \frac{4\delta^2}{81}\right\}$$

and rather tedious calculations yield

$$\text{MSE}\big(\{\hat{\pi}_{ij}\}\big) = \frac{1}{n}\left\{\frac{4}{9} + \frac{4}{9n}\right\} + \frac{4\delta^2}{81}\left\{1 - \frac{2}{n} + \frac{2}{n^2} - \frac{2}{n^3}\right\}.$$

Table 3.6 lists the total MSE values for various δ and n. When $\delta = 0$, $\text{MSE}(\{p_{ij}\}) = 8/9n$, whereas $\text{MSE}(\{\hat{\pi}_{ij}\}) \approx 4/9n$ for large n. The independence estimator is then much better than the sample proportions. When the table is close to independence ($\delta \approx 0$) and n is not large, MSE is only about half as large for the independence estimator. When $\delta \neq 0$, the inconsistency of $\{\hat{\pi}_{ij}\}$ is reflected by $\text{MSE}(\{\hat{\pi}_{ij}\}) \to 4\delta^2/81$ [whereas $\text{MSE}(\{p_{ij}\}) \to 0$] as $n \to \infty$. When the table is close to independence, however, the independence estimator has a lower total MSE even for moderately large n (e.g., for $n = 500$ when $\delta = 0.1$).

3.4 TWO-WAY TABLES WITH ORDERED CLASSIFICATIONS

The X^2 and G^2 chi-squared tests ignore some information when used to test independence between ordinal classifications. When rows and/or columns are ordered, more powerful tests usually exist.

3.4.1 Linear Trend Alternative to Independence

When the row variable X and the column variable Y are ordinal, a positive or negative trend in the association is common. One approach to inference, described later in this section, uses an ordinal measure of monotone trend.

A more popular analysis assigns scores to categories and summarizes the *linear trend*.

A test statistic that is sensitive to positive or negative linear trends utilizes correlation information. Let $u_1 \leq u_2 \leq \cdots \leq u_I$ denote scores for the rows, and let $v_1 \leq v_2 \leq \cdots \leq v_J$ denote column scores. The scores have the same ordering as the categories. They assign distances between categories and actually treat the measurement scale as interval, with greater distances between categories that are farther apart.

The sum $\sum_i \sum_j u_i v_j n_{ij}$ weights cross-products of scores by their frequency. It relates to the covariation of X and Y. For the scores chosen, the correlation r between X and Y equals the standardization of this sum to the -1 to $+1$ scale (in fact, r equals this sum when both sets of scores are linearly transformed for the n subjects to have a mean of 0 and standard deviation of 1). The larger the correlation is in absolute value, the farther the data fall from independence in this linear dimension.

A statistic for testing independence against the two-sided alternative of nonzero true correlation is

$$M^2 = (n - 1)r^2. \tag{3.15}$$

This statistic increases as $|r|$ or n do. For large samples, it is approximately chi-squared with df = 1 (Mantel 1963). Large values contradict independence, so as with X^2 and G^2, the P-value is the right-tailed probability above the value observed. A small P-value does not imply that the association is linear, merely that searching for a linear component to the association helped to build power against H_0. The test treats the variables symmetrically.

3.4.2 Job Satisfaction Example Revisited

Table 2.8 showed job satisfaction and income for 96 subjects. The ordinary chi-squared statistics for testing independence are $X^2 = 6.0$ and $G^2 = 6.8$ with df = 9 (P-values = 0.74 and 0.66). These statistics show little evidence of association, but they ignore the ordering of rows and columns. With scores (1, 2, 3, 4) for job satisfaction and scores {7.5, 20, 32.5, 60} for income that approximate midpoints of categories in thousands of dollars, the correlation is $r = 0.200$. The linear trend test statistic $M^2 = (96 - 1)(0.200)^2 = 3.81$. This shows some evidence of association ($P = 0.051$). The evidence is stronger for the one-sided (positive trend) alternative, using $M = \sqrt{n - 1}\, r = 1.95$ ($P = 0.026$).

The nontrivial evidence of positive association may be surprising, since X^2 and G^2 have such unimpressive values. When a positive or negative trend exists, analyses designed to detect that trend can provide much smaller P-values than analyses that ignore it.

3.4.3 Monotone Trend Alternatives to Independence

Ordinal variables do not have a specified metric. Detecting a linear trend alternative to independence requires assigning scores to X and Y, treating them as interval variables. Alternatively, a strict ordinal analysis with the weaker alternative of monotonicity uses an ordinal measure of association, such as gamma (Section 2.4.4).

For large random samples, sample gamma has approximately a normal sampling distribution. The standard error (SE) follows from the delta method (Problem 3.27). Gamma is the basis of an ordinal test of independence using test statistic $z = \hat{\gamma}/\text{SE}$. A confidence interval describes the strength of positive or negative monotone association.

For Table 2.8 on income and job satisfaction, in Section 2.4.5 we showed that $\hat{\gamma} = 0.221$. The sample has a weak tendency for job satisfaction to be higher at higher income levels. Software (e.g., PROC FREQ in SAS) reports a standard error of 0.117 for gamma. There is some evidence that $\gamma > 0$, since $z = 0.221/0.117 = 1.89$ ($P = 0.03$ for the one-sided alternative). An approximate 95% confidence interval for γ is $0.221 \pm 1.96(0.117)$, or $(-0.01, 0.45)$. The true association between income and job satisfaction is at best moderately positive.

3.4.4 Extra Power with Ordinal Tests

For testing independence, X^2 and G^2 refer to the most general alternative, whereby cell probabilities exhibit *any* type of statistical dependence. Their df value of $(I-1)(J-1)$ reflects an alternative hypothesis that has $(I-1)(J-1)$ more parameters than the null hypothesis—the nonredundant odds ratios that describe the association [such as (2.10)]. These statistics are designed to detect any pattern for these parameters. In achieving this generality, they sacrifice sensitivity for detecting particular patterns.

By contrast, the analyses for ordinal row and column variables attempt to describe association using a single parameter. For instance, M^2 uses the correlation. When a chi-squared test statistic refers to a single parameter [such as M^2 or $(\hat{\gamma}/\text{SE})^2$ do], it has df = 1. When the association truly has a positive or negative trend, an ordinal test has a power advantage over the tests using X^2 or G^2. Since df equals the mean of the chi-squared distribution, a relatively large M^2 value with df = 1 falls farther out in its right-hand tail than a comparable value of X^2 or G^2 with df $= (I-1)(J-1)$; falling farther out in the tail produces a smaller P-value. The potential discrepancy in power increases as I and J increase. In Section 6.4 we present the theory behind such a power comparison.

3.4.5 Choice of Scores

Often, it is unclear how to assign scores to statistics that require them, such as M^2. Cochran (1954) noted that "any set of scores gives a *valid* test,

provided that they are constructed without consulting the results of the experiment. If the set of scores is poor, in that it badly distorts a numerical scale that really does underlie the ordered classification, the test will not be sensitive. The scores should therefore embody the best insight available about the way in which the classification was constructed and used." Ideally, the scale is chosen by a consensus of experts, and subsequent interpretations use that same scale.

How sensitive are analyses to the choice or scores? There is no simple answer, but different scoring systems can give quite different results (e.g., Graubard and Korn 1987). For most data sets, different choices of monotone scores give similar results. Scores that are linear transforms of each other, such as (1, 2, 3, 4) and (0, 2, 4, 6), have the same absolute correlation and hence the same M^2. Results *may* depend on the scores, however, when the data are highly unbalanced, with some categories having many more observations than others.

Table 3.7 illustrates the potential dependence. It refers to a prospective study of maternal drinking and congenital malformations. After the first three months of pregnancy, the women in the sample completed a questionnaire about alcohol consumption. Following childbirth, observations were recorded on the presence or absence of congenital sex organ malformations. When a variable is nominal but has only two categories, statistics that treat it as ordinal are still valid. For instance, we can artificially regard malformation as ordinal, treating "present" as "high" and "absent" as "low." With only two rows, any set of distinct row scores is a linear transformation of any other set and gives the same M^2 value. Alcohol consumption, measured as the average number of drinks per day, is an ordinal explanatory variable. This groups a naturally continuous variable, and we first use the scores $\{v_1 = 0, v_2 = 0.5, v_3 = 1.5, v_4 = 4.0, v_5 = 7.0\}$, the last score being somewhat arbitrary. For this choice, $M^2 = 6.57$, for which the P-value is 0.010. By contrast, for the equally spaced row scores (1, 2, 3, 4, 5), $M^2 = 1.83$, giving a much weaker conclusion ($P = 0.18$).

An alternative approach uses the data to form the scores automatically, by using ranks as the category scores. All subjects in a category receive the average of the ranks that would apply for a complete ranking of the sample from 1 to n. These are called *midranks*. The 17,114 subjects at level 0 for

TABLE 3.7 Example for which Results Depend on Choice of Scores

	Alcohol Consumption (average number of drinks per day)				
Malformation	0	< 1	1–2	3–5	≥ 6
Absent	17,066	14,464	788	126	37
Present	48	38	5	1	1

Source: Reprinted with permission from the Biometric Society (Graubard and Korn 1987).

alcohol consumption share ranks 1 through 17,114. Each receives the average of these ranks, which is the midrank $(1 + 17{,}114)/2 = 8557.5$. Similarly, the midranks for the last four categories are 24,365.5, 32,013, 32,473, and 32,555.5. These scores yield $M^2 = 0.35$ and a weaker conclusion yet ($P = 0.55$).

Why does this happen? Adjacent categories having relatively few observations necessarily have similar midranks. The midranks are similar for the final three categories, since those categories have few observations compared with the first two categories. This scoring scheme treats alcohol consumption level 1–2 drinks (category 3) as much closer to consumption level $\geq$ 6 drinks (category 5) than to consumption level 0 drinks (category 1). This seems inappropriate. It is usually better to select scores that reflect distances between categories. When uncertain about this choice, a sensitivity analysis should be performed, selecting two or three sensible choices and checking whether results are similar. Equally spaced scores often provide a reasonable compromise when the category labels do not suggest obvious choices, such as the categories (liberal, moderate, conservative) for political philosophy.

When X and Y are both ordinal and M^2 uses midrank scores, the correlation on which M^2 is based is called *Spearman's rho*.

3.4.6 Trend Tests for $I \times 2$ and $2 \times J$ Tables

When I or J equal 2, the tests based on linear or monotonic trend simplify to well-established procedures. With binary X, $2 \times J$ tables occur in comparisons of two groups, such as when the rows represent two treatments. Using scores $\{u_1 = 0, u_2 = 1\}$ for levels of X, the covariation measure $\sum_i \sum_j u_i v_j n_{ij}$ in M^2 simplifies to $\sum_j v_j n_{2j}$. This term sums the scores on Y for all subjects in row 2. Divided by the number of subjects in row 2, it gives the mean score for that row. In fact, M^2 is then directed toward detecting differences between the two row means of the scores on Y.

With midrank scores for Y, the test using M^2 for $2 \times J$ tables is sensitive to differences in mean ranks for the two rows. This test is called the *Wilcoxon* or *Mann–Whitney test*. Most nonparametric statistics textbooks present this test for fully ranked response data, whereas the $2 \times J$ table is an extended case in which sets of subjects in the same category of Y are tied and use midranks. The large-sample version of that nonparametric test uses a standard normal z statistic. The square of the statistic is equivalent to M^2, using arbitrary row scores and midranks for the columns. It is also asymptotically equivalent to test statistics based on the numbers of concordant and discordant pairs, such as the one using gamma.

When Y has two levels, the table has size $I \times 2$. The linear trend statistic then refers to a linear trend in the probability of either response category, such as the probability of malformation as a function of alcohol consumption. The test in that case, often called the *Cochran–Armitage trend test*, is presented in Section 5.3.5.

3.4.7 Nominal–Ordinal Tables

The tests using the correlation or gamma are appropriate when both classifications are ordinal. When one is nominal with more than two categories, other statistics are needed. One is based on summarizing the variation among means on the ordinal variable in the various categories of the nominal variable. We defer discussion of this case to Section 7.5.3, Note 3.6, and Problem 3.28.

3.5 SMALL-SAMPLE TESTS OF INDEPENDENCE

The inferential methods of the preceding four sections are large-sample methods. When n is small, alternative methods use *exact* small-sample distributions rather than large-sample approximations. In this section we describe small-sample tests of independence, starting with one that R. A. Fisher proposed for 2×2 tables.

3.5.1 Fisher's Exact Test for 2×2 Tables

In Section 3.5.7 we show that a distribution not depending on unknown parameters results from conditioning on the marginal totals of the contingency table. These are usually not naturally fixed. For Poisson sampling nothing is fixed, for multinomial sampling only n is fixed, and for independent binomial sampling in the two rows only the row marginal totals are fixed. In any of these cases, under H_0: independence, conditioning on both sets of marginal totals yields the hypergeometric distribution

$$p(t) = P(n_{11} = t) = \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{+1} - t}}{\binom{n}{n_{+1}}}. \tag{3.16}$$

This formula expresses the distribution of $\{n_{ij}\}$ in terms of only n_{11}. Given the marginal totals, n_{11} determines the other three cell counts. The range of possible values for n_{11} is $m_- \le n_{11} \le m_+$, where $m_- = \max(0, n_{1+} + n_{+1} - n)$ and $m_+ = \min(n_{1+}, n_{+1})$.

For 2×2 tables, independence is equivalent to the odds ratio $\theta = 1$. To test H_0: $\theta = 1$, the P-value is the sum of certain hypergeometric probabilities. To illustrate, consider H_a: $\theta > 1$. For the given marginal totals, tables having larger n_{11} have larger sample odds ratios and hence stronger evidence in favor of H_a. Thus, the P-value equals $P(n_{11} \ge t_o)$, where t_o denotes the observed value of n_{11}. This test for 2×2 tables is called *Fisher's exact test* (Fisher 1934, 1935a,c; Irwin 1935; Yates 1934).

3.5.2 Fisher's Tea Drinker

R. A. Fisher (1935a) described the following experiment from his days at Rothamsted Experiment Station, an agriculture research lab north of London. Muriel Bristol, a colleague of Fisher's, claimed that when drinking tea she could distinguish whether milk or tea was added to the cup first (she preferred milk first). To test her claim, Fisher asked her to taste eight cups of tea, four of which had milk added first and four of which had tea added first. She knew there were four cups of each type and had to predict which four had the milk added first. The order of presenting the cups to her was randomized.

Table 3.8 shows a possible result. Distinguishing the order of pouring better than with pure guessing corresponds to $\theta > 1$, reflecting a positive association between order of pouring and the prediction. We conduct Fisher's exact test of H_0: $\theta = 1$ against H_a: $\theta > 1$.

The experimental design fixed both marginal distributions, since Dr. Bristol had to predict which four cups had milk added first. Thus, the hypergeometric applies naturally for the null distribution of n_{11}. The P-value for Fisher's exact test is the null probability of Table 3.8 and of tables having even more evidence in favor of her claim. The observed table, $t_o = 3$ correct choices of the cups having milk added first, has null probability

$$\frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = 0.229.$$

The only table that is more extreme in the direction of H_a has $n_{11} = 4$ correct. It has a probability of 0.014. The P-value is $P(n_{11} \geq 3) = 0.243$. This result does not establish an association between the actual order of pouring and her predictions. It is difficult to do so with such a small sample. According to Fisher's daughter (Box 1978, p. 134), in reality Bristol did convince Fisher of her ability.

TABLE 3.8 Fisher's Tea Tasting Experiment

	Guess Poured First		
Poured First	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	

Source: Based on experiment described by Fisher (1935a).

3.5.3 Two-Sided *P*-Values for Fisher's Exact Test

For the one-sided alternative, the same *P*-value results using tables ordered according to larger n_{11}, larger odds ratio, or larger difference of proportions (Davis 1986a). For the two-sided alternative, different criteria can have different *P*-values.

For a two-sided *P*-value, a popular approach sums $P(n_{11} = t)$ in (3.16) for counts t such that $p(t) \leq p(t_o)$; that is, the *P*-value is $P = P[\,p(n_{11}) \leq p(t_0)]$ for the observed value t_o. Another possibility sums $p(t)$ for tables that are farther from H_0; that is,

$$P = P\big[\,|n_{11} - E(n_{11})| \geq |t_0 - E(n_{11})|\,\big],$$

where the hypergeometric $E(n_{11}) = n_{1+}n_{+1}/n$. This is identical to $P(X^2 \geq X_o^2)$ for observed Pearson statistic X_o^2. A third approach takes $P = 2\min[P(n_{11} \geq t_o),\ P(n_{11} \leq t_o)]$, but this can exceed 1. A fourth approach takes $P = \min[P(n_{11} \geq t_o),\ P(n_{11} \leq t_o)]$ plus an attainable probability in the other tail that is as close as possible to, but not greater than, that one-tailed probability.

Each approach has advantages and disadvantages (Blaker 2000; Davis 1986a; Dupont 1986; Lloyd 1988b; Mantel 1987b; Yates and discussants 1984). They can provide different results because of the discreteness and potential skewness. The approach of ordering tables by a distance measure from H_0, such as X^2, extends naturally to $I \times J$ tables.

In practice, two-sided tests are much more common than one-sided. Partly this is so that researchers can avoid charges of bias in giving evidence that supports their predicted direction for an effect. To conduct a test of size 0.05 when one truly believes that the effect has a particular direction, it is safest to conduct the one-sided test at the 0.025 level to guard against criticism. For instance, in the 1998 document *Biostatistical Principles for Clinical Trials*, the International Conference on Harmonization (ICH E9) stated: "The approach of setting type I errors for one-sided tests at half the conventional type I error used in two-sided tests is preferable in regulatory settings. This promotes consistency with two-sided confidence intervals that are generally appropriate for estimating the possible size of the difference between two treatments."

3.5.4 Discreteness and Conservatism Issues

The hypergeometric distribution (3.16) is highly discrete for small samples, as n_{11} and hence the *P*-value can assume relatively few values. It is usually not possible to achieve a fixed significance level (size) such as 0.05.

In the tea-tasting experiment, for instance, n_{11} can equal only 4, 3, 2, 1, 0. The one-sided *P*-values are restricted to 0.014, 0.243, 0.757, 0.986, and 1.0. If

one rejects H_0 when the P-value does not exceed 0.05, then 0.05 is not the probability of type I error. Only the P-value of 0.014 does not exceed 0.05; thus, when H_0 is true, the probability of falsely rejecting it is 0.014, not 0.05. In this sense, the traditional approach to hypothesis testing is conservative: The true probability of type I error is less than the nominal level.

It *is* possible to achieve any fixed significance level by data-unrelated randomization on the boundary of the critical region, in deciding whether to reject H_0. For the tea-tasting experiment, suppose that we reject H_0 when $n_{11} = 4$, we reject H_0 with probability 0.157 when $n_{11} = 3$, and we do not reject H_0 otherwise; that is, when $n_{11} = 3$, we generate a uniform random variable U over [0, 1] and reject H_0 if $U < 0.157$. For expectation taken with respect to the null hypergeometric distribution of n_{11}, the significance level equals

$$\begin{aligned} P(\text{reject } H_0) &= E\big[P(\text{reject } H_0 | n_{11})\big] \\ &= 1.0(0.014) + 0.157(0.229) + 0.0 \times P(n_{11} \le 2) = 0.05. \end{aligned}$$

With the randomization extension, Tocher (1950) showed that Fisher's test is uniformly most powerful unbiased (UMPU).

In practice, randomization having nothing to do with the data is unacceptable. We recommend simply reporting the P-value. To reduce conservativeness, report the mid-P-value (Section 1.4.5). The test is no longer guaranteed to have true P(type I error) no greater than the nominal value, but in practice it is rarely much greater. For the one-sided test with the tea-tasting data,

$$\text{mid-}P\text{-value} = (1/2)P(n_{11} = 3) + P(n_{11} > 3) = 0.129.$$

3.5.5 Small-Sample Unconditional Test of Independence*

A common sampling assumption for analyses comparing two groups on a binary response is that the rows are independent binomial samples. Then, only $\{n_{i+}\}$ are naturally fixed. For Poisson and multinomial sampling schemes, neither marginal distribution is fixed. For such cases it may seem artificial to condition on *both* sets of marginal counts. An alternative small-sample test, designed for independent binomial samples, conditions on only the row totals.

Under binomial sampling with parameter π_i in row i, consider testing H_0: $\pi_1 = \pi_2$ using some test statistic T, such as the Pearson X^2. For fixed $\{n_{i+}\}$, T can take a discrete set of values, one of which is the observed value t_o. Given $\pi_1 = \pi_2 = \pi$, the P-value is $P_\pi(T \ge t_o)$, calculated using the product of the two binomial probability mass functions. This is the sum of the product binomial probabilities for those pairs of binomial samples that have $T \ge t_o$. Since π is unknown, the actual P-value is defined as

$$P = \sup_{0 \le \pi \le 1} P_\pi(T \ge t_o).$$

This is an *unconditional* small-sample test of independence. Like Fisher's exact test, the true size is no greater than the nominal value (e.g., if we reject when $P \leq 0.05$, the actual P(type I error) is no greater than 0.05).

We illustrate using test statistic X^2 for the 2×2 table having entries (3, 0/0, 3), by row, with fixed row totals (3, 3) as binomial sample sizes. The sample $X^2 = 6.0$. This X^2 value for the observed table and for table (0, 3/3, 0) is the maximum possible. For a given value π for $\pi_1 = \pi_2$, the probability of the first table is $[\pi^3(1-\pi)^0][\pi^0(1-\pi)^3] = \pi^3(1-\pi)^3$ (3 successes and 0 failures in the first row and 0 successes and 3 failures in the second), the product of two binomial probabilities. Similarly, the probability of the second table is $(1-\pi)^3\pi^3$. Thus, the P-value is $P_\pi(X^2 \geq 6) = 2\pi^3(1-\pi)^3$, the sum of the product binomial probabilities for those two tables. The supremum of this over $0 \leq \pi \leq 1$ occurs at $\pi = \frac{1}{2}$, giving overall P-value equal to $2(0.5)^3(0.5)^3 = 0.031$. By contrast, the two-sided Fisher's exact test has P-value equal to $2\binom{3}{0}\binom{3}{3}/\binom{6}{3} = 0.100$.

Barnard (1945, 1947) first proposed an unconditional test comparing binomial parameters, although he later (1949) refuted it in favor of Fisher's exact test. Several authors have since proposed related tests (e.g., Haber 1986; Suissa and Shuster 1985).

3.5.6 Conditional versus Unconditional Tests*

Since Barnard introduced the unconditional test, statisticians have debated the proper way to conduct small-sample analyses of 2×2 tables. Fisher criticized the unconditional approach, arguing that possible samples with quite different numbers of successes than observed were not relevant. In Fisher's (1945) view, "... the existence of these less informative possibilities should not affect our judgment of significance based on the series actually observed.... The fact that such an unhelpful outcome as these might occur... is surely no reason for enhancing our judgment of significance in cases where it has not occurred; ... it is only the sampling distribution of samples of the same type that can supply a rational test of significance." Sprott (2000, Sec. 6.4.4) recently provided a similar argument.

An adaptation of the unconditional approach by Berger and Boos (1994) addresses this criticism somewhat. They took the supremum for the P-value over a confidence interval of values for the nuisance parameter π rather than over all possible values. Their unconditional P-value is

$$P = \sup_{\pi \in C_\gamma} P_\pi(T \geq t_o) + \gamma,$$

where C_γ is a $100(1-\gamma)\%$ confidence interval for π. Here, γ is taken to be very small (e.g., 0.001), and the test maintains the guaranteed upper bound on size.

Other arguments in favor of conditioning on both sets of marginal totals are that the conditional approach provides a simple way to eliminate nuisance parameters in a variety of problems (e.g., generalizing to other contingency table problems), and the margins contain little information about the association (Haber 1989; Yates 1984). Zhu and Reid (1994) noted that some information loss occurs in conditioning on the margins except when $\theta = 1$. Arguments against conditioning partly concern the increased discreteness that occurs. The few possible values for n_{11} make it difficult to obtain a small P-value. In repeated use with a nominal significance level, the actual type I error probability may be much smaller than the nominal value and the power may suffer. Finally, for inference about nonnull values (e.g, confidence intervals), we will see that the conditional approach applies only with the odds ratio and not other measures.

The conservatism problem is partly unavoidable. Statistics having discrete distributions are necessarily conservative in terms of achieving nominal significance levels. Because an unconditional test fixes only one margin, however, it has many more tables in the reference set for its sampling distribution. That distribution is less discrete, and a richer array of possible P-values occurs than with Fisher's exact test. An unconditional test tends to be less conservative and more powerful than Fisher's exact test. A disadvantage is that computations are very intensive for more complex problems, such as larger tables.

If a table truly has two independent binomial samples, the unconditional approach seems sensible. See Kempthorne (1979) for a cogent argument. The conditional approach is useful for other cases. In a randomized clinical trial a convenience sample of n subjects is randomly allocated to two treatments. The samples are not binomials, as they are not random samples from two populations of interest. One could focus on the sample alone and consider the probability of a result at least as extreme as observed if there truly is no treatment effect. For instance, out of all possible ways of choosing n_{1+} of the n subjects for treatment 1, for what proportion would n_{11} be at least as large as observed? Under the null hypothesis of no treatment effect, the same overall response distribution (n_{+1}, n_{+2}) of successes and failures occurs regardless of the allocation of subjects to treatments. Thus, the column margin is also naturally fixed. This argument leads to hypergeometric null probabilities and Fisher's exact test (Greenland 1981). This argument does not extend, however, to nonnull effect values and hence to confidence intervals.

When both sets of marginal totals are naturally fixed, such as in Table 3.8, the high degree of discreteness is unavoidable and Fisher's exact test is the best procedure. Regardless of which margins are naturally fixed, using the mid-P-value helps reduce conservative effects of discreteness.

3.5.7 Derivation of Exact Conditional Distribution*

We now show how the conditional test for independence yields the hypergeometric distribution. We do this for $I \times J$ tables, since we next discuss

extensions of Fisher's exact test for them. We assume independent multinomial sampling within rows, as often applies in comparing I treatment groups. Then row totals $\{n_{i+}\}$ are fixed, and we estimate the I conditional distributions $\{\pi_{j|i},\ j = 1, \ldots, J\}$. Under H_0: independence, $\pi_{j|1} = \pi_{j|2} = \cdots = \pi_{j|I} = \pi_{+j}$, for $j = 1, \ldots, J$. The product of the I multinomial probability functions then simplifies to

$$\prod_i \left(\frac{n_{i+}!}{\prod_j n_{ij}!} \prod_j \pi_{j|i}^{n_{ij}} \right) = \frac{(\prod_i n_{i+}!)(\prod_j \pi_{+j}^{n_{+j}})}{\prod_i \prod_j n_{ij}!}. \tag{3.17}$$

This distribution for $\{n_{ij}\}$ depends on $\{\pi_{+j}\}$. These are nuisance parameters, since they do not describe the association. Fisher introduced the standard way of eliminating nuisance parameters, by conditioning on their sufficient statistics. From the definition of sufficiency, the resulting conditional distribution does not depend on those parameters.

The contribution of $\{\pi_{+j}\}$ to the product multinomial distribution (3.17) depends on the data only through $\{n_{+j}\}$, which are their sufficient statistics. The $\{n_{+j}\}$ have the multinomial $(n, \{\pi_{+j}\})$ distribution, namely

$$\frac{n!}{\prod_j n_{+j}!} \prod_j \pi_{+j}^{n_{+j}}. \tag{3.18}$$

The joint probability function of $\{n_{ij}\}$ and $\{n_{+j}\}$ is identical to the probability function of $\{n_{ij}\}$, since $\{n_{ij}\}$ determines $\{n_{+j}\}$. Thus, the probability function of $\{n_{ij}\}$, conditional on $\{n_{+j}\}$, equals the probability function (3.17) of $\{n_{ij}\}$ divided by the probability function (3.18) evaluated at $\{n_{+j}\}$, or

$$\frac{(\prod_i n_{i+}!)(\prod_j n_{+j}!)}{n! \prod_i \prod_j n_{ij}!}. \tag{3.19}$$

This is the *multiple hypergeometric* distribution. It applies to the set of $\{n_{ij}\}$ having the same $\{n_{i+}\}$ and $\{n_{+j}\}$ as the observed table. For 2×2 tables, it is the hypergeometric distribution (3.16).

When a table has a single multinomial sample, the unknown parameters are $\{\pi_{ij}\}$. For testing independence ($\pi_{ij} = \pi_{i+}\pi_{+j}$ all i and j), distribution (3.19) results from conditioning on the row and column totals. These are sufficient statistics for $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$, which determine the null distribution. For either sampling model, both sets of margins are fixed after the conditioning. The end result (3.19) does not depend on unknown parameters and thus permits exact probability calculations.

3.5.8 Exact Tests of Independence for $I \times J$ Tables*

Exact tests for $I \times J$ tables utilize the multiple hypergeometric distribution. Freeman and Halton (1951) defined the P-value as the probability of the set

TABLE 3.9 Example for Exact Conditional Test

	Smoking Level (cigarettes/day)		
	0	1–24	> 25
Control	25	25	12
Myocardial infarction	0	1	3

Source: Reprinted with permission, based on Table 5 in S. Shapiro et al., *Lancet* 743–746 (1979).

of tables with the given margins that are no more likely to occur than the table observed. Other exact tests order the tables using a statistic describing distance from H_0. Yates (1934) used X^2. The P-value is then the null value of $P(X^2 \geq X_o^2)$ for observed value X_o^2. When classifications have ordered categories, an ordinal statistic is more relevant. For the alternative hypothesis of a positive association, we could use $P(T \geq t_o)$, where T is the correlation or gamma and where t_o denotes its observed value.

We illustrate an exact test for ordered categories with Table 3.9, which cross-classifies level of smoking and myocardial infarction for a sample of young women in a case–control study. The second row contains small counts, and large-sample tests may be inappropriate. Given the marginal counts, the only table having greater evidence of positive association between smoking and myocardial infarction has counts (25,26,11) for row 1 and (0,0,4) in row 2. Conditional on both sets of margins, the null probability of the observed table and this more extreme table [based on formula (3.19)] equals 0.018. Although the sample contains only four myocardial infarction patients, evidence exists of a positive association. The evidence is stronger than using X^2, which ignores the ordering of categories. The exact $P(X^2 \geq X_o^2) = P(X^2 \geq 6.96) = 0.052$.

Special algorithms and software for computing exact tests for $I \times J$ tables are widely available (e.g., Mehta and Patel 1983; see also Appendix A). We recommend these tests when asymptotic approximations may be invalid. Computing time increases exponentially as n, I, or J increase. However, one can use Monte Carlo to sample randomly from the set of tables with the given margins. The estimated P-value is then the sample proportion of tables having test statistic value at least as large as the value observed.

As I and/or J increase, the number of possible values for any test statistic T tends to increase. Thus, the conservativeness issue for conditional tests becomes less problematic.

3.6 SMALL-SAMPLE CONFIDENCE INTERVALS FOR 2 × 2 TABLES*

Small-sample methods also apply to estimation. Exact distributions depending only on the parameter of interest result from the same arguments. These

distributions are the basis of confidence intervals for measures such as the odds ratio.

3.6.1 Small-Sample Inference for the Odds Ratio

For multinomial sampling, the distribution of $\{n_{ij}\}$ depends on n and cell probabilities $\{\pi_{ij}\}$. For 2×2 tables, the odds ratio is

$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{\pi_{11}(1 - \pi_{1+} - \pi_{+1} + \pi_{11})}{(\pi_{1+} - \pi_{11})(\pi_{+1} - \pi_{11})}.$$

Hence, π_{11} is a function of θ and $\{\pi_{1+}, \pi_{+1}\}$. The same argument applies to any π_{ij}, so the multinomial distribution of $\{n_{ij}\}$ can use parameters $\{\theta, \pi_{1+}, \pi_{+1}\}$. Conditional on $\{n_{1+}, n_{+1}\}$, the distribution of $\{n_{ij}\}$ depends only on θ. Since n_{11} determines all other cell counts, given the marginal totals, the conditional distribution of $\{n_{ij}\}$ is specified by some function $P(n_{11} = t) = f(t; n_{1+}, n_{+1}, n, \theta)$. This distribution (Fisher 1935c) is the *noncentral hypergeometric*,

$$f(t; n_{1+}, n_{+1}, n, \theta) = \frac{\binom{n_{1+}}{t}\binom{n - n_{1+}}{n_{+1} - t}\theta^t}{\sum_{u=m_-}^{m_+}\binom{n_{1+}}{u}\binom{n - n_{1+}}{n_{+1} - u}\theta^u} \tag{3.20}$$

for $m_- \le t \le m_+$.

A confidence interval for θ results from inverting the test of H_0: $\theta = \theta_0$, having observed $n_{11} = t_o$. For H_a: $\theta > \theta_0$, the P-value is

$$P = \sum_{t \ge t_o} f(t; n_{1+}, n_{+1}, n, \theta_0).$$

For testing against H_0: $\theta < \theta_0$,

$$P = \sum_{t \le t_o} f(t; n_{1+}, n_{+1}, n, \theta_0).$$

When $\theta_0 = 1$, these are one-sided Fisher's exact tests. Cornfield (1956) constructed a confidence interval using the *tail method*. The lower endpoint is θ_0 for which $P = \alpha/2$ in testing against H_a: $\theta > \theta_0$. The upper endpoint is θ_0 for which $P = \alpha/2$ for H_a: $\theta < \theta_0$. The interval is the set of θ_0 for which both one-sided P-values $\ge \alpha/2$.

As in Fisher's exact test, the conditional approach to interval estimation is necessarily conservative because of discreteness. The actual confidence coefficient, defined as the infimum of the coverage probabilities for all possible θ, has the nominal confidence level as a lower bound. Less conservative

behavior and shorter intervals result from inverting a single two-sided test rather than inverting two one-sided tests (Agresti and Min 2001; Baptista and Pike 1977). An alternative approach with independent binomial samples inverts nonnull unconditional small-sample tests. Because of the reduced discreteness, such intervals are also usually shorter.

The *conditional ML estimate* of θ is the value of θ that maximizes probability (3.20). Differentiating the log likelihood with respect to θ shows that this estimate satisfies the equation $n_{11} = E(n_{11})$ in θ, where the expectation refers to distribution (3.20). This equation has a unique solution $\hat{\theta}$ and is solved using iterative methods (Cornfield 1956). This estimator differs from the *unconditional ML estimator* $\hat{\theta} = n_{11}n_{22}/n_{12}n_{21}$, which uses the ML estimates of $\{\pi_{ij}\}$ for the multinomial distribution of $\{n_{ij}\}$. Using statistical software, we can calculate conditional ML estimates and small-sample confidence intervals for odds ratios (e.g., for SAS, see Table A.2).

3.6.2 Tea Tasting Example

We illustrate with Table 3.8 from Fisher's tea-tasting experiment. The conditional ML estimate of θ is 6.4. Software provides the Cornfield tail-method interval (0.2, 626.2) with confidence coefficient guaranteed ≥ 0.95. Not surprisingly, it is very wide because of the small sample. Inverting a family of two-sided "exact" conditional score tests gives a more precise interval, (0.3, 306.2). The unconditional approach is not appropriate here because of the sampling design. [If the table were two binomial samples, that approach gives interval (0.4, 234.4) by inverting "exact" unconditional score tests.]

3.6.3 Impact of Discreteness on Exact Confidence Intervals

Small-sample inference is "exact" in the sense that the conditional distribution is free of nuisance parameters. Confidence intervals and tests use exact probability calculations rather than approximate ones. However, their operating characteristics are conservative because of discreteness.

Large-sample methods do not have the guarantee of bounds on error probabilities. They can be conservative or liberal, and thus their results can appear quite different from exact methods. For example, for the tea-tasting data (Table 3.8), the P-value for the Pearson chi-squared test equals 0.157, compared to 0.486 for the two-sided exact test. The 95% large-sample confidence interval (3.2) for the odds ratio is (0.4, 220.9), compared to Cornfield's exact interval of (0.2, 626.2). Normally, one would prefer an exact method over an approximate one. When the conditional distribution is highly discrete, however, the choice is not so obvious. Exact methods then can be quite conservative, especially with small samples.

For highly discrete data, it seems sensible to use adjustments of exact methods based on the mid-P-value. Confidence intervals with the conditional approach then invert hypergeometric tests of $\theta = \theta_0$ using the mid-P-value. Although not guaranteed to have error probabilities no greater than the

nominal level, this method usually comes closer than the exact method to the desired level. Compared to large-sample methods, it has the advantage of working well as the degree of discreteness diminishes, since it then is essentially the same as the corresponding exact method using an ordinary *P*-value.

Inference based on the mid-*P*-value compromises between the conservativeness of exact methods and the uncertain adequacy of large-sample methods. For interval estimation of the odds ratio, this method tends to be a bit conservative, but for small samples can yield much shorter intervals than the Cornfield exact interval. For the tea-tasting data, for instance, the 95% confidence interval based on inverting two one-sided hypergeometric tests using the mid-*P*-value is (0.31, 309), compared to the Cornfield interval of (0.21, 626).

3.6.4 Small-Sample Inference for Difference of Proportions

The conditional approach to eliminating nuisance parameters works when those parameters have sufficient statistics. However, we'll see (Section 6.7.9) that reduced sufficient statistics occur only for certain models. For binary data, such models must have odds ratios as parameters. For 2×2 tables, the conditional approach cannot yield confidence intervals for differences or ratios of proportions. The unconditional approach is more complex but does not require sufficient statistics. We used it in Section 3.5.5 for testing $\pi_1 - \pi_2 = 0$ with independent binomial samples.

A small-sample confidence interval inverts the corresponding unconditional test of H_0: $\pi_1 - \pi_2 = \delta_0$, for any fixed $-1 < \delta_0 < 1$. The probability function for the table is the product of $\text{bin}(n_1, \pi_1)$ and $\text{bin}(n_2, \pi_2)$ mass functions. One can express this in terms of $\delta = \pi_1 - \pi_2$ and a nuisance parameter λ. For instance, if $\lambda = \pi_1 + \pi_2$, one substitutes $\pi_1 = (\lambda + \delta)/2$ and $\pi_2 = (\lambda - \delta)/2$. For $\delta = \delta_0$ and a fixed value of λ, one then uses this binomial product to calculate the probability that the test statistic is at least as large as observed. The *P*-value is the supremum of such probabilities calculated over all possible values for λ. This provides a family of tests for the various values of δ_0. The confidence interval for $\pi_1 - \pi_2$ is the set of δ_0 for which this *P*-value exceeds α.

This approach can be quite conservative. For details regarding various test statistics, see Agresti and Min (2001), Coe and Tamhane (1993), Santner and Snell (1980), and Santner and Yamagami (1993). It is better to invert a single two-sided test, as in Coe and Tamhane (1993), than to invert two separate one-sided tests.

3.7 EXTENSIONS FOR MULTIWAY TABLES AND NONTABULATED RESPONSES

The methods of this chapter extend to multiway contingency tables. For instance, tests of independence for two-way tables extend to tests of condi-

tional independence in three-way tables. In future chapters we present such methods with models that provide a basis for defining relevant parameters and their statistical inferences. The methods then apply in a greater variety of situations, such as when some explanatory variables are continuous rather than categorical.

3.7.1 Categorical Data Need Not Be Contingency Tables

Examples so far have presented categorical data in the format of contingency tables. However, this book has broader focus than contingency table analysis. Models for categorical response variables can have continuous as well as categorical explanatory variables. Even when all or most variables are categorical, source data files are not usually contingency tables but have the form of a line of data for each subject. The first three lines in a data file containing responses of a survey of subjects measuring gender, race, education (1 = less than high school, 2 = high school or some college, 3 = college graduate), and opinion about homosexuality (1 = tolerant, 2 = homophobic) might be:

subject	gender	race	education	opinion
1	f	w	2	1
2	m	b	3	1
3	m	w	1	2

Software can read data files of this type and then conduct analyses that may involve forming contingency tables.

In the next chapter we introduce the modeling framework used in the rest of the book. All the methods that we've studied in this chapter result from inferences for parameters in simple versions of these models.

NOTES

Section 3.1: Confidence Intervals for Association Parameters

3.1. Adaptations of Woolf's interval (3.2) for $\log\theta$ to handle zero cell counts include Agresti (1999) and Gart (1966, 1971). Goodman (1964a) presented simultaneous confidence intervals for all odds ratios in an $I \times J$ table. Brown and Benedetti (1977) and Goodman and Kruskal (1963, 1972) provided standard errors for many association measures. Goodman and Kruskal (1963, 1972) extended (3.9) for independent multinomial sampling.

3.2. Agresti and Caffo (2000) showed that as in the single-sample case (Problem 1.24), the Wald interval (3.4) for $\pi_1 - \pi_2$ behaves much better after adding two pseudo-observations of each type (one of each type in each sample).

Section 3.2: Testing Independence in Two-Way Contingency Tables

3.3. For hypergeometric sampling, $\{\hat{\mu}_{ij}\}$ in tests of independence are *exact* (rather than *estimated*) expected values. Specifically,

$$E(n_{11}) = \frac{n_{1+}n_{+1}}{n} \quad \text{and} \quad \text{var}(n_{11}) = \frac{n_{1+}n_{+1}n_{2+}n_{+2}}{n^2(n-1)}.$$

Haldane (1940) derived $E(X^2) = (I-1)(J-1)n/(n-1)$ and a complex formula for $\text{var}(X^2)$; Dawson (1954) provided a simplified expression. Lewis et al. (1984) derived the third central moment. Watson (1959) showed that the conditional distribution of X^2 also has the limiting chi-squared distribution.

3.4. Diaconis and Efron (1985) presented inference based on a uniform distribution over all possible tables of the same I, J, and n; their *volume test* considers the proportion of such tables having $X^2 \le X_o^2$.

3.5. Specialized methods are necessary for complex sampling designs. Sequential methods are useful in biomedical applications (Jennison and Turnbull 2000, Chap. 12). Social science applications often incorporate clustering and/or stratification. LaVange et al. (2001) and Rao and Thomas (1988) surveyed analyses of categorical data for complex sampling methods. Gleser and Moore (1985) showed that positive dependence causes null distributions of Pearson statistics to stochastically increase. See also Bedrick (1983), Clogg and Eliason (1987), Fay (1985), Holt et al. (1980), Koehler and Wilson (1986), Rao and Scott (1987), Scott and Wild (2001), Shuster and Downing (1976), Tavaré and Altham (1983), and methods of Chapter 12.

Other modifications are necessary when some data are missing. Watson (1956) was perhaps the first to study this. Lipsitz and Fitzmaurice (1996) derived score tests of independence and conditional independence for contingency tables, assuming ignorable nonresponse, and showed that the test statistics have the usual asymptotic chi-squared null distributions. See Schafer (1997, Chap. 7) for a survey of methods.

Section 3.4: Two-Way Tables with Ordered Classifications

3.6. Bhapkar (1968) and Yates (1948) proposed statistics similar to M^2 and also proposed statistics for singly-ordered tables. Graubard and Korn (1987) listed 14 tests for $2 \times J$ tables that utilize a correlation-type statistic. See also Nair (1987) and Williams (1952). Cohen and Sackrowitz (1991, 1992) evaluated decision-theoretic aspects, such as admissibility, of tests based on gamma and local log odds ratios. Rayner and Best (2001) considered nonparametrics methods in a contingency table format.

Section 3.5: Small-Sample Tests of Independence

3.7. Yates (1934) mentioned that Fisher suggested the hypergeometric to him for an exact test. He proposed a continuity-corrected version of X^2,

$$X_c^2 = \sum\sum \frac{\left(|n_{ij} - \hat{\mu}_{ij}| - 0.5\right)^2}{\hat{\mu}_{ij}},$$

to approximate the exact test. Haber (1980, 1982), Plackett (1964), and Yates (1984) discussed its appropriateness. Since software now makes Fisher's exact test feasible even with large samples, this correction is no longer needed.

3.8. The UMPU property of Fisher's exact test follows from conditioning on a sufficient statistic that is complete and has distribution in the exponential family (Lehmann 1986, Secs. 4.5–4.7). Fleiss (1981), Gail and Gart (1973), and Suissa and Shuster (1985) studied sample size for obtaining fixed power in Fisher's test. The controversy over conditioning includes Barnard (1945, 1947, 1949, 1979), Berkson (1978), Fisher (1956), Howard (1998), Kempthorne (1979), Lloyd (1988a), Pearson (1947), Rice (1988), Routledge (1992), Suissa and Shuster (1984, 1985), and Yates (1984). Yates and discussants also addressed the choice of two-sided P-value. Discussion of unconditional methods includes Chan (1998), Martín Andrés and Silva Mato (1994), and Røhmel and Mansmann (1999). Altham (1969) and Howard (1998) discussed Bayesian analyses for 2×2 tables (see Section 15.2.3). Agresti (1992, 2001) surveyed small-sample methods.

3.9. For discussion of inference using the mid-P-value, see Berry and Armitage (1995), Hirji (1991), Hwang and Wells (2002), Hwang and Yang (2001), Mehta and Walsh (1992), and Routledge (1994). Similar benefits can accrue from alternative proposed P-values. One approach, useful when several tables have the same value for a test statistic, uses the table probability to create a more finely partitioned sample space; for tables having the observed test statistic value, only those contribute to the P-value that are no more likely than the observed table (Cohen and Sackrowitz 1992; Kim and Agresti 1995). This depends on more than the sufficient statistic, and in some cases a Rao–Blackwellized version is the mid-P-value (Hwang and Wells 2002). Ordinary P-values obtained with higher-order asymptotic methods without continuity corrections for discreteness yield performance similar to that of the mid-P-value (Pierce and Peters 1999; Strawderman and Wells 1998).

3.10. For exact treatment of $I \times J$ tables, see Mehta and Patel (1983). For ordered categories, see also Agresti et al. (1990). For Monte Carlo estimation of exact P-values, see Agresti et al. (1979), Booth and Butler (1999), Diaconis and Sturmfels (1998), Forster et al. (1996), Mehta et al. (1988), and Patefield (1982). Gail and Mantel (1977) and Good (1976) gave approximate formulas for the number of tables having certain fixed margins. Freidlin and Gastwirth (1999) extended the unconditional approach to a test for trend in $I \times 2$ tables and a test of conditional independence with several 2×2 tables.

Section 3.6: Small-Sample Confidence Intervals for 2 × 2 Tables

3.11. Suppose that (θ, λ) has minimal sufficient statistic (T, U), where λ is a nuisance parameter. Cox and Hinkley (1974, p. 35) defined U to be *ancillary* for θ if its distribution depends only on λ, and the distribution of T given U depends only on θ. For 2×2 tables with odds ratio θ and $\lambda = (\pi_{1+}, \pi_{+1})$, let $T = n_{11}$ and $U = (n_{1+}, n_{+1})$. Then U is not ancillary, because its distribution depends on θ as well as λ. Using a definition due to Godambe, Bhapkar (1989) referred to the marginals U as *partial ancillary* for θ. This means that the distribution of the data, given U, depends only on θ, and that for fixed θ, the family of distributions of U for various λ is complete. Liang (1984) gave an alternative definition referring to conditional and unconditional inference being equally efficient.

PROBLEMS

Applications

3.1 Refer to Table 2.9. Construct and interpret a 95% confidence interval for the population (**a**) odds ratio, (**b**) difference of proportions, and (**c**) relative risk between seat-belt use and type of injury.

3.2 Refer to Table 2.5 on lung cancer and smoking. Construct a confidence interval for a relevant measure of association. Interpret.

3.3 In professional basketball games during 1980–1982, when Larry Bird of the Boston Celtics shot a pair of free throws, 5 times he missed both, 251 times he made both, 34 times he made only the first, and 48 times he made only the second (Wardrop 1995). Is it plausible that the successive free throws are independent?

3.4 Refer to Table 3.10.

a. Using X^2 and G^2, test the hypothesis of independence between party identification and race. Report the P-values and interpret.

b. Use residuals to describe the evidence of association.

c. Partition chi-squared into components regarding the choice between Democrat and Independent and between these two combined and Republican. Interpret.

d. Summarize association by constructing a 95% confidence interval for the odds ratio between race and whether a Democrat or Republican. Interpret.

TABLE 3.10 Data for Problem 3.4

	Party Identification		
Race	Democrat	Independent	Republican
Black	103	15	11
White	341	105	405

Source: 1991 General Social Survey, National Opinion Research Center.

3.5 Refer to Table 3.10. In the same survey, gender was cross-classified with party identification. Table 3.11 shows some results. Explain how to interpret all the results on this printout.

3.6 In a study of the relationship between stage of breast cancer at diagnosis (local or advanced) and a woman's living arrangement, of 144 women living alone, 41.0% had an advanced case; of 209 living with spouse, 52.2% were advanced; of 89 living with others, 59.6% were advanced. The authors reported the P-value for the relationship as 0.02 (D. J. Moritz and W. A. Satariano, *J. Clin. Epidemiol.* **46**: 443–454, 1993). Reconstruct the analysis performed to obtain this P-value.

TABLE 3.11 Results for Problem 3.5

Frequency Expected	dem	indep	repub
female	279 261.42	73 70.653	225 244.93
male	165 182.58	47 49.347	191 171.07

Statistic	DF	Value	Prob
Chi-Square	2	7.0095	0.0301
Likelihood Ratio Chi-Square	2	7.0026	0.0302

Observ	Resraw	Reschi	StReschi	Observ	Resraw	Reschi	StReschi
1	17.584	1.088	2.293	4	−17.584	−1.301	−2.293
2	2.347	0.279	0.465	5	−2.347	−0.334	−0.464
3	−19.931	−1.274	−2.618	6	19.931	1.524	2.618

3.7 Refer to Table 2.1. Partition G^2 for testing whether the incidence of heart attacks is independent of aspirin intake into two components. Interpret.

3.8 *Project Blue Book: Analysis of Reports of Unidentified Aerial Objects* was published by the U.S. Air Force (Air Technical Intelligence Center at Wright-Patterson Air Force Base) in May 1955 to analyze reports of unidentified flying objects (UFOs). In its Table II, the report classified 1765 sightings later regarded as known objects and 434 sightings later regarded as unknown, according to the object color (nine categories). The report states: "The chi-square test is applicable only to distributions which have the same number of elements," so the investigators multiplied all counts in the known category by (434/1765), so each row has 434 observations, before computing X^2. They reported $X^2 = 26.15$ with df = 8. Explain why this is incorrect. What should X^2 equal? (*Hint:* For their adjusted table, first show that the contribution to X^2 is the same for each cell in a column, and then show the effect on those contributions of multiplying each count in one row by a constant.)

3.9 Table 3.12 classifies a sample of psychiatric patients by their diagnosis and by whether their treatment prescribed drugs.

a. Obtain standardized Pearson residuals for independence, and interpret.

b. Partition chi-squared into three components to describe differences and similarities among the diagnoses, by comparing **(i)** the first two rows, **(ii)** the third and fourth rows, and **(iii)** the last row to the first and second rows combined and the third and fourth rows combined.

TABLE 3.12 Data for Problem 3.9

Diagnosis	Drugs	No Drugs
Schizophrenia	105	8
Affective disorder	12	2
Neurosis	18	19
Personality disorder	47	52
Special symptoms	0	13

Source: Reprinted with permission from E. Helmes and G. C. Fekken, *J. Clin. Psychol.* **42:** 569–576 (1986).

3.10 Refer to Table 7.8. For the combined data for the two genders, yielding a single 4×4 table, $X^2 = 11.5$ ($P = 0.24$), whereas using row scores (3, 10, 20, 35) and column scores (1, 3, 4, 5), $M^2 = 7.04$ ($P = 0.008$). Explain why the results are so different.

3.11 A study on educational aspirations of high school students (S. Crysdale, *Internat. J. Compar. Sociol.* **16**: 19–36, 1975) measured aspirations with the scale (some high school, high school graduate, some college, college graduate). The student counts in these categories were (11, 52, 23, 22) when family income was low, (9, 44, 13, 10) when family income was middle, and (9, 41, 12, 27) when family income was high.

a. Test independence of educational aspirations and family income using X^2 or G^2. Explain the deficiency of this test for these data.

b. Find the standardized Pearson residuals. Do they suggest any association pattern?

c. Conduct an alternative test that may be more powerful. Interpret.

3.12 Refer to Table 8.15. Obtain a 95% confidence interval for gamma. Interpret the association between schooling and attitude toward abortion.

3.13 Table 3.13 shows the results of a retrospective study comparing radiation therapy with surgery in treating cancer of the larynx. The response

TABLE 3.13 Data for Problem 3.13

	Cancer Controlled	Cancer Not Controlled
Surgery	21	2
Radiation therapy	15	3

Source: Reprinted with permission from W. M. Mendenhall, R. R. Million, D. E. Sharkey, and N. J. Cassisi, *Internat. J. Radiat. Oncol. Biol. Phys.* **10:** 357–363 (1984), Pergamon Press plc.

TABLE 3.14 SAS Output for Problem 3.13

```
          Fisher's Exact Test
---------------------------------
Cell (1,1) Frequency (F)              21
Left-sided Pr <= F                0.8947
Right-sided Pr >= F               0.3808
Table Probability (P)             0.2755
Two-sided Pr<= P                  0.6384
---------------------------------

Odds Ratio                        2.1000

Asymptotic Conf Limits:     95% Lower Conf Limit   0.3116
                            95% Upper Conf Limit  14.1523
Exact Conf Limits:          95% Lower Conf Limit   0.2089
                            95% Upper Conf Limit  27.5522
```

indicates whether the cancer was controlled for at least two years following treatment. Table 3.14 shows SAS output.

a. Report and interpret the P-value for Fisher's exact test with **(i)** H_a: $\theta > 1$, and **(ii)** H_a: $\theta \neq 1$. Explain how the P-values are calculated.

b. Interpret the confidence intervals for θ. Explain the difference between them and how they were calculated.

c. Find and interpret the one-sided mid-P-value. Give advantages and disadvantages of this type of P-value.

3.14 A study considered the effect of prednisolone on severe hypercalcaemia in women with metastatic breast cancer (B. Kristensen et al., *J. Intern. Med.* **232**: 237–245, 1992). Of 30 patients, 15 were randomly selected to receive prednisolone. The other 15 formed a control group. Normalization in their level of serum-ionized calcium was achieved by 7 of the treated patients and none of the control group. Analyze whether results were significantly better for treatment than for control. Interpret.

3.15 For Problem 3.14, obtain a 95% confidence interval for the odds ratio using **(a)** the Woolf (i.e., Wald) interval, **(b)** Cornfield's "exact" approach, **(c)** the profile likelihood. In each case, note the effect of the zero cell count. Summarize advantages and disadvantages of each approach.

3.16 Refer to the tea-tasting data (Table 3.8). Construct the null distributions of the ordinary P-value and the mid-P-value for Fisher's exact test with H_a: $\theta > 1$. Find and compare their expected values.

3.17 Consider a 3×3 table having entries, by row, of (4, 2, 0 / 2, 2, 2 / 0, 2, 4). Conduct an exact test of independence, using X^2. Assuming ordered rows and columns and using equally spaced scores, conduct an ordinal exact test. Explain why results differ so much.

3.18 An advertisement by Schering Corp. in 1999 for the allergy drug Claritin mentioned that in a pediatric randomized clinical trial, symptoms of nervousness were shown by 4 of 188 patients on loratadine (Claritin), 2 of 262 patients taking placebo, and 2 of 170 patients on choropheniramine. In each part below, explain which method you used, and why.

a. Is there inferential evidence that nervousness depends on drug?

b. For the Claritin and placebo groups, construct and interpret a 95% confidence interval for the **(i)** odds ratio and **(ii)** difference of proportions suffering nervousness.

3.19 Refer to Problem 2.19 on sexual fun. Analyze these data. Present a short report summarizing results and interpretations.

Theory and Methods

3.20 Is $\hat{\theta}$ the midpoint of large- and small-sample confidence intervals for θ? Why or why not?

3.21 For comparing two binomial samples, show that the standard error (3.1) of a log odds ratio increases as the absolute difference of proportions of successes and failures for a given sample increases.

3.22 Using the delta method, show that the Wald confidence interval for the logit of a binomial parameter π is

$$\log[\hat{\pi}/(1-\hat{\pi})] \pm z_{\alpha/2}/\sqrt{n\hat{\pi}(1-\hat{\pi})}\,.$$

Explain how to use this interval to obtain one for π itself. [Newcombe (2001) noted that the sample logit is also the midpoint of the score interval for π, on the logit scale. He showed that this logit interval contains the score interval.]

3.23 For two parameters, a confidence interval for $\theta_1 - \theta_2$ based on single-sample estimate $\hat{\theta}_i$ and interval (ℓ_i, u_i) for θ_i, $i = 1, 2$, is

$$\left(\hat{\theta}_1 - \hat{\theta}_2 - \sqrt{\left(\hat{\theta}_1 - \ell_1\right)^2 + \left(u_2 - \hat{\theta}_2\right)^2},\quad \hat{\theta}_1 - \hat{\theta}_2 + \sqrt{\left(u_1 - \hat{\theta}_1\right)^2 + \left(\hat{\theta}_2 - \ell_2\right)^2}\right).$$

Newcombe (1998b) proposed an interval for $\pi_1 - \pi_2$ using the score interval (ℓ_i, u_i) for π_i that performs much better than the Wald interval (3.4). It is $(\hat{\pi}_1 - \hat{\pi}_2 - z_{\alpha/2}s_L,\ \hat{\pi}_1 - \hat{\pi}_2 + z_{\alpha/2}s_U)$, with

$$s_L = \sqrt{\frac{\ell_1(1-\ell_1)}{n_1} + \frac{u_2(1-u_2)}{n_2}}, \qquad s_U = \sqrt{\frac{u_1(1-u_1)}{n_1} + \frac{\ell_2(1-\ell_2)}{n_2}}.$$

Show that it has the general form above of an interval for $\theta_1 - \theta_2$.

3.24 For multinomial sampling, use the asymptotic variance of $\log \hat{\theta}$ to show that for Yule's Q (Problem 3.26) the asymptotic variance of $\sqrt{n}(\hat{Q} - Q)$ is $\sigma^2 = (\sum_i \sum_j \pi_{ij}^{-1})(1 - Q^2)^2/4$ (Yule 1900, 1912).

3.25 Refer to Problem 2.23. For multinomial sampling, show how to obtain a confidence interval for AR by first finding one for $\log(1 - AR)$ (Fleiss 1981, p. 76).

3.26 For multinomial probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$ with a contingency table of arbitrary dimensions, suppose that a measure $g(\boldsymbol{\pi}) = \nu/\delta$. Show that the asymptotic variance of $\sqrt{n}[g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi})]$ is $\sigma^2 = [\sum_i \pi_i \eta_i^2 - (\sum_i \pi_i \eta_i)^2]/\delta^4$, where $\eta_i = \delta(\partial\nu/\partial\pi_i) - \nu(\partial\delta/\partial\pi_i)$ (Goodman and Kruskal, 1972).

3.27 For ordinal variables, consider gamma (2.14). Let

$$\pi_{ij}^{(c)} = \sum_{a<i}\sum_{b<j} \pi_{ab} + \sum_{a>i}\sum_{b>j} \pi_{ab}, \qquad \pi_{ij}^{(d)} = \sum_{a<i}\sum_{b>j} \pi_{ab} + \sum_{a>i}\sum_{b<j} \pi_{ab},$$

where i and j are fixed in the summations. Show that $\Pi_c = \sum_i \sum_j \pi_{ij}\pi_{ij}^{(c)}$ and $\Pi_d = \sum_i \sum_j \pi_{ij}\pi_{ij}^{(d)}$. Use the delta method to show that the large-sample normality (3.9) applies for $\hat{\gamma}$, with (Goodman and Kruskal 1963)

$$\phi_{ij} = 4\left[\Pi_d \pi_{ij}^{(c)} - \Pi_c \pi_{ij}^{(d)}\right]/(\Pi_c + \Pi_d)^2, \qquad \sum_i \sum_j \pi_{ij}\phi_{ij} = 0,$$

$$\sigma^2 = \frac{16}{(\Pi_c + \Pi_d)^4} \sum_i \sum_j \pi_{ij}\left[\Pi_d \pi_{ij}^{(c)} - \Pi_c \pi_{ij}^{(d)}\right]^2.$$

3.28 An $I \times J$ table has ordered columns and unordered rows. *Ridits* (Bross 1958) are data-based column scores. The jth sample ridit is the average cumulative proportion within category j,

$$\hat{r}_j = \sum_{k=1}^{j-1} p_{+k} + \left(\frac{1}{2}\right) p_{+j}.$$

The sample mean ridit in row i is $\hat{R}_i = \sum_j \hat{r}_j p_{j|i}$. Show that $\sum_j p_{+j}\hat{r}_j = 0.50$ and $\sum_i p_{i+}\hat{R}_i = 0.50$. [For ridit analyses, see Agresti (1984, Secs. 9.3 and 10.2), Bross (1958), Fleiss (1981, Sec. 9.4), and Landis et al. (1978).]

3.29 Show that $X^2 = n\sum\sum(p_{ij} - p_{i+}p_{+j})^2/p_{i+}p_{+j}$. Thus, X^2 can be large when n is large, regardless of whether the association is practically important. Explain why this test, like other tests, simply indicates the degree of evidence against H_0 and does not describe strength of association. ("Like fire, the chi-square test is an excellent servant and a bad master," Sir Austin Bradford Hill, *Proc. Roy. Soc. Med.* **58**: 295–300, 1965.)

3.30 For testing H_0: $\pi_1 = \pi_2$ using independent binomial variates y_1 and y_2 with n_1 and n_2 trials, the score statistic is

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1 - \hat{\pi})(1/n_1 + 1/n_2)}},$$

where $\hat{\pi} = (y_1 + y_2)/(n_1 + n_2)$ is the *pooled estimate* of $\pi_1 = \pi_2$ under H_0. Show that $z^2 = X^2$.

3.31 For a 2×2 table, consider H_0: $\pi_{11} = \theta^2$, $\pi_{12} = \pi_{21} = \theta(1 - \theta)$, $\pi_{22} = (1 - \theta)^2$.

a. Show that the marginal distributions are identical and that independence holds.

b. For a multinomial sample, under H_0 show that $\hat{\theta} = (p_{1+} + p_{+1})/2$.

c. Explain how to test H_0. Show that df = 2 for the test statistic.

d. Refer to Problem 3.3. Are Larry Bird's pairs of free throws plausibly independent *and* identically distributed?

3.32 For a 2×2 table, show that:

a. The four Pearson residuals may take different values.

b. All four standardized Pearson residuals have the same absolute value. (This is sensible, since df = 1.)

c. The square of each standardized Pearson residual equals X^2. [*Note:* $X^2 = n(n_{11}n_{22} - n_{12}n_{21})^2/(n_{1+}n_{2+}n_{+1}n_{+2})$ for 2×2 tables. See Mirkin (2001) for alternative X^2 formulas for $I \times J$ tables.]

3.33 For testing independence, show that $X^2 \leq n \min(I - 1, J - 1)$. Hence $V^2 = X^2/[n\min(I - 1, J - 1)]$ falls between 0 and 1 (Cramér 1946). For 2×2 tables, X^2/n is often called *phi-squared*; it equals Goodman and Kruskal's tau (Problem 2.38). Other measures based on X^2 include the *contingency coefficient* $[X^2/(X^2 + n)]^{1/2}$ (Pearson 1904).

3.34 For counts $\{n_i\}$, the *power divergence statistic* for testing goodness of fit (Cressie and Read 1984; Read and Cressie 1988) is

$$\frac{2}{\lambda(\lambda + 1)} \sum n_i \left[\left(n_i/\hat{\mu}_i \right)^{\lambda} - 1 \right] \qquad \text{for } -\infty < \lambda < \infty.$$

a. For $\lambda = 1$, show that this equals X^2.

b. As $\lambda \to 0$, show that it converges to G^2. [*Hint:* $\log t = \lim_{h \to 0} (t^h - 1)/h$.]

c. As $\lambda \to -1$, show that it converges to $2\Sigma \hat{\mu}_i \log(\hat{\mu}_i/n_i)$, the *minimum discrimination information* statistic (Gokhale and Kullback 1978).

d. For $\lambda = -2$, show that it equals $\Sigma(n_i - \hat{\mu}_i)^2/n_i$, the *Neyman modified chi-squared* statistic (Neyman 1949).

e. For $\lambda = -\frac{1}{2}$, show that it equals $4\Sigma(\sqrt{n_i} - \sqrt{\hat{\mu}_i})^2$, the *Freeman–Tukey* statistic (Freeman and Tukey 1950).

[Under regularity conditions, their asymptotic distributions are identical (see Drost et al. 1989). The chi-squared null approximation works best for λ near $\frac{2}{3}$.]

3.35 Use a partitioning argument to explain why G^2 for testing independence cannot increase after combining two rows (or two columns) of a contingency table. (*Hint:* Argue that G^2 for full table = G^2 for collapsed table + G^2 for table of the two rows that are combined in the collapsed table.)

3.36 Motivate partitioning (3.14) by showing that the multiple hypergeometric distribution (3.19) for $\{n_{ij}\}$ factors as the product of hypergeometric distributions for the separate component tables (Lancaster, 1949).

3.37 Explain why $\{n_{+j}\}$ are sufficient for $\{\pi_{+j}\}$ in (3.17).

3.38 Assume independence, and let $p_{ij} = n_{ij}/n$ and $\hat{\pi}_{ij} = p_{i+}p_{+j}$.

a. Show that p_{ij} and $\hat{\pi}_{ij}$ are unbiased for $\pi_{ij} = \pi_{i+}\pi_{+j}$.

b. Show that $\text{var}(p_{ij}) = \pi_{i+}\pi_{+j}(1 - \pi_{i+}\pi_{+j})/n$.

c. Using $E(p_{i+}p_{+j})^2 = E(p_{i+}^2)E(p_{+j}^2)$ and $E(p_{i+}^2) = \text{var}(p_{i+}) + [E(p_{i+})]^2$, show that

$$\text{var}(\hat{\pi}_{ij}) = \left\{\pi_{i+}\pi_{+j}\left[\pi_{i+}(1 - \pi_{+j}) + \pi_{+j}(1 - \pi_{i+})\right]\right\}\Big/n + \pi_{i+}(1 - \pi_{i+})\pi_{+j}(1 - \pi_{+j})/n^2.$$

d. As $n \to \infty$, show that $\lim \text{var}(\sqrt{n}\,\hat{\pi}_{ij}) \leq \lim \text{var}(\sqrt{n}\,p_{ij})$, with equality only if $\pi_{ij} = 1$ or 0. Hence, if the model holds or if it nearly holds, the model estimator is better than the sample proportion.

3.39 Show that the sample value of the uncertainty coefficient (2.13) satisfies $\hat{U} = -G^2/2n(\Sigma p_{+j} \log p_{+j})$. [Haberman (1982) gave its standard error.]

3.40 When a test statistic has a continuous distribution, the P-value has a null uniform distribution, $P(P\text{-value} \leq \alpha) = \alpha$ for $0 < \alpha < 1$. For Fisher's exact test, explain why under the null, $P(P\text{-value} \leq \alpha) \leq \alpha$ for $0 < \alpha < 1$. (*Hint:* $P(P\text{-value} \leq \alpha) = E[P(P\text{-value} \leq \alpha | n_{1+}, n_{+1}, n)]$.)

3.41 Refer to Note 3.3 about moments of the hypergeometric distribution (3.16). Letting $\rho = n_{+1}/n$, show that n_{11} has the same mean as a binomial random variable for n_{1+} trials with success probability ρ, and that it has its variance multiplied by a finite population correction factor $(n - n_{1+})/(n - 1)$. (The hypergeometric is similar to the binomial when n_{1+} is small compared to n.)

3.42 A contingency table for two independent binomial variables has counts (3, 0 / 0, 3) by row. For H_0: $\pi_1 = \pi_2$ and H_a: $\pi_1 > \pi_2$, show that the P-value equals $\frac{1}{64}$ for the exact unconditional test and $\frac{1}{20}$ for Fisher's

exact test. [For discussion of this example, see Little (1989), G. Barnard's remarks at the end of Yates (1984), and Sprott (2000, Sec. 6.4.4).]

3.43 Refer to Problem 3.42 and exact tests using X^2 with H_a: $\pi_1 \neq \pi_2$. Explain why the unconditional P-value, evaluated at $\pi = 0.5$, is related to Fisher conditional P-values for various tables by

$$P(X^2 \geq 6) = \sum_{k=0}^{6} P\big(X^2 \geq 6 | n_{+1} = k\big) P(n_{+1} = k) .$$

Thus, the unconditional P-value of $\frac{1}{32}$ is a weighted average of the Fisher P-value for the observed column margins and P-values of 0 corresponding to the impossibility of getting results as extreme as observed if other margins had occurred (i.e., $\frac{1}{32} = 0.10\left[\binom{6}{3}(1/2)^6\right]$). The Fisher quote in Section 3.5.6 gave his view about this.

3.44 Consider exact tests of independence, given the marginals, for the $I \times I$ table having $n_{ii} = 1$ for $i = 1, \ldots, I$, and $n_{ij} = 0$ otherwise. Show that **(a)** tests that order tables by their probabilities, X^2, or G^2 have P-value = 1.0, and **(b)** the one-sided test that orders tables by an ordinal statistic such as r or $C - D$ has P-value = $(1/I!)$.

3.45 A Monte Carlo scheme randomly samples M separate $I \times J$ tables having the observed margins to approximate $P_o = P(X^2 \geq X_o^2)$ for an exact test. Let $\hat{P}$ be the sample proportion of the M tables with $X^2 \geq X_o^2$. Show that $P(|\hat{P} - P_o| \leq B) = 1 - \alpha$ requires that $M \approx z_{\alpha/2}^2 P_o(1 - P_o)/B^2$.

3.46 Show that the conditional ML estimate of θ satisfies $n_{11} = E(n_{11})$ for distribution (3.18).

Categorical Data Analysis, Second Edition. Alan Agresti

ISBN: 0-471-36093-7

CHAPTER 9

Building and Extending Loglinear/Logit Models

In Chapters 5 through 7 we presented logistic regression models, which use the logit link for binomial or multinomial responses. In Chapter 8 we presented loglinear models for contingency tables, which use the log link for Poisson cell counts. Equivalences between them were discussed in Section 8.5.3. In this chapter we discuss building and extending these models with contingency tables.

In Section 9.1 we present graphs that show a model's association and conditional independence patterns. In Section 9.2 we discuss selection and comparison of loglinear models. Diagnostics for checking models, such as residuals, are presented in Section 9.3.

The loglinear models of Chapter 8 treat all variables as nominal. In Section 9.4 we present loglinear models of association between ordinal variables. In Sections 9.5 and 9.6 we present generalizations that replace fixed scores by parameters. In the final section we discuss complications that occur with sparse contingency tables.

9.1 ASSOCIATION GRAPHS AND COLLAPSIBILITY

A graphical representation for associations in loglinear models indicates the pairs of conditionally independent variables. This representation helps reveal implications of models. Our presentation derives partly from Darroch et al. (1980), who used mathematical graph theory to represent certain loglinear models (called *graphical models*) having a conditional independence structure.

9.1.1 Association Graphs

An *association graph* has a set of vertices, each vertex representing a variable. An edge connecting two variables represents a conditional association be-

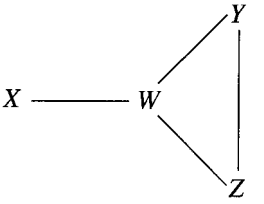


FIGURE 9.1 Association graph for model (WX, WY, WZ, YZ).

tween them. For instance, loglinear model (WX, WY, WZ, YZ) lacks XY and XZ terms. It assumes independence between X and Y and between X and Z, conditional on the remaining two variables. Figure 9.1 portrays this model's association graph. The four variables form the vertices. The four edges represent pairwise conditional associations. Edges do not connect X and Y or X and Z, the conditionally independent pairs.

Two loglinear models with the same pairwise associations have the same association graph. For instance, this association graph is also the one for model (WX, WYZ), which adds a three-factor WYZ interaction.

A *path* in an association graph is a sequence of edges leading from one variable to another. Two variables X and Y are said to be *separated* by a subset of variables if all paths connecting X and Y intersect that subset. For instance, in Figure 9.1, W separates X and Y, since any path connecting X and Y goes through W. The subset $\{W, Z\}$ also separates X and Y. A fundamental result states that two variables are conditionally independent given *any* subset of variables that separates them (Kreiner 1987; Whittaker 1990, p. 67). Thus, not only are X and Y conditionally independent given W and Z, but also given W alone. Similarly, X and Z are conditionally independent given W alone.

9.1.2 Collapsibility in Three-Way Contingency Tables

In Section 2.3.3 we showed that conditional associations in partial tables usually differ from marginal associations. Under certain *collapsibility conditions*, however, they are the same.

> For three-way tables, XY marginal and conditional odds ratios are identical if either Z and X are conditionally independent or if Z and Y are conditionally independent.

The conditions state that the variable treated as the control (Z) is conditionally independent of X or Y, or both. These conditions occur for loglinear models (XY, YZ) and (XY, XZ). Thus, the fitted XY odds ratio is identical in the partial tables and the marginal table for models with association graphs

$$X \text{ —— } Y \text{ —— } Z \quad \text{and} \quad Y \text{ —— } X \text{ —— } Z$$

or even simpler models, but not for the model with graph

$$X \text{ —— } Z \text{ —— } Y$$

in which an edge connects Z to both X and Y. The proof follows directly from the formulas for models (XY, YZ) and (XY, XZ) (Problem 9.26).

We illustrate for the student survey (Table 8.3) from Section 8.2.4, with A = alcohol use, C = cigarette use, and M = marijuana use. Model (AM, CM) specifies AC conditional independence, given M. It has association graph

$$A \text{ —— } M \text{ —— } C.$$

Consider the AM association. Since C is conditionally independent of A, the AM fitted conditional odds ratios are the same as the AM fitted marginal odds ratio collapsed over C. From Table 8.5, both equal 61.9. Similarly, the CM association is collapsible. The AC association is not, because M is conditionally dependent with both A and C in model (AM, CM). Thus, A and C may be marginally dependent, even though they are conditionally independent. In fact, from Table 8.5, the fitted AC marginal odds ratio for this model is 2.7.

For model (AC, AM, CM), no pair is conditionally independent. No collapsibility conditions are fulfilled. Table 8.5 showed that each pair has quite different fitted marginal and conditional associations for this model. When a model contains all two-factor effects, effects may change after collapsing over any variable.

9.1.3 Collapsibility and Logit Models

The collapsibility conditions apply also to logit models. For instance, suppose that a clinical trial studies the association between a binary treatment variable X $(x_1 = 1,\ x_2 = 0)$ and a binary response Y, using data from K centers (Z). The logit model

$$\text{logit}\big[P(Y = 1 \,|\, X = i, Z = k)\big] = \alpha + \beta x_i + \beta_k^Z$$

has the same treatment effect β for each center. Since this model corresponds to loglinear model (XY, XZ, YZ), this effect may differ after collapsing the $2 \times 2 \times K$ table over centers. The estimated XY conditional odds ratio, $\exp(\hat{\beta})$, typically differs from the sample odds ratio in the marginal 2×2 table.

Next, consider the simpler model that lacks center effects,

$$\text{logit}\big[P(Y = 1 \,|\, X = i, Z = k)\big] = \alpha + \beta x_i .$$

For a given treatment, the success probability is identical for each center. The model satisfies a collapsibility condition, because it states that Z is

conditionally independent of Y, given X. This logit model is equivalent to loglinear model (XY, XZ), for which the XY association is collapsible. So, when center effects are negligible and the simpler model fits nearly as well, the estimated treatment effect is approximately the marginal XY odds ratio.

9.1.4 Collapsibility and Association Graphs for Multiway Tables

Bishop et al. (1975, p. 47) provided a parametric collapsibility condition with multiway tables:

> Suppose that a model for a multiway table partitions variables into three mutually exclusive subsets, A, B, C, such that B separates A and C. After collapsing the table over the variables in C, parameters relating variables in A and parameters relating variables in A to variables in B are unchanged.

We illustrate using model (WX, WY, WZ, YZ) (Figure 9.1). Let $A = \{X\}$, $B = \{W\}$, and $C = \{Y, Z\}$. Since the XY and XZ terms do not appear, all parameters linking set A with set C equal zero, and B separates A and C. If we collapse over Y and Z, the WX association is unchanged. Next, identify $A = \{Y, Z\}$, $B = \{W\}$, $C = \{X\}$. Then, conditional associations among W, Y, and Z remain the same after collapsing over X.

This result also implies that when any variable is independent of all other variables, collapsing over it does not affect any other model terms. For instance, associations among W, X, and Y in model (WX, WY, XY, Z) are the same as in (WX, WY, XY).

When set B contains more than one variable, although parameter values are unchanged in collapsing over set C, the ML estimates of those parameters may differ slightly. A stronger collapsibility definition also requires that the estimates be identical. This condition of commutativity of fitting and collapsing holds if the model contains the highest-order term relating variables in B to each other. Asmussen and Edwards (1983) discussed this property, which relates to *decomposability* of tables (Note 8.2).

9.2 MODEL SELECTION AND COMPARISON

Strategies for selecting and comparing loglinear models are similar to those for logistic regression discussed in Section 6.1. A model should be complex enough to fit well but also relatively simple to interpret, smoothing rather than overfitting the data.

9.2.1 Considerations in Model Selection

The potentially useful models are usually a small subset of the possible models. A study designed to answer certain questions through confirmatory analyses may plan to compare models that differ only by the inclusion of certain terms. Also, models should recognize distinctions between response

and explanatory variables. The modeling process should concentrate on terms linking responses and terms linking explanatory variables to responses. The model should contain the most general interaction term relating the explanatory variables. From the likelihood equations, this has the effect of equating the fitted totals to the sample totals at combinations of their levels. This is natural, since one normally treats such totals as fixed. Related to this, certain marginal totals are often fixed by the sampling design. Any potential model should include those totals as sufficient statistics, so likelihood equations equate them to the fitted totals.

Consider Table 8.8 with I = automobile injury and S = seat-belt use as responses and G = gender and L = location as explanatory variables. Then we treat $\{n_{g+\ell+}\}$ as fixed at each combination for G and L. For example, 20,629 women had accidents in urban locations, so the fitted counts should have 20,629 women in urban locations. To ensure this, a loglinear model should contain the GL term, which implies from its likelihood equations that $\{\hat{\mu}_{g+\ell+} = n_{g+\ell+}\}$. Thus, the model should be at least as complex as (GL, S, I) and focus on the effects of G and L on S and I as well as the SI association.

If S is also explanatory and only I is a response, $\{n_{g+\ell s}\}$ should be fixed. With a single categorical response, relevant loglinear models correspond to logit models for that response. One should then use logit rather than loglinear models, when the main focus is describing effects on that response.

For exploratory studies, a search among potential models may provide clues about associations and interactions. One approach first fits the model having single-factor terms, then the model having two-factor and single-factor terms, then the model having three-factor and lower terms, and so on. Fitting such models often reveals a restricted range of good-fitting models. In Section 8.4.2 we used this strategy with the automobile injury data set. Automatic search mechanisms among possible models, such as backward elimination, may also be useful but should be used with care and skepticism. Such a strategy need not yield a meaningful model.

9.2.2 Model Building for the Dayton Student Survey

In Sections 8.2.4 and 8.3.2 we analyzed the use of alcohol (A), cigarettes (C), and marijuana (M) by a sample of high school seniors. The study also classified students by gender (G) and race (R). Table 9.1 shows the five-dimensional contingency table. In selecting a model, we treat A, C, and M as responses and G and R as explanatory. Thus, a model should contain the GR term, which forces the GR fitted marginal totals to equal the sample marginal totals

Table 9.2 displays goodness-of-fit tests for several models. Because many cell counts are small, the chi-squared approximation for G^2 may be poor, but this index is useful for comparing models. The first model listed contains only the GR association and assumes conditional independence for the other nine pairs of associations. It fits horribly, which is no surprise. Model 2, with all two-factor terms, on the other hand, seems to fit well. Model 3, containing all

TABLE 9.1 Alcohol, Cigarette, and Marijuana Use for High School Seniors

		Marijuana Use							
		Race = White				Race = Other			
Alcohol Use	Cigarette Use	Female		Male		Female		Male	
		Yes	No	Yes	No	Yes	No	Yes	No
Yes	Yes	405	268	453	228	23	23	30	19
	No	13	218	28	201	2	19	1	18
No	Yes	1	17	1	17	0	1	1	8
	No	1	117	1	133	0	12	0	17

Source: Harry Khamis, Wright State University.

TABLE 9.2 Goodness-of-Fit Tests for Loglinear Models for Table 9.1

Model[a]	G^2	df
1. Mutual independence + GR	1325.1	25
2. Homogeneous association	15.3	16
3. All three-factor terms	5.3	6
4a. (2)–AC	201.2	17
4b. (2)–AM	107.0	17
4c. (2)–CM	513.5	17
4d. (2)–AG	18.7	17
4e. (2)–AR	20.3	17
4f. (2)–CG	16.3	17
4g. (2)–CR	15.8	17
4h. (2)–GM	25.2	17
4i. (2)–MR	18.9	17
5. ($AC, AM, CM, AG, AR, GM, GR, MR$)	16.7	18
6. ($AC, AM, CM, AG, AR, GM, GR$)	19.9	19
7. (AC, AM, CM, AG, AR, GR)	28.8	20

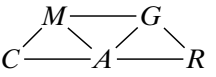
[a] G, gender; R, race; A, alcohol use; C, cigarette use; M, marijuana use.

the three-factor interaction terms, also fits well, but the improvement in fit is not great (difference in G^2 of $15.3 - 5.3 = 10.0$ based on df $= 16 - 6 = 10$). Thus, we consider models without three-factor terms. Beginning with model 2, we eliminate two-factor terms. We use backward elimination, sequentially taking out terms for which the resulting increase in G^2 is smallest, when refitting the model.

Table 9.2 shows the start of this process. Nine pairwise associations are candidates for removal from model 2 (all except GR), shown in models 4a through 4i. The smallest increase in G^2, compared to model 2, occurs in removing the CR term (i.e., model 4g). The increase is $15.8 - 15.3 = 0.5$, with df $= 17 - 16 = 1$, so this elimination seems sensible. After removing it,

the smallest additional increase results from removing the CG term (model 5), resulting in $G^2 = 16.7$ with df = 18, and a change in G^2 of 0.9 based on df = 1. Removing next the MR term (model 6) yields $G^2 = 19.9$ with df = 19, a change in G^2 of 3.2 based on df = 1.

Further removals have a more severe effect. For instance, removing the AG term increases G^2 by 5.3, with df = 1, for a P-value of 0.02. One cannot take such P-values literally, since the data suggested these tests, but it seems safest not to drop additional terms. [See Westfall and Wolfinger (1997) and Westfall and Young (1993) for methods of adjusting P-values to account for multiple tests]. Model 6, denoted by $(AC, AM, CM, AG, AR, GM, GR)$, has association graph



Every path between C and $\{G, R\}$ involves a variable in $\{A, M\}$. Given the outcome on alcohol use and marijuana use, the model states that cigarette use is independent of both gender and race. Collapsing over the explanatory variables race and gender, the conditional associations between C and A and between C and M are the same as with the model (AC, AM, CM) fitted in Section 8.2.4.

Removing the GM term from this model yields model 7 in Table 9.2. Its association graph reveals that A separates $\{G, R\}$ from $\{C, M\}$. Thus, all pairwise conditional associations among A, C, and M in model 7 are identical to those in model (AC, AM, CM), collapsing over G and R. In fact, model 7 does not fit poorly ($G^2 = 28.8$ with df = 20) considering the large sample size. (Its sample dissimilarity index is $\hat{\Delta} = 0.036$.) Hence, one might collapse over gender and race in studying associations among the primary variables. An advantage of the full five-variable model is that it estimates effects of gender and race on these responses, in particular the effects of race and gender on alcohol use and the effect of gender on marijuana use.

9.2.3 Loglinear Model Comparison Statistics

Consider two loglinear models, M_1 and M_0, with M_0 a special case of M_1. By Sections 4.5.4 and 5.4.3, the likelihood-ratio statistic for testing M_0 against M_1 is $G^2(M_0 \mid M_1) = G^2(M_0) - G^2(M_1)$. We used this statistic above in comparing pairs of models.

Let $\mathbf{n}$ denote a column vector of the observed cell counts $\{n_i\}$. Let $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\mu}}_1$ denote vectors of the fitted values $\{\hat{\mu}_{0i}\}$ and $\{\hat{\mu}_{1i}\}$ for M_0 and M_1. The deviance $G^2(M_0)$ for the simpler model partitions into

$$G^2(M_0) = G^2(M_1) + G^2(M_0 \mid M_1). \tag{9.1}$$

Just as $G^2(M)$ measures the distance of fitted values for M from $\mathbf{n}$, $G^2(M_0 \mid M_1)$ measures the distance of fit $\hat{\boldsymbol{\mu}}_0$ from fit $\hat{\boldsymbol{\mu}}_1$. In this sense,

decomposition (9.1) expresses a certain orthogonality: The distance of $\mathbf{n}$ from $\hat{\boldsymbol{\mu}}_0$ equals the distance of $\mathbf{n}$ from $\hat{\boldsymbol{\mu}}_1$ plus the distance of $\hat{\boldsymbol{\mu}}_1$ from $\hat{\boldsymbol{\mu}}_0$.

The model comparison statistic equals

$$G^2(M_0 \mid M_1) = 2\sum_i n_i \log(n_i/\hat{\mu}_{0i}) - 2\sum_i n_i \log(n_i/\hat{\mu}_{1i})$$
$$= 2\sum_i n_i \log(\hat{\mu}_{1i}/\hat{\mu}_{0i}). \tag{9.2}$$

The two loglinear models have the matrix form (8.17), or

$$\log \boldsymbol{\mu}_0 = \mathbf{X}_0\boldsymbol{\beta}_0 \quad \text{and} \quad \log \boldsymbol{\mu}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 .$$

Since M_0 is simpler than M_1, one can express $\log \boldsymbol{\mu}_0 = \mathbf{X}_0\boldsymbol{\beta}_0 = \mathbf{X}_1\boldsymbol{\beta}_1^*$, where $\boldsymbol{\beta}_1^*$ equals $\boldsymbol{\beta}_0$ with 0 elements appended corresponding to the extra parameters in $\boldsymbol{\beta}_1$ but not in $\boldsymbol{\beta}_0$. Then, from (9.2),

$$G^2(M_0 \mid M_1) = 2\mathbf{n}'(\log \hat{\boldsymbol{\mu}}_1 - \log \hat{\boldsymbol{\mu}}_0) = 2\mathbf{n}'\left[\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1^*\right]$$
$$= 2\hat{\boldsymbol{\mu}}_1'\left[\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1^*\right] = 2\hat{\boldsymbol{\mu}}_1'(\log \hat{\boldsymbol{\mu}}_1 - \log \hat{\boldsymbol{\mu}}_0)$$
$$= 2\sum \hat{\mu}_{1i}\log(\hat{\mu}_{1i}/\hat{\mu}_{0i}), \tag{9.3}$$

where the replacement of $\mathbf{n}$ by $\hat{\boldsymbol{\mu}}_1$ follows from the likelihood equations $\mathbf{n}'\mathbf{X}_1 = \hat{\boldsymbol{\mu}}_1'\mathbf{X}_1$ for M_1 [Recall (8.22)]. Statistic (9.3) has the same form as $G^2(M_0)$, but with $\{\hat{\mu}_{1i}\}$ playing the role of the observed data. Note that $G^2(M_0)$ is the special case of $G^2(M_0 \mid M_1)$ with M_1 saturated.

The Pearson difference $X^2(M_0) - X^2(M_1)$ does not have Pearson form. It is not even necessarily nonnegative. A more appropriate Pearson statistic for comparing models is

$$X^2(M_0 \mid M_1) = \sum (\hat{\mu}_{1i} - \hat{\mu}_{0i})^2/\hat{\mu}_{0i}. \tag{9.4}$$

This has the usual form with $\{\hat{\mu}_{1i}\}$ in place of $\{n_i\}$. Statistics (9.3) and (9.4) depend on the data only through the fitted values and thus only through sufficient statistics for M_1.

When M_0 holds, $G^2(M_0)$ and $G^2(M_1)$ have asymptotic chi-squared distributions, and $G^2(M_0 \mid M_1)$ is asymptotically chi-squared with df equal to the difference between df for M_0 and M_1. Haberman (1977a) showed that $G^2(M_0 \mid M_1)$ and $X^2(M_0 \mid M_1)$ have the same null large-sample behavior, even for fairly sparse tables. (Under certain conditions, their difference converges in probability to 0 as n increases.) When M_1 holds but M_0 does not, $G^2(M_1)$ still has its asymptotic chi-squared distribution, but the other two statistics tend to grow unboundedly as n increases.

9.2.4 Partitioning Chi-Squared with Model Comparisons

Equation (9.1) utilizes the property by which a chi-squared statistic with df > 1 partitions into components. We used such partitionings in tests for trend with ordinal predictors in linear logit or linear probability models (Section 5.3.5) and with ordinal responses in cumulative logit models (Section 7.2). More generally, this property applies with a set of nested models to test a sequence of hypotheses. The separate tests for comparing pairs of models are asymptotically independent.

For example, a chi-squared decomposition with $J - 1$ models justifies the partitioning of G^2 stated in Section 3.3.3 for $2 \times J$ tables. For $j = 2, \ldots, J$, let M_j denote the model that satisfies

$$\theta_i = (\mu_{1i}\,\mu_{2,i+1})/(\mu_{1,i+1}\,\mu_{2i}) = 1, \quad i = 1, \ldots, j-1.$$

For M_j, the $2 \times j$ table consisting of columns 1 through j satisfies independence. Model M_J is independence in the complete $2 \times J$ table. Model M_h is a special case of M_j whenever $h > j$. By (9.2),

$$\begin{aligned} G^2(M_J) &= G^2(M_J \mid M_{J-1}) + G^2(M_{J-1}) \\ &= G^2(M_J \mid M_{J-1}) + G^2(M_{J-1} \mid M_{J-2}) + G^2(M_{J-2}) \\ &= \cdots = G^2(M_J \mid M_{J-1}) + \cdots + G^2(M_3 \mid M_2) + G^2(M_2). \end{aligned}$$

From (9.3), $G^2(M_j \mid M_{j-1})$ has the G^2 form with the fitted values for model M_{j-1} playing the role of the observed data. Substitution of fitted values for the two models into (9.3) shows that $G^2(M_j \mid M_{j-1})$ is identical to G^2 for testing independence in a 2×2 table; the first column combines column 1 through $j - 1$ of the original table, and the second column is column j of the original table.

With several preplanned comparisons, simultaneous test procedures lessen the probability of attributing importance to sample effects that simply reflect chance variation. These procedures use adjusted significance levels. For a set of s tests for nested models, when each test has level $1 - (1 - \alpha)^{1/s}$, the overall asymptotic P(type I error) $\leq \alpha$ (Goodman 1969a). For instance, suppose that we test the fit of (WXZ, WY, XY, ZY), compare that model to (WX, WZ, XZ, WY, XY, ZY), and compare that model to (WX, WZ, XZ, WY, ZY). To ensure overall $\alpha = 0.05$ for the $s = 3$ tests, use level $1 - (0.95)^{1/3} = 0.017$ for each.

9.2.5 Identical Marginal and Conditional Tests of Independence

A test using $G^2(M_0 \mid M_1)$ simplifies dramatically when both models have direct estimates. In that case, the models have independence linkages neces-

sary to ensure collapsibility. A test of conditional independence has the same result as the test of independence applied to the marginal table. Sundberg (1975) proved the following: When two direct models M_0 and M_1 are identical except for a pairwise association term, $G^2(M_0 \mid M_1)$ is identical to G^2 for testing independence in the marginal table for that pair of variables. Bishop (1971) and Goodman (1970, 1971b) have related discussion.

For instance, $G^2[(X, Y, Z) \mid (XY, Z)]$ tests $\lambda^{XY} = 0$ in model (XY, Z). Thus, it tests XY conditional independence under the assumption that X and Y are jointly independent of Z. Using the two sets of fitted values, from (9.3), it equals

$$2\sum_i \sum_j \sum_k \frac{n_{ij+}n_{++k}}{n} \log \frac{n_{ij+}n_{++k}/n}{n_{i++}n_{+j+}n_{++k}/n^2}$$

$$= 2\sum_i \sum_j n_{ij+} \log \frac{n_{ij+}}{n_{i++}n_{+j+}/n},$$

which equals $G^2[(X, Y)]$ for testing independence in the marginal XY table. This is not surprising. The collapsibility conditions imply that for model (XY, Z), the marginal XY association is the same as the conditional XY association.

9.3 DIAGNOSTICS FOR CHECKING MODELS

The model comparison test using $G^2(M_0 \mid M_1)$ is useful for detecting whether an extra term improves a model fit. Cell residuals provide a cell-specific indication of model lack of fit.

9.3.1 Residuals for Loglinear Models

In Section 4.5.5 we noted that residuals for the independence model (Section 3.3.1) extend to any Poisson GLM. For cell i in a contingency table with observed count n_i and fitted value $\hat{\mu}_i$, the *Pearson residual* is

$$e_i = \frac{n_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}. \tag{9.5}$$

These relate to the Pearson statistic by $\sum e_i^2 = X^2$.

Like the Pearson residual (6.1) for binomial models, the asymptotic variances of $\{e_i\}$ are less than 1.0. They average (residual df)/(number of

cells). Haberman (1973a) defined the standardized Pearson residual,

$$r_i = e_i/\sqrt{1 - \hat{h}_i}\,,$$

where the leverage $\hat{h}_i$ is a diagonal element of the estimated hat matrix (Section 4.5.5). This has an asymptotic standard normal distribution and is preferable to the Pearson residual. A closed-form expression applies for loglinear models having direct estimates (Haberman 1978, p. 275). Alternative residuals use components of the deviance (Section 4.5.5).

9.3.2 Student Survey Example Revisited

For Table 9.1 cross-classifying alcohol, cigarette, and marijuana use by gender and race, we suggested in Section 9.2.2 that the model with all two-factor associations is plausible. For it, the only large standardized Pearson residual equals 3.2, resulting from a fitted value of 3.1 in the cell having a count of 8. Further comparisons suggested that the simpler model $(AC, AM, CM, AG, AR, GM, GR)$ is adequate. Its only large standardized residual equals 3.3, referring to a fitted value of 2.9 in that cell. The number of nonwhite males who did not use alcohol or marijuana but who smoked cigarettes is somewhat greater than either model predicts. The standardized Pearson residuals do not suggest problems with either model, considering the large sample size and many cells studied.

9.3.3 Correspondence between Loglinear and Logit Residuals

In Section 8.5 we showed that logit models in contingency tables are equivalent to certain loglinear models. However, a Pearson residual for a logit model differs from a Pearson residual for a loglinear model. The numerators comparing the ith observed and fitted binomial or Poisson count are the same, since the model fitted values are the same. However, the logit model uses a fitted binomial standard deviation in the denominator [see (6.1)], whereas the loglinear model uses a fitted Poisson standard deviation [see (9.5)]. Thus, the logit Pearson residual exceeds the loglinear Pearson residual (9.5).

Once standardized by dividing by estimated standard errors, the standardized Pearson residuals are identical for the two models. This is another reason for preferring standardized residuals over ordinary Pearson residuals.

9.4 MODELING ORDINAL ASSOCIATIONS

The loglinear models presented so far have a serious limitation—they treat all classifications as nominal. If the order of a variable's categories changes in

TABLE 9.3 Opinions about Premarital Sex and Availability of Teenage Birth Control

	Teenage Birth Control[a]			
Premarital Sex	Strongly Disagree	Disagree	Agree	Strongly Agree
Always wrong	81	68	60	38
	(42.4)[1]	(51.2)	(86.4)	(67.0)
	7.6[2]	3.1	−4.1	−4.8
	(80.9)[3]	(67.6)	(69.4)	(29.1)
Almost always wrong	24	26	29	14
	(16.0)	(19.3)	(32.5)	(25.2)
	2.3	1.8	−0.8	−2.8
	(20.8)	(23.1)	(31.5)	(17.6)
Wrong only sometimes	18	41	74	42
	(30.1)	(36.3)	(61.2)	(47.4)
	−2.7	1.0	2.2	−1.0
	(24.4)	(36.1)	(65.7)	(48.8)
Not wrong at all	36	57	161	157
	(70.6)	(85.2)	(143.8)	(111.4)
	−6.1	−4.6	2.4	6.8
	(33.0)	(65.1)	(157.4)	(155.5)

[a][1]Independence model fit; [2]standardized Pearson residuals for the independence model fit; [3]linear-by-linear association model fit.

Source: 1991 General Social Survey, National Opinion Research Center.

any way, the fit is the same. For ordinal classifications, these models ignore important information.

Refer to Table 9.3. Subjects were asked their opinion about a man and woman having sexual relations before marriage (always wrong, almost always wrong, wrong only sometimes, not wrong at all). They were also asked whether methods of birth control should be available to teenagers between the ages of 14 and 16 (strongly disagree, disagree, agree, strongly agree). For the loglinear model of independence, denoted by I, $G^2(I) = 127.6$ with df = 9. The model fits poorly. Yet, adding the ordinary association term makes it saturated and unhelpful.

Table 9.3 also contains fitted values and standardized residuals for independence. The residuals in the corners stand out. Sample counts are much larger than independence predicts where both responses are the most negative possible or the most positive possible. By contrast, the counts are much smaller than fitted values where one response is the most positive and the other is the most negative. Cross-classifications of ordinal variables often exhibit their greatest deviations from independence in the corner cells. This pattern for Table 9.3 indicates lack of fit in the form of a positive trend.

Subjects who are more willing to make birth control available to teenagers also tend to feel more tolerant about premarital sex.

Models for ordinal variables use association terms that permit trends. The models are more complex than the independence model, yet unsaturated. Models with association and interaction terms exist in situations in which nominal models are saturated. Tests with ordinal models have improved power for detecting trends.

9.4.1 Linear-by-Linear Association in Two-Way Tables

For two-way tables, a simple model for two ordinal variables assigns ordered row scores $u_1 \leq u_2 \leq \cdots \leq u_I$ and column scores $v_1 \leq v_2 \leq \cdots \leq v_J$. The model is

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j, \tag{9.6}$$

with constraints such as $\lambda_I^X = \lambda_J^Y = 0$. This is the special case of the saturated model (8.2) in which $\lambda_{ij}^{XY} = \beta u_i v_j$. It requires only one parameter to describe association, whereas the saturated model requires $(I-1)(J-1)$.

Independence occurs when $\beta = 0$. The term $\beta u_i v_j$ represents the deviation of $\log \mu_{ij}$ from independence. The deviation is linear in the Y scores at a fixed level of X and linear in the X scores at a fixed level of Y. In column j, for instance, the deviation is a linear function of X, having form (slope) $\times$ (score for X), with slope βv_j. Because of this property, (9.6) is called the *linear-by-linear association model* (abbreviated, $L \times L$). The model has its greatest departures from independence in the corners of the table. Birch (1965), Goodman (1979a), and Haberman (1974b) introduced special cases.

The direction and strength of the association depend on β. When $\beta > 0$, Y tends to increase as X increases. Expected frequencies are larger than expected (under independence) in cells where X and Y are both high or both low. When $\beta < 0$, Y tends to decrease as X increases. When the data display a positive or negative trend, the $L \times L$ model usually fits much better than the independence model.

For the 2×2 table using the cells intersecting rows a and c with columns b and d, direct substitution shows that the model has

$$\log \frac{\mu_{ab}\,\mu_{cd}}{\mu_{ad}\,\mu_{cb}} = \beta(u_c - u_a)(v_d - v_b). \tag{9.7}$$

This log odds ratio is stronger as $|\beta|$ increases and for pairs of categories that are farther apart. Simple interpretations result when $u_2 - u_1 = \cdots = u_I - u_{I-1}$ and $v_2 - v_1 = \cdots = v_J - v_{J-1}$. When $\{u_i = i\}$ and $\{v_j = j\}$, for instance, the *local odds ratios* (2.10) for adjacent rows and adjacent columns have common value e^{β}. Goodman (1979a) called this case *uniform association*. Figure 9.2 portrays local odds ratios having uniform value.

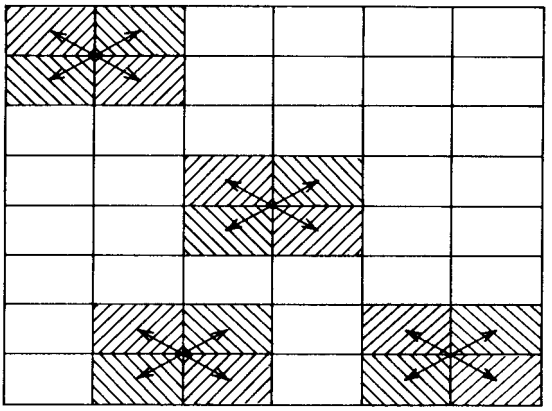


FIGURE 9.2 Constant odds ratio implied by uniform association model. (*Note:* β = the constant log odds ratio for adjacent rows and adjacent columns.)

The choice of scores affects the interpretation of β. Often, the response scale discretizes an inherently continuous scale. It is sensible to choose scores that approximate distances between midpoints of categories for the underlying scale, such as we did in measuring alcohol consumption for a linear logit model in Section 3.4.5. It is sometimes useful to standardize the scores, subtracting the mean and dividing by the standard deviation, so

$$\sum u_i \pi_{i+} = \sum v_j \pi_{+j} = 0$$
$$\sum u_i^2 \pi_{i+} = \sum v_j^2 \pi_{+j} = 1.$$

Then, β represents the log odds ratios for standard deviation distances in the X and Y directions. The $L \times L$ model tends to fit well when an underlying continuous distribution is approximately bivariate normal. For standardized scores, β is then comparable to $\rho/(1-\rho^2)$, where ρ is the underlying correlation. For weak associations, $\beta \approx \rho$ (see Becker 1989b; Goodman 1981a, b, 1985).

9.4.2 Corresponding Logit Model for Adjacent Responses

A logit formulation of the $L \times L$ model treats Y as a response and X as explanatory. Let $\pi_{j|i} = P(Y = j \,|\, X = i)$. Using logits for adjacent response categories (Section 7.4.1),

$$\log \frac{\pi_{j+1|i}}{\pi_{j|i}} = \log \frac{\mu_{i,j+1}}{\mu_{ij}} = \left(\lambda_{j+1}^Y - \lambda_j^Y\right) + \beta(v_{j+1} - v_j)u_i.$$

For unit-spaced $\{v_j\}$, this simplifies to

$$\log \frac{\pi_{j+1|i}}{\pi_{j|i}} = \alpha_j + \beta u_i$$

where $\alpha_j = \lambda_{j+1}^Y - \lambda_j^Y$. The same linear logit effect β applies simultaneously for all $(J-1)$ pairs of adjacent response categories: The odds $Y = j+1$ instead of $Y = j$ multiply by e^β for each unit change in X. In using equal-interval response scores, we implicitly assume that the effect of X is the same on each of the $J-1$ adjacent-categories logits for Y.

9.4.3 Likelihood Equations and Model Fitting

The Poisson log-likelihood $L(\boldsymbol{\mu}) = \Sigma_i \Sigma_j n_{ij} \log \mu_{ij} - \Sigma_i \Sigma_j \mu_{ij}$ simplifies for the $L \times L$ model (9.6) to

$$L(\boldsymbol{\mu}) = n\lambda + \sum_i n_{i+}\lambda_i^X + \sum_j n_{+j}\lambda_j^Y + \beta \sum_i \sum_j u_i v_j n_{ij} - \sum_i \sum_j \exp\left(\lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j\right).$$

Differentiating $L(\boldsymbol{\mu})$ with respect to $(\lambda_i^X, \lambda_j^Y, \beta)$ and setting the three partial derivatives equal to zero yields likelihood equations

$$\hat{\mu}_{i+} = n_{i+},\ i = 1, \ldots, I, \qquad \hat{\mu}_{+j} = n_{+j},\ j = 1, \ldots, J,$$

$$\sum_i \sum_j u_i v_j \hat{\mu}_{ij} = \sum_i \sum_j u_i v_j n_{ij}.$$

Iterative methods such as Newton–Raphson yield the ML fit.

Let $p_{ij} = n_{ij}/n$ and $\hat{\pi}_{ij} = \hat{\mu}_{ij}/n$. The third likelihood equation implies that

$$\sum_i \sum_j u_i v_j \hat{\pi}_{ij} = \sum_i \sum_j u_i v_j p_{ij}.$$

Since marginal distributions and hence marginal means and variances are identical for fitted and observed distributions, the third equation implies the correlation between the scores for X and Y is the same for both distributions. The fitted counts display the same positive or negative trend as the data.

Since $\{u_i\}$ and $\{v_j\}$ are fixed, the $L \times L$ model (9.6) has only one more parameter (β) than the independence model. Its residual

$$\text{df} = IJ - [1 + (I-1) + (J-1) + 1] = IJ - I - J,$$

unsaturated for all but 2×2 tables.

9.4.4 Sex Opinions Example

Table 9.3 also reports fitted values for the linear-by-linear association model applied to Table 9.3, using scores $\{1, 2, 3, 4\}$ for rows and columns. Table 9.4

TABLE 9.4 Output for Fitting Linear-by-Linear Association Model to Table 9.3

```
        Criteria For Assessing Goodness Of Fit
        Criterion              DF      Value
        Deviance                8     11.5337
        Pearson Chi-Square      8     11.5085

                          Standard   Wald 95% Conf.    Chi-
Parameter      Estimate    Error        Limits       Square  Pr > ChiSq
Intercept        0.4735    0.4339   -0.3769  1.3239    1.19    0.2751
premar     1     1.7537    0.2343    1.2944  2.2129   56.01    <.0001
premar     2     0.1077    0.1988   -0.2820  0.4974    0.29    0.5880
premar     3    -0.0163    0.1264   -0.2641  0.2314    0.02    0.8972
premar     4     0.0000    0.0000    0.0000  0.0000       .       .
birth      1     1.8797    0.2491    1.3914  2.3679   56.94    <.0001
birth      2     1.4156    0.1996    1.0243  1.8068   50.29    <.0001
birth      3     1.1551    0.1291    0.9021  1.4082   80.07    <.0001
birth      4     0.0000    0.0000    0.0000  0.0000       .       .
linlin           0.2858    0.0282    0.2305  0.3412  102.46    <.0001

                    LR Statistics
          Source      DF     Chi-Square     Pr > ChiSq
          linlin       1         116.12         >.0001
```

shows software output. To get this, we added a variable (denoted "linlin") to the independence model having values equal to the product of row and column number. Compared to the independence model, for which $G^2(I) = 127.6$ with df = 9, the $L \times L$ model fits dramatically better [$G^2(L \times L) = 11.5$, df = 8]. This is especially noticeable in the corners, where it predicts the greatest departures from independence.

The ML estimate $\hat{\beta} = 0.286$ (SE = 0.028) indicates that subjects having more favorable attitudes about teen birth control also tend to have more tolerant attitudes about premarital sex. The estimated local odds ratio is $\exp(\hat{\beta}) = \exp(0.286) = 1.33$. A 95% Wald confidence interval is $\exp(0.286 \pm 1.96 \times 0.028)$, or (1.26, 1.41). The strength of association seems weak. From (9.7), however, nonlocal odds ratios are stronger. The estimated odds ratio for the four corner cells equals

$$\exp\big[\,\hat{\beta}(u_4 - u_1)(v_4 - v_1)\big] = \exp[0.286(4-1)(4-1)] = 13.1.$$

This also results from the corner fitted values, $(80.9 \times 155.5)/(29.1 \times 33.0) = 13.1$.

Two sets of scores having the same spacings yield the same $\hat{\beta}$ and the same fit. Any other sets of equally spaced scores yield the same fit but an appropriately rescaled $\hat{\beta}$. For instance, using row scores $\{2, 4, 6, 8\}$ with $\{v_j = j\}$ also yields $G^2 = 11.5$, but $\hat{\beta} = 0.143$ with SE = 0.014 (both half as

large). For Table 9.3, one might regard categories 2 and 3 as farther apart than categories 1 and 2, or categories 3 and 4. Scores such as $\{1, 2, 4, 5\}$ for rows and columns recognize this. The $L \times L$ model then has $G^2 = 8.8$ (df = 8) and $\hat{\beta} = 0.146$ (SE = 0.014).

One need not regard the scores as approximations for distances between categories or as reasonable scalings of ordinal variables in order for the models to be valid. They simply imply a certain pattern for the odds ratios. If the $L \times L$ model fits well with equally spaced row and column scores, the uniform local odds ratio describes the association regardless of whether the scores are sensible indexes of true distances between categories.

For scores $\{u_i = i\}$ with Table 9.3, the marginal mean and standard deviation for premarital sex are 2.81 and 1.26. The standardized scores are $\{(i - 2.81)/1.26\}$, or $(-1.44, -0.65, 0.15, 0.95)$. The standardized equal-interval scores for birth control are $(-1.65, -0.69, 0.27, 1.23)$. For these scores, $\hat{\beta} = 0.374$. By solving $\hat{\beta} = \hat{\rho}/(1 - \hat{\rho}^2)$ for $\hat{\rho}$, $\hat{\rho} = 0.333$. If there is an underlying bivariate normal distribution, we estimate the correlation to be 0.333.

9.4.5 Directed Ordinal Test of Independence

For the linear-by-linear association model, H_0: independence is H_0: $\beta = 0$. The likelihood-ratio test statistic equals

$$G^2(I|L \times L) = G^2(I) - G^2(L \times L) .$$

Designed to detect positive or negative trends, it has df = 1. For Table 9.3, $G^2(I|L \times L) = 127.6 - 11.5 = 116.1$. This has $P < 0.0001$, extremely strong evidence of an association. The Wald statistic $z^2 = (\hat{\beta}/\text{SE})^2 = (0.286/0.0282)^2 = 102.5$ (df = 1) also shows strong evidence. The correlation statistic (3.15) presented in Section 3.4.1 for testing independence is the score statistic for H_0: $\beta = 0$ in this model. It equals 112.6 (df = 1).

When the $L \times L$ model holds, the ordinal test using $G^2(I|L \times L)$ is asymptotically more powerful than the test using $G^2(I)$. This is true for the same reason given in Section 6.4.2 for the linear logit model. The power of a chi-squared test increases when df decrease, for fixed noncentrality. When the $L \times L$ model holds, the noncentrality is the same for $G^2(I|L \times L)$ and $G^2(I)$; thus $G^2(I|L \times L)$ is more powerful, since its df = 1 compared to $(I - 1)(J - 1)$ for $G^2(I)$. The power advantage increases as I and J increase, since the noncentrality remains focused on df = 1 for $G^2(I|L \times L)$ but df also increases for $G^2(I)$.

9.5 ASSOCIATION MODELS*

Generalizations of the linear-by-linear association model apply to multiway tables or treat scores as parameters rather than fixed. The models are called *association models*, because they focus on the association structure.

9.5.1 Row and Column Effects Models

We first present a model that treats X as nominal and Y as ordinal. It is appropriate for two-way tables with ordered columns, using scores $v_1 \leq v_2 \leq \cdots \leq v_J$. Since the rows are unordered, they do not have scores. Replacing the ordered values $\{\beta u_i\}$ in the linear-by-linear term $\beta u_i v_j$ in model (9.6) by unordered parameters $\{\mu_i\}$ gives

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \mu_i v_j. \tag{9.8}$$

Constraints are needed such as $\lambda_I^X = \lambda_J^Y = \mu_I = 0$. The $\{\mu_i\}$ are called *row effects*. The model is called the *row effects model*.

Model (9.8) has $I - 1$ more parameters (the $\{\mu_i\}$) than the independence model. Independence is the special case $\mu_1 = \cdots = \mu_I$. A corresponding *column effects model* has association term $u_i \nu_j$. It treats X as ordinal with scores $\{u_i\}$ and Y as nominal with parameters $\{\nu_j\}$. The row effects and column effects models were developed by Goodman (1979a), Haberman (1974b), and Simon (1974).

9.5.2 Logit Model for Adjacent Responses

With $\{v_{j+1} - v_j = 1\}$, the row effects model has adjacent-categories logit form

$$\log \frac{P(Y = j + 1 \,|\, X = i)}{P(Y = j \,|\, X = i)} = \alpha_j + \mu_i. \tag{9.9}$$

The effect in row i is identical for each pair of adjacent responses. Plots of these logits against i $(i = 1, \ldots, I)$ for different j are parallel. Goodman (1983) referred to model (9.9) as the *parallel odds* model.

Differences among $\{\mu_i\}$ compare rows with respect to their conditional distributions on Y. When $\mu_i = \mu_h$, rows h and i have identical conditional distributions. If $\mu_i > \mu_h$, Y is stochastically higher in row i than row h.

The likelihood equations for the row effects model (9.8) are $\{\hat{\mu}_{i+} = n_{i+}\}$, $\{\hat{\mu}_{+j} = n_{+j}\}$, and

$$\sum_j v_j \hat{\mu}_{ij} = \sum v_j n_{ij}, \quad i = 1, \ldots, I.$$

Let $\hat{\pi}_{j|i} = \hat{\mu}_{ij}/\hat{\mu}_{i+}$ and $p_{j|i} = n_{ij}/n_{i+}$. Since $\hat{\mu}_{i+} = n_{i+}$, the third likelihood equation is $\sum_j v_j \hat{\pi}_{j\,|\,i} = \sum_j v_j p_{j\,|\,i}$. For the conditional distribution within each row, the mean column score is the same for the fitted and sample distributions. The likelihood equations are solved using iterative methods.

TABLE 9.5 Observed Frequencies and Fitted Values for Political Ideology Data

	Political Ideology[a]			
Party Affiliation	Liberal	Moderate	Conservative	Total
Democrat	143	156	100	399
	(102.0)[1]	(161.4)	(135.6)	
	(136.6)[2]	(168.7)	(93.6)	
Independent	119	210	141	470
	(120.2)	(190.1)	(159.7)	
	(123.8)	(200.4)	(145.8)	
Republican	15	72	127	214
	(54.7)	(86.6)	(72.7)	
	(16.6)	(68.9)	(128.6)	

[a][1] Independence model; [2] row effects model.

Source: Based on data in R. D. Hedlund, *Public Opinion Quart.* **41**: 498–514 (1978).

9.5.3 Political Ideology Example

Table 9.5 displays the relationship between political ideology and political party affiliation for a sample of voters in a presidential primary in Wisconsin. The table shows fitted values for the independence (I) model and the row effects (R) model with $\{v_j = j\}$.

Table 9.6 shows output. Goodness-of-fit tests show that independence is inadequate. Adding the row effects parameters much improves the fit ($G^2(I) = 105.7$, df = 4; $G^2(R) = 2.8$, df = 2). Also, testing H_0: $\mu_1 = \mu_2 = \mu_3$ using $G^2(I \mid R) = 102.9$ (df = 2) shows very strong evidence of an association. In Table 9.5, the improved fit is especially noticeable at the ends of the ordinal scale, where the model has greatest deviation from independence.

The output uses dummy variables for the first two categories of each classification. The interaction term equals the product of the score for ideology and a parameter for party. Thus, the row effect estimates satisfy $\hat{\mu}_3 = 0$, and the other two estimates contrast the first two parties with Republicans. The estimates are $\hat{\mu}_1 = -1.213$ and $\hat{\mu}_2 = -0.943$. The further $\hat{\mu}_i$ falls in the negative direction, the greater the tendency for the party i to locate at the liberal end of the ideology scale, relative to Republicans. In this sample the Republicans are much more conservative than the other two groups, and the Democrats (row 1) are the most liberal. From (9.9) the model predicts constant odds ratios for adjacent columns of political ideology. For instance, since $\hat{\mu}_3 - \hat{\mu}_1 = 1.213$, the estimated odds that Republicans were conservative instead of moderate, or moderate instead of liberal, were $\exp(1.213) = 3.36$ times the corresponding estimated odds for Democrats. Figure 9.3 shows the parallelism of the estimated logits for the row effects model.

The loglinear model does not distinguish between response and explanatory variables. Instead, one could use a cumulative logit model to describe

TABLE 9.6 Output for Fitting Row Effects Model to Table 9.5

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value
Deviance	2	2.8149
Pearson Chi-Square	2	2.8039

Parameter		Estimate	Std Error	Wald 95% Conf. Limits		Chi-Square	Pr > ChiSq
Intercept		4.8565	0.0858	4.6883	5.0246	3204.02	<.0001
party	Democ	3.3230	0.3188	2.6981	3.9479	108.63	<.0001
party	Indep	2.9536	0.3149	2.3364	3.5707	87.98	<.0001
party	Repub	0.0000	0.0000	0.0000	0.0000	.	.
ideology	1	−2.0488	0.2216	−2.4831	−1.6145	85.50	<.0001
ideology	2	−0.6244	0.1139	−0.8476	−0.4013	30.08	<.0001
ideology	3	0.0000	0.0000	0.0000	0.0000	.	.
score*party	Democ	−1.2134	0.1304	−1.4690	−0.9577	86.56	<.0001
score*party	Indep	−0.9426	0.1260	−1.1896	−0.6956	55.95	<.0001
score*party	Repub	0.0000	0.0000	0.0000	0.0000	.	.

LR Statistics

Source	DF	Chi-Square	Pr > ChiSq
score*party	2	102.85	<.0001

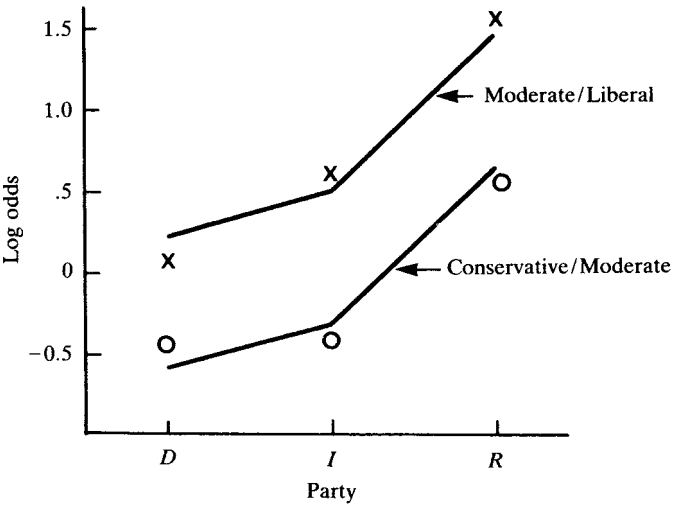


FIGURE 9.3 Observed and predicted logits for adjacent response categories.

the effects of party affiliation on ideology, or a baseline-category logit model to describe linear effects of ideology on party affiliation.

9.5.4 Ordinal Variables in Models for Multiway Tables

Multidimensional tables with ordinal responses can use generalizations of association models. In three dimensions, the rich collection of models includes (1) association models that are more parsimonious than the nominal model (XY, XZ, YZ), and (2) models permitting heterogeneous association that, unlike model (XYZ), are unsaturated.

Models for association that are special cases of (XY, XZ, YZ) replace λ association terms by structured terms that account for ordinality. For instance, when both X and Y are ordinal, alternatives to λ_{ij}^{XY} are a linear-by-linear term $\beta u_i v_j$, a row effects term $\mu_i v_j$, or a column effects term $u_i \nu_j$; these provide a stochastic ordering of conditional distributions within rows and within columns, or just within rows, or just within columns. With a linear-by-linear term, the model is

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta u_i v_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \qquad (9.10)$$

The conditional local odds ratios (8.13) then satisfy

$$\log \theta_{ij(k)} = \beta(u_{i+1} - u_i)(v_{j+1} - v_j) \quad \text{for all } k.$$

The association is the same in different partial tables, with *homogeneous linear-by-linear XY association*.

When the association is heterogeneous, structured terms for ordinal variables make effects simpler to interpret than in the saturated model. For instance, the *heterogeneous linear-by-linear XY association model*

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_k u_i v_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \qquad (9.11)$$

allows the XY association to change across levels of Z. With unit-spaced scores,

$$\log \theta_{ij(k)} = \beta_k \qquad \text{for all } i \text{ and } j.$$

It has uniform association within each level of Z, but heterogeneity among levels of Z in the strength of association. Fitting it corresponds to fitting the $L \times L$ model (9.6) separately at each level of Z.

9.5.5 Air Pollution and Breathing Examples

Table 9.7 displays associations among smoking status (S), breathing test results (B), and age (A) for workers in certain industrial plants in Houston,

TABLE 9.7 Cross-Classification of Industrial Workers by Breathing Test Results

Age	Smoking Status	Breathing Test Results		
		Normal	Borderline	Abnormal
< 40	Never smoked	577	27	7
	Former smoker	192	20	3
	Current smoker	682	46	11
40–59	Never smoked	164	4	0
	Former smoker	145	15	7
	Current smoker	245	47	27

Texas. The loglinear model (SA, SB, BA) fits poorly ($G^2 = 25.9$, df = 4). Thus, simpler models such as homogeneous linear-by-linear SB association are not plausible ($G^2 = 29.1$, df = 7, using equally spaced scores). The heterogeneous linear-by-linear SB association model fits much better with only one additional parameter ($G^2 = 10.8$, df = 6). With integer scores for S and B, $\hat{\beta}_1 = 0.115$ for the younger group and $\hat{\beta}_2 = 0.781$ for the older group, with SE = 0.167 for the difference. The effect of smoking seems much stronger for the older group, with estimated local odds ratio of exp(0.781) = 2.18 compared to exp(0.115) = 1.12 for the younger group. Here, it may be more natural to use logit models with B as the response variable (Problem 7.11).

When strata are ordered, roughly a linear trend may exist across strata in certain log odds ratios as Table 9.8 illustrates. The data refer to a sample of coal miners, measured on B = breathlessness, W = wheeze, and A = age, where B and W are response variables. One could use a separate logit model to describe effects of age on each response. To study whether the BW association varies by age, we fit model (BW, AB, AW). It has residual $G^2 = 26.7$, with df = 8. Table 9.8 reports the standardized Pearson residuals. They show a decreasing tendency as age increases.

This suggests the model

$$\log \mu_{ijk} = (BW, AB, AW) + kI(i = j = 1)\,\delta, \tag{9.12}$$

where I is the indicator function. It amends the homogeneous association model by adding δ in the cell for $\mu_{111}, \ldots, 9\delta$ in the cell for μ_{119}. Then, the BW log odds ratio changes linearly in the age category. The model fit has $\hat{\delta} = -0.131$ (SE = 0.029). The estimated BW log odds ratio at level k of age is $3.676 - 0.131k$, decreasing from 3.55 to 2.50. The model has residual $G^2 = 6.80$ (df = 7). McCullagh and Nelder (1989, Sec. 6.6) showed other analyses.

TABLE 9.8 Coal Miners Classified by Breathlessness, Wheeze, and Age

	Breathlessness				
	Yes		No		
Age	Wheeze Yes	Wheeze No	Wheeze Yes	Wheeze No	Std. Pearson Residual[a]
20–24	9	7	95	1841	0.75
25–29	23	9	105	1654	2.20
30–34	54	19	177	1863	2.10
35–39	121	48	257	2357	1.77
40–44	169	54	273	1778	1.13
45–49	269	88	324	1712	−0.42
50–54	404	117	245	1324	0.81
55–59	406	152	225	967	−3.65
60–64	372	106	132	526	−1.44

[a]Residual refers to yes–yes and no–no cells; reverse sign for yes–no and no–yes cells.

Source: Reprinted with permission from Ashford and Sowden (1970).

9.5.6 Other Ordinal Tests of Conditional Independence

Tests of conditional independence of ordinal classifications can generalize $G^2(I \mid L \times L)$. For instance, one can compare the XY conditional independence model (XZ, YZ) to the homogeneous linear-by-linear XY association model (9.10). It tests $\beta = 0$ in that model, with df = 1. This is an alternative to the ordinal test of conditional independence in Section 7.5.3. Like Mantel's score statistic (7.21), this statistic uses correlation information, since $\sum_k(\sum_i \sum_j u_i v_j n_{ijk})$ is the sufficient statistic for β in model (9.10). In fact, the Mantel statistic provides the score test of H_0: $\beta = 0$ in that model.

Exact, small-sample tests can use likelihood-ratio, score, or Wald statistics for such models. Computations require special algorithms (Agresti et al. 1990; Kim and Agresti 1997).

9.6 ASSOCIATION MODELS, CORRELATION MODELS, AND CORRESPONDENCE ANALYSIS*

The linear-by-linear association $(L \times L)$ model is a special case of the row effects (R) model, which has parameter row scores, and the column effects (C) model, which has parameter column scores. These models are special cases of a more general model with row *and* column parameter scores.

9.6.1 Multiplicative Row and Column Effects Model

Replacing $\{u_i\}$ and $\{v_j\}$ in the $L \times L$ model (9.6) by parameters yields the *row and column effects* (RC) model (Goodman 1979a)

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta \mu_i \nu_j. \tag{9.13}$$

Identifiability requires location and scale constraints on $\{\mu_i\}$ and $\{\nu_j\}$. The residual df $= (I-2)(J-2)$. This model is *not* loglinear, because the predictor is a multiplicative (rather than linear) function of parameters μ_i and ν_j. It treats classifications as nominal; the same fit results from a permutation of rows or columns. Parameter interpretation is simplest when at least one variable is ordinal, through the local log odds ratios

$$\log \theta_{ij} = \beta(\mu_{i+1} - \mu_i)(\nu_{j+1} - \nu_j).$$

Although it may seem appealing to use parameters instead of arbitrary scores, the *RC* model presents complications that do not occur with loglinear models. The likelihood may not be concave and may have local maxima. Independence is a special case, but it is awkward to test independence using the *RC* model. Haberman (1981) showed that the null distribution of $G^2(I) - G^2(RC)$ is not chi-squared but rather that of the maximum eigenvalue from a Wishart matrix.

When one set of parameter scores is fixed, the *RC* model simplifies to the *R* or *C* model. Goodman (1979a) suggested an iterative model-fitting algorithm that exploits this. A cycle of the algorithm has two steps. First, for some initial guess of $\{\nu_j\}$, it estimates the row scores as in the *R* model. Then, treating the estimated row scores from the first step as fixed, it estimates the column scores as in the *C* model. Those estimates serve as fixed column scores in the first step of the next cycle, for reestimating the row scores in the *R* model. There is no guarantee of convergence to ML estimates, but this seems to happen when the model fits well. Haberman (1995) provided more sophisticated fitting methods for association models.

Goodman (1985) expressed the association term in the saturated model in a form that generalizes the $\beta\mu_i\nu_j$ term in the *RC* model, namely,

$$\lambda_{ij}^{XY} = \sum_{k=1}^{M} \beta_k \mu_{ik} \nu_{jk} \tag{9.14}$$

where $M = \min(I-1, J-1)$. The parameters satisfy constraints such as

$$\begin{aligned} \sum_i \mu_{ik}\pi_{i+} = \sum_j \nu_{jk}\pi_{+j} &= 0 \qquad \text{for all } k, \\ \sum_i \mu_{ik}^2\pi_{i+} = \sum_j \nu_{jk}^2\pi_{+j} &= 1 \qquad \text{for all } k, \\ \sum_i \mu_{ik}\mu_{ih}\pi_{i+} = \sum_j \nu_{jk}\nu_{jh}\pi_{+j} &= 0 \qquad \text{for all } k \neq h. \end{aligned} \tag{9.15}$$

When $\beta_k = 0$ for $k > M^*$, model (9.14) is called the $RC(M^*)$ model. See Becker (1990) for ML model fitting. The *RC* model (9.13) is the case $M^* = 1$.

TABLE 9.9 Cross-Classification of Mental Health Status and Socioeconomic Status

Parents' Socioeconomic Status	Mental Health Status			
	Well	Mild Symptom Formation	Moderate Symptom Formation	Impaired
A (high)	64	94	58	46
B	57	94	54	40
C	57	105	65	60
D	72	141	77	94
E	36	97	54	78
F (low)	21	71	54	71

Source: Reprinted with permission from L. Srole et al. *Mental Health in the Metropolis: The Midtown Manhattan Study*, (New York: NYU Press, 1978), p. 289.

9.6.2 Mental Health Status Example

Table 9.9 describes the relationship between child's mental impairment and parents' socioeconomic status for a sample of residents of Manhattan (Goodman 1979a). The *RC* model fits well ($G^2 = 3.6$, df = 8). For scaling (9.15), the ML estimates are $(-1.11, -1.12, -0.37, 0.03, 1.01, 1.82)$ for the row scores, $(-1.68, -0.14, 0.14, 1.41)$ for the column scores, and $\hat{\beta} = 0.17$. Nearly all estimated local log odds ratios are positive, indicating a tendency for mental health to be better at higher levels of parents' SES.

Ordinal loglinear models also fit well. For equal-interval scores, $G^2(L \times L) = 9.9$ (df = 14). The statistic $G^2(L \times L \,|\, RC) = 6.3$ (df = 6) tests that row and column scores in the *RC* model are equal-interval. The parameter scores do not provide a significantly better fit. It is sufficient to use a uniform local odds ratio to describe the table. For unit-spaced scores, $\hat{\beta} = 0.091$ (SE = 0.015), so the fitted local odds ratio is $\exp(0.091) = 1.09$. There is strong evidence of positive association, but the degree of association is rather weak, at least locally.

9.6.3 Correlation Models

A *correlation model* for two-way tables has many features in common with the *RC* model (Goodman 1985). In its simplest form, it is

$$\pi_{ij} = \pi_{i+}\pi_{+j}(1 + \lambda\mu_i\nu_j), \tag{9.16}$$

where $\{\mu_i\}$ and $\{\nu_i\}$ are score parameters satisfying

$$\sum \mu_i\pi_{i+} = \sum \nu_j\pi_{+j} = 0 \quad \text{and} \quad \sum \mu_i^2\pi_{i+} = \sum \nu_j^2\pi_{+j} = 1.$$

The parameter λ is the correlation between the scores for joint distribution (9.16).

The correlation model is also called the *canonical correlation model*, because ML estimates of the scores maximize the correlation for (9.16). The general canonical correlation model is

$$\pi_{ij} = \pi_{i+}\pi_{+j}\left(1 + \sum_{k=1}^{M} \lambda_k \mu_{ik} \nu_{jk}\right)$$

where $0 \leq \lambda_M \leq \cdots \leq \lambda_1 \leq 1$ and with constraints such as in (9.15). The parameter λ_k is the correlation between $\{\mu_{ik},\ i = 1, \ldots, I\}$ and $\{\nu_{jk},\ j = 1, \ldots, J\}$. The $\{\mu_{i1}\}$ and $\{\nu_{j1}\}$ are standardized scores that maximize the correlation λ_1 for the joint distribution; $\{\mu_{i2}\}$ and $\{\nu_{j2}\}$ are standardized scores that maximize the correlation λ_2, subject to $\{\mu_{i1}\}$ and $\{\mu_{i2}\}$ being uncorrelated and $\{\nu_{j1}\}$ and $\{\nu_{j2}\}$ being uncorrelated, and so on.

Unsaturated models result from replacing M by $M^* < \min(I - 1, J - 1)$. Gilula and Haberman (1986) and Goodman (1985) discussed ML fitting. When λ is close to zero in (9.16), Goodman (1981a, 1985, 1986) noted that ML estimates of λ and the score parameters are similar to those of β and the score parameters in the *RC* model. Correlation models can also use fixed scores instead of parameter scores.

Goodman discussed advantages of association models over correlation models. The correlation model is not defined for all possible combinations of score values because of the constraint $0 \leq \pi_{ij} \leq 1$, ML fitted values do not have the same marginal totals as the observed data, and the model is not simply generalizable to multiway tables. Gilula and Haberman (1988) analyzed multiway tables with correlation models by treating explanatory variables as a single variable and response variables as a second variable.

9.6.4 Correspondence Analysis

Correspondence analysis is a graphical way to represent associations in two-way contingency tables. The rows and columns are represented by points on a graph, the positions of which indicate associations. Goodman (1985, 1986) noted that coordinates of the points are reparameterizations of $\{\mu_{ik}\}$ and $\{\nu_{jk}\}$ in the general canonical correlation model. Correspondence analysis uses adjusted scores

$$x_{ik} = \lambda_k \mu_{ik}, \qquad y_{jk} = \lambda_k \nu_{jk}.$$

These are close to zero for dimensions k in which the correlation λ_k is close to zero. A correspondence analysis graph uses the first two dimensions, plotting (x_{i1}, x_{i2}) for each row and (y_{j1}, y_{j2}) for each column.

TABLE 9.10 Scores from Correspondence Analysis Applied to Table 9.9

	Dimension				Dimension		
Column Score	1	2	3	Row Score	1	2	3
1	0.260	0.012	0.023	1	0.181	−0.018	0.028
2	0.030	0.024	−0.019	2	0.185	−0.011	−0.026
3	−0.013	−0.069	−0.002	3	0.059	−0.021	−0.010
4	−0.236	0.019	0.016	4	−0.008	0.042	0.011
				5	−0.164	0.044	−0.009
				6	−0.287	−0.061	0.005

Source: Reprinted with permission from the Institute of Mathematical Statistics, based on Goodman (1985).

Goodman (1985, 1986) used Table 9.9 to illustrate the similarities of correspondence analysis to analyses using correlation models and association models. For the general canonical correlation model, $M = \min(I - 1, J - 1) = 3$. Its estimated squared correlations are (0.0260, 0.0014, and 0.0003). The association is rather weak. Table 9.10 contains estimated row and column scores for the correspondence analysis of these three dimensions. Both sets of scores in the first dimension fall in a monotone increasing pattern, except for a slight discrepancy between the first two row scores. This indicates an overall positive association. The scores for the second and third dimension are close to zero, reflecting the relatively small $\hat{\lambda}_2$ and $\hat{\lambda}_3$.

Figure 9.4 exhibits the results of the correspondence analysis. The horizontal axis has estimates for the first dimension, and the vertical axis has estimates for the second dimension. Six points (circles) represent the six rows, with point i giving $(\hat{x}_{i1}, \hat{x}_{i2})$. Similarly, four points (squares) display the estimates $(\hat{y}_{j1}, \hat{y}_{j2})$. Both sets of points lie close to the horizontal axis, since the first dimension is more important than the second.

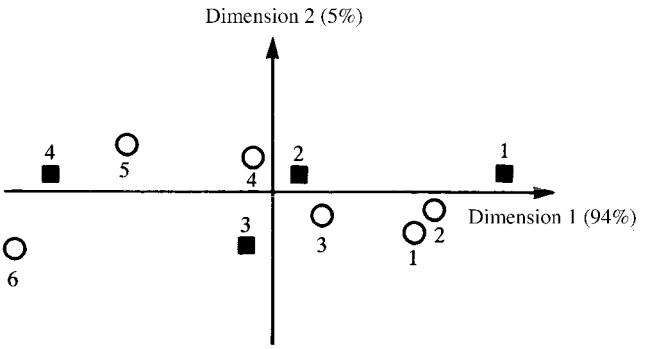


FIGURE 9.4 Graphical display of scores from first two dimensions of correspondence analysis. [Based on Escoufier (1982); reprinted with permission.]

Row points that are close together represent rows with similar conditional distributions across the columns. Close column points represent columns with similar conditional distributions across rows. Row points close to column points represent combinations that are more likely than expected under independence. Figure 9.4 shows a tendency for subjects at the high end of one scale to be at the high end of the other and for subjects at the low end of one to be at the low end of the other.

Correspondence analysis is used mainly as a descriptive tool. Goodman (1986) developed inferential methods for it. For Table 9.9, inferential analysis reveals that the first dimension, accounting for 94% of the total squared correlation, is adequate for describing the association. Goodman argued for choosing the unsaturated model employing only one dimension and having graphics display fitted scores for that dimension alone. Then, correspondence analysis is equivalent to a ML analysis using correlation model (9.16). The estimated scores for that model are $(-1.09, -1.17, -0.37, 0.05, 1.01, 1.80)$ for the rows and $(-1.60, -0.19, 0.09, 1.48)$ for the columns. The model fits well ($G^2 = 2.75, \text{df} = 8$). The quality of fit and the estimated scores are similar to those we saw in Section 9.6.2 for the *RC* model. More parsimonious correlation models also fit these data well, such as ones using equally spaced scores.

All analyses of Table 9.9 have yielded similar conclusions about the association. They all neglect, however, that mental health is a natural response variable. It may make more sense to use an ordinal logit model.

Like correlation models, a severe limitation of correspondence analysis is nontrivial generalization to multiway tables. Greenacre (1993) showed displays of several pairwise associations in a single plot.

9.6.5 Model Selection and Score Choice for Ordinal Variables

The past three sections showed several ways to use category orderings in model building. With allowance for ordinal effects, the variety of potential models is much greater than standard loglinear models. To choose among models, one approach uses the standard models for guidance. If a standard model fits well, simplify by replacing some parameters with structured terms for ordinal classifications.

Association, correlation, and correspondence analysis models have scores for categories of ordinal variables. Parameter interpretations are simplest for equally spaced scores. With parameter scores, the resulting ML estimates of scores need not be monotone. Constrained versions of the models force monotonicity by maximizing the likelihood subject to order restrictions (e.g., Agresti et al. 1987; Ritov and Gilula 1991). Disadvantages exist, however, of treating scores as parameters. The model becomes less parsimonious, and tests of effects may be less powerful because of a greater df value (recall Section 6.4.3). When one variable alone is a response, cumulative link models

(Sections 7.2 and 7.3) for that response do not require preassigned or parameter scores.

9.7 POISSON REGRESSION FOR RATES

Loglinear models need not refer to contingency tables. In Section 4.3 we introduced Poisson regression for modeling counts. When outcomes occur over time, space, or some other index of size, it is more relevant to model their *rate* of occurrence than their raw number.

9.7.1 Analyzing Rates Using Loglinear Models with Offsets

When a response count n_i has index equal to t_i, the sample rate is n_i/t_i. Its expected value is μ_i/t_i. With an explanatory variable x, a loglinear model for the expected rate has form

$$\log(\mu_i/t_i) = \alpha + \beta x_i. \tag{9.17}$$

This model has equivalent representation

$$\log \mu_i - \log t_i = \alpha + \beta x_i.$$

As noted in Section 8.7.4, the adjustment term, $-\log t_i$, to the log link of the mean is called an *offset*. The fit correspond to using $\log t_i$ as a predictor on the right-hand side and forcing its coefficient to equal 1.0.

For model (9.17), the expected response count satisfies

$$\mu_i = t_i \exp(\alpha + \beta x_i).$$

The mean is proportional to the index, with proportionality constant depending on the value of x. The identity link is also sometimes useful. The model is then

$$\mu_i/t_i = \alpha + \beta x_i, \quad \text{or} \quad \mu_i = \alpha t_i + \beta x_i t_i.$$

This does not require an offset. It corresponds to an ordinary Poisson GLM using identity link with t_i and $x_i t_i$ as explanatory variables and no intercept. It provides additive, rather than multiplicative, predictor effects. It is less useful with many predictors, as the fitting process may fail because of negative fitted counts at some iteration.

9.7.2 Modeling Death Rates for Heart Valve Operations

Laird and Olivier (1981) analyzed patient survival after heart valve replacement operations. A sample of 109 patients were classified by type of heart

TABLE 9.11 Data on Heart Valve Replacement Operations

Age		Type of Heart Valve	
		Aortic	Mitral
< 55	Deaths	4	1
	Time at risk	1259	2082
	Death rate	0.0032	0.0005
$55 +$	Deaths	7	9
	Time at risk	1417	1647
	Death rate	0.0049	0.0055

Source: Reprinted with permission, based on data in Laird and Olivier (1981).

valve (aortic, mitral) and by age (< 55, ≥ 55). Follow-up observations occurred until the patient died or the study ended. Operations occurred throughout the study period, and follow-up observations covered lengths of time varying from 3 to 97 months. The response was whether the subject died and the follow-up time. For subjects who died, this is the time after the operation until death; for the others, it is the time until the study ended or the subject withdrew from it.

Table 9.11 lists the numbers of deaths during the follow-up period, by valve type and age. These counts are the first layer of a three-way contingency table that classifies valve type, age, and whether died (yes, no). The subjects not tabulated in Table 9.11 were not observed to die. They are *censored*, since we know only a lower bound for how long they lived after the operation. It is inappropriate to analyze that $2 \times 2 \times 2$ table using binary GLMs for the probability of death, since subjects had differing times at risk; it is not sensible to treat a subject who could be observed for 3 months and a subject who could be observed for 97 months as identical trials with the same probability. To use age and valve type as predictors in a model for frequency of death, the proper baseline is not the number of subjects but rather the total time that subjects were at risk. Thus, we model the *rate* of death.

The *time at risk* for a subject is their follow-up time of observation. For a given age and valve type, the total time at risk is the sum of the times at risk for all subjects in that cell (those who died and those censored). Table 9.11 lists those total times in months. The sample rate, also shown in that table, divides the number of deaths by total time at risk. For instance, 4 deaths in 1259 months of observation occurred for younger subjects with aortic valve replacement, so their sample rate is $4/1259 = 0.0032$.

We now model effects of age and valve type on the rate. Let a be a dummy variable for age, with $a_1 = 0$ for the younger age group and $a_2 = 1$ for the older group. Let v be a dummy variable for valve type, with $v_1 = 0$ for aortic and $v_2 = 1$ for mitral. Let n_{ij} denote the number of deaths for age a_i and valve type v_j, with expected value μ_{ij} for total time at risk t_{ij}. Given t_{ij},

TABLE 9.12 Fit to Table 9.11 for Poisson Regression Models

Age		Log Link		Identity Link	
		Aortic	Mitral	Aortic	Mitral
< 55	Number of deaths	2.28	2.72	3.16	1.19
	Death rate	0.0018	0.0013	0.0025	0.0006
55 +	Number of deaths	8.72	7.28	9.17	7.48
	Death rate	0.0062	0.0044	0.0065	0.0046

the expected rate is μ_{ij}/t_{ij}. The model

$$\log(\mu_{ij}/t_{ij}) = \alpha + \beta_1 a_i + \beta_2 v_j \tag{9.18}$$

assumes a lack of interaction in the effects.

Model fitting uses standard iterative methods, treating $\{n_{ij}\}$ as independent Poisson variates with means $\{\mu_{ij}\}$. This is done conditional on $\{t_{ij}\}$. Table 9.12 presents the fitted death counts and estimated rates. The estimated effects are

$$\hat{\beta}_1 = 1.221 \quad (\text{SE} = 0.514), \qquad \hat{\beta}_2 = -0.330 \quad (\text{SE} = 0.438).$$

There is evidence of an age effect. Given valve type, the estimated rate for the older age group is exp(1.221) = 3.4 times that for the younger age group. The 95% Wald confidence interval for β_1 of $1.221 \pm 1.96(0.514)$ translates to $(1.2, 9.3)$ for the true multiplicative effect $\exp(\beta_1)$. (The likelihood-ratio confidence interval is (1.3, 10.4).) The study contains much censored data. Of the 109 patients, only 21 died during the study period. Both effect estimates are imprecise. Note, though, that the analysis uses all 109 patients through their contributions to the times at risk.

Goodness-of-fit statistics comparing $\{n_{ij}\}$ to fitted values $\{\hat{\mu}_{ij}\}$ are $G^2 = 3.2$ and $X^2 = 3.1$. The residual df = 1, since the four response counts have three parameters. The mild evidence of lack of fit corresponds to evidence of interaction between valve type and age. However, the model without valve-type effects [i.e., $\beta_2 = 0$ in (9.18)] fits nearly as well, with $G^2 = 3.8$ and $X^2 = 3.8$ (df = 2). Models omitting age effects fit poorly.

The corresponding model with identity link

$$\mu_{ij} = \alpha t_{ij} + \beta_1 a_i t_{ij} + \beta_2 v_j t_{ij}$$

shows a good fit, with $G^2 = 1.1$ and $X^2 = 1.1$ (df = 1). Table 9.12 shows the fit. Substantive conclusions are similar. The estimate $\hat{\beta}_1 = 0.0040$ (SE = 0.0014) then represents an estimated difference in death rates between the older and younger age groups for each valve type.

9.7.3 Modeling Survival Times*

A method for modeling survival times relates to the Poisson loglinear model for rates. This method focuses on times until death rather than on numbers of deaths. Let T denote the time to some event, such as death or such as product failure in a reliability study. Let $f(t)$ denote the probability density function (pdf) and $F(t)$ the cdf of T. A connection exists between ML estimation using a Poisson likelihood for numbers of events and a negative exponential likelihood for T (Aitkin and Clayton 1980).

A subject having $T = t$ contributes $f(t)$ to the likelihood. For a subject whose censoring time equals t, we know only that $T > t$. Thus, this subject contributes $P(T > t) = 1 - F(t)$. Using the indicator $w_i = 1$ for death and 0 for censoring for subject i, the survival-time likelihood for n independent observations is

$$\prod_{i=1}^{n} f(t_i)^{w_i}\left[1 - F(t_i)\right]^{1-w_i}.$$

The log likelihood equals

$$\sum_i w_i \log\left[f(t_i)\right] + \sum_i (1 - w_i)\log\left[1 - F(t_i)\right]. \tag{9.19}$$

Further analysis requires a parametric form for f and a model for the dependence of its parameters on explanatory variables.

Most survival models focus on the *rate* at which death occurs rather than on $E(T)$. The *hazard function*

$$h(t) = \frac{f(t)}{1 - F(t)} = \lim_{\epsilon \downarrow 0} \frac{P\left[t < T < t + \epsilon \,|\, T > t\right]}{\epsilon}$$

represents the instantaneous rate of death for subjects who have survived to time t. A simple density for survival modeling is the negative exponential. The pdf is

$$f(t) = \lambda e^{-\lambda t}, \qquad t > 0.$$

The cdf is $F(t) = 1 - e^{-\lambda t}$ for $t > 0$, and $E(T) = \lambda^{-1}$. The hazard function is

$$h(t) = \lambda, \qquad t > 0,$$

constant for all t.

Now we include explanatory variables $\mathbf{x}$. Suppose that the hazard function for a negative exponential survival distribution is

$$h(t; \mathbf{x}) = \lambda \exp(\boldsymbol{\beta}'\mathbf{x}). \tag{9.20}$$

That is, the distribution for T has parameter depending on $\mathbf{x}$ through (9.20). The choice of functional form (9.20) for explanatory variable effects ensures the hazard is nonnegative at all $\mathbf{x}$. For instance, loglinear model (9.18) corresponds to a multiplicative model of type (9.20) for the rate itself.

Now, consider the log likelihood (9.19) with $f(t)$ equal to the negative exponential density with parameter $\lambda \exp(\boldsymbol{\beta}'\mathbf{x})$. For subject i, let

$$\mu_i = t_i \lambda \exp(\boldsymbol{\beta}'\mathbf{x}_i).$$

With this substitution, the log likelihood simplifies to

$$\sum_i w_i \log \mu_i - \sum_i \mu_i - \sum_i w_i \log t_i.$$

The first two terms involve $\boldsymbol{\beta}$. This part is identical to the log likelihood for independent Poisson variates $\{w_i\}$ with expected values $\{\mu_i\}$. In this application $\{w_i\}$ are binary rather than Poisson, but that is irrelevant to the process of maximizing with respect to $\boldsymbol{\beta}$. This process is equivalent to maximizing the likelihood for the Poisson loglinear model

$$\log \mu_i - \log t_i = \log \lambda + \boldsymbol{\beta}'\mathbf{x}_i$$

with offset $\log(t_i)$, using observations $\{w_i\}$. When we sum terms in the log likelihood for subjects having a common value of $\mathbf{x}$, the observed data are the numbers of deaths (Σw_i) at each setting of $\mathbf{x}$, and the offset is the log of (Σt_i) at each setting.

The assumption of constant hazard over time is often not sensible. As products wear out, their failure rate increases. A generalization divides the time scale into disjoint time intervals and assumes constant hazard in each, namely,

$$h(t;\mathbf{x}) = \lambda_k \exp(\boldsymbol{\beta}'\mathbf{x})$$

for t in interval k, $k = 1, \ldots$. A separate hazard rate applies to each piece of the time scale. Consider the contingency table for numbers of deaths, in which one dimension is a discrete time scale and other dimensions represent categorical explanatory variables. Holford (1980) and Laird and Olivier (1981) showed that Poisson loglinear models and likelihoods for this table are equivalent to loglinear hazard models and likelihoods that assume piecewise exponential hazards for the survival times.

For short time intervals, the piecewise exponential approach is essentially nonparametric, making no assumption about the dependence of the hazard on time. This suggests the generalization of model (9.20) that replaces λ by an unspecified function $\lambda(t)$, so that

$$h(t;\mathbf{x}) = \lambda(t)\exp(\boldsymbol{\beta}'\mathbf{x}).$$

This is the Cox *proportional hazards* model. Its ratio of hazards

$$h(t;\mathbf{x}_1)/h(t;\mathbf{x}_2) = \exp\left[\boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)\right]$$

is the same for all t.

TABLE 9.13 Number of Deaths from Lung Cancer

Follow-up Time Interval (months)	Disease Stage:	Histology[a] I 1	I 2	I 3	II 1	II 2	II 3	III 1	III 2	III 3
0–2		9	12	42	5	4	28	1	1	19
		(157	134	212	77	71	130	21	22	101)
2–4		2	7	26	2	3	19	1	1	11
		(139	110	136	68	63	72	17	18	63)
4–6		9	5	12	3	5	10	1	3	7
		(126	96	90	63	58	42	14	14	43)
6–8		10	10	10	2	4	5	1	1	6
		(102	86	64	55	42	21	12	10	32)
8–10		1	4	5	2	2	0	0	0	3
		(88	66	47	50	35	14	10	8	21)
10–12		3	3	4	2	1	3	1	0	3
		(82	59	39	45	32	13	8	8	14)
12 +		1	4	1	2	4	2	0	2	3
		(76	51	29	42	28	7	6	6	10)

[a] Values in parentheses represent total follow-up.

Source: Reprinted with permission from the Biometric Society, based on Holford (1980).

9.7.4 Lung Cancer Survival Example*

Table 9.13 describes survival for 539 males diagnosed with lung cancer. The prognostic factors are histology (H) and stage (S) of disease. For a piecewise exponential hazard approach, the time scale for follow-up (T) was divided into two-month intervals.

Let μ_{ijk} denote the expected number of deaths and t_{ijk} the total time at risk for histology i and state of disease j, in follow-up time interval k. The model

$$\log(\mu_{ijk}/t_{ijk}) = \lambda + \lambda_i^H + \lambda_j^S + \lambda_k^T \tag{9.21}$$

has residual $G^2 = 43.9$ (df = 52). All models assuming no interaction between follow-up time interval and either prognostic factor are proportional hazards models, since they have the same effects of histology and stage of disease for each time interval. Table 9.14 summarizes results of fitting several such models. Although stage of disease is an important prognostic factor, histology did not contribute significant additional information.

For model (9.21), the effects of stage of disease satisfy

$$\hat{\lambda}_2^S - \hat{\lambda}_1^S = 0.470 \quad (\text{SE} = 0.174),$$

$$\hat{\lambda}_3^S - \hat{\lambda}_1^S = 1.324 \quad (\text{SE} = 0.152).$$

TABLE 9.14 Results for Poisson Regression Models of Proportional Hazards Form with Table 9.13

Effects[a]	G^2	df
T	170.7	56
$T + H$	143.1	54
$T + S$	45.8	54
$T + S + H$	43.9	52
$T + S + H + S \times H$	41.5	48

[a] T, time scale for follow-up; H, histology; S, disease stage.

For instance, at a fixed follow-up time for a given histology, the estimated death rate at the third stage of disease is $\exp(1.324) = 3.8$ times that at the first stage. Adding interaction terms between stage and time does not significantly improve the fit (change in $G^2 = 14.9$, change in df $= 12$). The $\{\hat{\lambda}_j^S\}$ are very similar for the simpler model without the histology effects.

9.7.5 Analyzing Weighted Data*

The process of fitting a loglinear model with an offset is also useful in other applications. For expected frequencies $\{\mu_i\}$ and fixed constants $\{t_i\}$, consider a model

$$\log(\mu_i/t_i) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots .$$

Standard loglinear models have $\{t_i = 1\}$. The general form is useful for the analysis of categorical data with sampling designs more complex than simple random sampling.

Many surveys have sampling designs employing stratification and/or clustering. Case weights inflate or deflate the influence of each observation according to features of that design. Adding the case weights for subjects in a particular cell i provides a total weighted frequency for that cell. The average cell weight z_i is defined to be the total weighted frequency divided by the cell count. Conditional on $\{z_i\}$, loglinear models for the weighted expected frequencies $\{z_i\,\mu_i = \mu_i/t_i\}$ with $t_i = z_i^{-1}$ express the model as a standard loglinear model for $\{\log \mu_i\}$, with offset $\{\log t_i = -\log z_i\}$. Fitting this model provides appropriate parameter estimates and standard errors (Clogg and Eliason 1987).

9.8 EMPTY CELLS AND SPARSENESS IN MODELING CONTINGENCY TABLES

Contingency tables having small cell counts are said to be *sparse*. We end this chapter by discussing effects of sparse tables on model fitting. Sparse

tables occur when the sample size n is small. They also occur when n is large but so is the number of cells. Sparseness is common in tables with many variables. The following discussion refers to a generic contingency table and model, with cell counts $\{n_i\}$ and expected frequencies $\{\mu_i\}$ for n observations in N cells.

9.8.1 Empty Cells: Sampling versus Structural Zeros

Sparse tables usually contain cells with $n_i = 0$. These *empty cells* are of two types: *sampling zeros* and *structural zeros*. In most cases, even though $n_i = 0$, $\mu_i > 0$. It is possible to have observations in the cell, and $n_i > 0$ with sufficiently large n. This empty cell is called a *sampling zero*. The empty cells in Table 9.1 for the student survey are sampling zeros.

An empty cell in which observations are impossible is called a *structural zero*. For such cells $\mu_i = 0$ and necessarily $\hat{\mu}_i = 0$ and $n_i = 0$ regardless of n. For a table that cross classifies cancer patients on their gender, race, and type of cancer, some cancers (e.g., prostate cancer, ovarian cancer) are gender specific. Thus, certain cells have structural zeros. Contingency tables with structural zeros are called *incomplete tables*.

Sampling zeros are part of the data set. A count of 0 is a permissible outcome for a Poisson or multinomial variate. It contributes to the likelihood function and model fitting. A structural zero, on the other hand, is not an observation and is not part of the data. Sampling zeros are much more common than structural zeros, and the remaining discussion refers to them.

9.8.2 Existence of Estimates in Loglinear / Logit Models

Sampling zeros can affect the existence of finite ML estimates of loglinear and logit model parameters. Haberman (1973b, 1974a), generalizing work by Birch (1963) and Fienberg (1970b), studied this. Let $\mathbf{n}$ denote the vector of cell counts and $\boldsymbol{\mu}$ their expected values. Haberman showed results 1 through 5 for Poisson sampling, but by result 6 they apply also to multinomial sampling.

1. The log-likelihood function is a strictly concave function of $\log \boldsymbol{\mu}$.
2. If a ML estimate of $\boldsymbol{\mu}$ exists, it is unique and satisfies the likelihood equations $\mathbf{X}'\mathbf{n} = \mathbf{X}'\hat{\boldsymbol{\mu}}$. Conversely, if $\hat{\boldsymbol{\mu}}$ satisfies the model and also the likelihood equations, it is the ML estimate of $\boldsymbol{\mu}$.
3. If all $n_i > 0$, ML estimates of loglinear model parameters exist.
4. Suppose that ML parameter estimates exist for a loglinear model that equates observed and fitted counts in certain marginal tables. Then those marginal tables have uniformly positive counts.
5. If ML estimates exist for a model M, they also exist for any special case of M.

6. For any loglinear model, the ML estimates $\hat{\boldsymbol{\mu}}$ are identical for multinomial and independent Poisson sampling, and those estimates exist in the same situations.

To illustrate, consider the saturated model. By results 2 and 3, when all $n_i > 0$, the ML estimate of $\boldsymbol{\mu}$ is $\mathbf{n}$. By result 4, parameter estimates do not exist when any $n_i = 0$. Model parameter estimates are contrasts of $\{\log \hat{\mu}_i\}$, and since $\hat{\boldsymbol{\mu}} = \mathbf{n}$ for the saturated model, the estimates are finite only when all $n_i > 0$.

For unsaturated models, by results 3 and 4 ML estimates exist when all $n_i > 0$ and do not exist when any count is zero in the set of sufficient marginal tables. Suppose that at least one $n_i = 0$ but the sufficient marginal counts are all positive. For hierarchical loglinear models, Glonek et al. (1988) showed that the positivity of the sufficient counts implies the existence of ML estimates if and only if the model is decomposable (Note 8.2), which includes the conditional independence models. Models having all pairs of variables associated, however, are more complex. For model (XY, XZ, YZ), for instance, ML estimates exist when only one $n_i = 0$ but may not exist when at least two cells are empty. For instance, ML estimates do not exist for Table 9.15, even though all sufficient statistics (the two-way marginal totals) are positive (Problem 9.47).

Haberman showed that the supremum of the likelihood function is finite. This motivated him to define *extended ML* estimators of $\boldsymbol{\mu}$. These always exist but may equal 0 and, falling on the boundary, need not have the same properties as regular ML estimators [see also Baker et al. (1985)]. A sequence of estimates satisfying the model that converges to the extended estimate has log likelihood approaching its supremum. In this extended sense, $\hat{\mu}_i = 0$ is the ML estimate of μ_i for the saturated model when $n_i = 0$, and one can have infinite loglinear parameter estimates.

When a sufficient marginal count for a factor equals zero, infinite estimates occur for that term. For instance, when a XY marginal total equals zero, infinite estimates occur among $\{\hat{\lambda}_{ij}^{XY}\}$ for loglinear models such as (XY, XZ, YZ), and infinite estimates occur among $\{\hat{\beta}_i^X\}$ for the effect of X on Y in logit models. Sometimes, however, not even infinite estimates exist. An example is estimating the log odds ratio when both entries in a row or column of a 2×2 table equal 0.

TABLE 9.15 Data for Which ML Estimates Do Not Exist for Model (XY, XZ, YZ)[a]

	Z:	1		2	
X	Y:	1	2	1	2
1		0	*	*	*
2		*	*	*	0

[a]Cells containing * may contain any positive numbers.

A value of ∞ (or $-\infty$) for a ML parameter estimate implies that ML fitted values equal 0 in some cells, and some odds ratio estimates equal ∞ or 0. One potential indicator is when the iterative fitting process does not converge, typically because an estimate keeps increasing from cycle to cycle. Most software, however, is fooled after a certain point in the iterative process by the nearly flat likelihood. It reports convergence, but because of the very slight curvature of the log likelihood, the estimated standard errors (based on inverting the information matrix of second partial derivatives) are extremely large and numerically unstable. Slight changes in the data then often cause dramatic changes in the estimates and their standard errors. A danger with sparse data is that one might not realize that a true estimated effect is infinite and, as a consequence, report estimated effects and results of statistical inferences that are invalid and highly unstable.

Many ML analyses are unharmed by empty cells. Even when a parameter estimate is infinite, this is not fatal to data analysis. The likelihood-ratio confidence interval for the true log odds ratio has one endpoint that is finite. For instance, when $n_{11} = 0$ but other $n_{ij} > 0$ in a 2×2 table, $\log \hat{\theta} = -\infty$ and a confidence interval has form $(-\infty, U)$ for some finite upper bound U. When the pattern of empty cells forces certain fitted values for a model to equal 0, this affects the df for testing model fit (Haslett 1990).

9.8.3 Clinical Trials Example

Table 9.16 shows results of a clinical trial conducted at five centers. The purpose was to compare an active drug to placebo for treating fungal infections, with a binary (success, failure) response. For these data, let Y = response, X = treatment ($x_1 = 1$ for active drug and $x_2 = 0$ for placebo), and Z = center.

Centers 1 and 3 had no successes. Thus, the 5×2 marginal table relating response to center, collapsed over treatment, contains zero counts. The last two columns of Table 9.16 show this marginal table. Infinite ML estimates occur for terms in loglinear or logit models containing the YZ association. An example is the logit model

$$\text{logit}\big[P(Y = 1 \,|\, X = i, Z = k)\big] = \beta x_i + \beta_k^Z.$$

(We omit the intercept, so the $\{\beta_k^Z\}$ need no constraint; then, these refer to center effects rather than contrasts between centers and a baseline center.) The likelihood function increases continually as β_1^Z and β_3^Z decrease toward $-\infty$; that is, as the logit decreases toward $-\infty$, so the fitted probability of success decreases toward the ML estimate of 0 for those centers.

The counts in the 2×2 marginal table relating response to treatment, shown in the bottom panel of Table 9.16, are all positive. The empty cells in Table 9.16 affect the center estimates, but not the treatment estimate, for this logit model. In the limit as the log likelihood increases, the fitted values have a log odds ratio $\hat{\beta} = 1.55$ (SE = 0.70). Most software reports this, but

TABLE 9.16 Clinical Trial Relating Treatment to Response with *XY* and *YZ* Marginal Tables[a]

		Response		*YZ* Marginal	
Center	Treatment	Success	Failure	Success	Failure
1	Active drug	0	5	0	14
	Placebo	0	9		
2	Active drug	1	12	1	22
	Placebo	0	10		
3	Active drug	0	7	0	12
	Placebo	0	5		
4	Active drug	6	3	8	9
	Placebo	2	6		
5	Active drug	5	9	7	21
	Placebo	2	12		
XY	Active drug	12	36		
marginal	Placebo	4	42		

[a] *X*, Treatment; *Y*, response; *Z*, center.

Source: Data courtesy of Diane Connell, Sandoz Pharmaceuticals Corporation.

instead of $\hat{\beta}_1^Z = \hat{\beta}_3^Z = -\infty$ reports large numbers with extremely large standard errors. For instance, PROC GENMOD in SAS reports values of about -26 for $\hat{\beta}_1^Z$ and $\hat{\beta}_3^Z$, with standard errors of about 200,000.

The treatment estimate $\hat{\beta} = 1.55$ also results from deleting centers 1 and 3 from the analysis. When a center contains responses of only one type, it provides no information about this odds ratio. (It does provide information about the size of some other measures, such as the difference of proportions.) In fact, such tables also make no contribution to standard tests of conditional independence, such as the Cochran–Mantel–Haenszel test (Section 6.3.2) and exact test (Section 6.7.5).

An alternative strategy in multicenter analyses combines centers of a similar type. Then, if each resulting partial table has responses with both outcomes, the inferences use all data. For Table 9.16, perhaps centers 1 and 3 are similar to center 2, since the success rate is very low for that center. Combining these three centers and refitting the model to this table and the tables for the other two centers yields $\hat{\beta} = 1.56$ (SE = 0.70). Usually, this strategy produces results similar to deleting the table with no outcomes of a particular type.

9.8.4 Effect of Small Samples on X^2 and G^2

Although empty cells and sparse tables need not affect parameter estimates of interest, they can cause sampling distributions of goodness-of-fit statistics to be far from chi-squared. The true sampling distributions converge to

chi-squared as $n \to \infty$, for a fixed number of cells N. The adequacy of the chi-squared approximation depends both on n and N.

Cochran studied the chi-squared approximation for X^2 in several articles. In 1954, he suggested that to test independence with df > 1, a minimum expected value $\mu_i \approx 1$ is permissible as long as no more than about 20% of $\mu_i < 5$. Koehler (1986), Koehler and Larntz (1980), and Larntz (1978) showed that X^2 applies with smaller n and more sparse tables than G^2. The distribution of G^2 is usually poorly approximated by chi-squared when n/N is less than 5. Depending on the sparseness, P-values based on referring G^2 to a chi-squared distribution can be too large or too small. When most μ_i are smaller than 0.5, treating G^2 as chi-squared gives a highly conservative test; when H_0 is true, reported P-values tend to be much larger than true ones. When most μ_i are between 0.5 and 4, G^2 tends to be too liberal; the reported P-value tends to be too small.

The size of n/N that produces adequate approximations for X^2 tends to decrease as N increases (Koehler and Larntz 1980). However, the approximation tends to be poor for sparse tables containing both small and moderately large μ_i (Haberman 1988). It is difficult to give a guideline that covers all cases. For other discussion, see Cressie and Read (1989) and Lawal (1984).

For fixed n and N, the chi-squared approximation is better for tests with smaller df. For instance, in testing conditional independence in $I \times J \times K$ tables, $G^2[(XZ, YZ) \,|\, (XY, XZ, YZ)]$ (with df $= (I-1)(J-1)$) is closer to chi-squared than $G^2(XZ, YZ)$ [with df $= K(I-1)(J-1)$]. The ordinal test of H_0: $\beta = 0$ with the homogeneous linear-by-linear XY association model (9.10) has df $= 1$, and behaves even better.

9.8.5 Model-Based Tests and Sparseness

From (9.3) and (9.4), the model-based statistics $G^2(M_0 \,|\, M_1)$ and $X^2(M_0 \,|\, M_1)$ depend on the data only through the fitted values, and hence only through minimal sufficient statistics for the more complex model. These statistics have null distributions converging to chi-squared as the expected values of the minimal sufficient statistics grow. For most loglinear models, these sufficient statistics refer to marginal tables. Marginal totals are more nearly normally distributed than are single cell counts. Thus, $G^2(M_0 \,|\, M_1)$ and $X^2(M_0 \,|\, M_1)$ converge to their limiting chi-squared distribution more quickly than does $G^2(M_0)$ and $X^2(M_0)$, which depend also on individual cell counts.

When $\{\hat{\mu}_i\}$ are small but the sufficient marginal totals for M_1 are mostly in at least the range 5 to 10, the chi-squared approximation is usually adequate for model comparison statistics. Haberman (1977a) provided theoretical justification.

9.8.6 Alternative Asymptotics and Alternative Statistics

When large-sample approximations are inadequate, exact small-sample methods are an alternative. When they are infeasible, it is often possible to

approximate exact distributions precisely using Monte Carlo methods (e.g., Booth and Butler 1999; Forster et al. 1996; Kim and Agresti 1997; Mehta et al. 1988).

An alternative approach uses sparse asymptotic approximations that apply when the number of cells N increases as n increases. For this approach, $\{\mu_i\}$ need not increase, as they must do in the usual (fixed N, $n \to \infty$) large-sample theory. For goodness-of-fit testing of a specified multinomial, Koehler and Larntz (1980) showed that a standardized version of G^2 has an approximate normal distribution for very sparse tables. Koehler (1986) presented limiting normal distributions for G^2 for use in testing models having direct ML estimates. McCullagh (1986) reviewed ways of handling sparse tables and presented an alternative approximation for G^2. Zelterman (1987) gave normal approximations for X^2 and proposed an alternative statistic.

9.8.7 Adding Constants to Cells of a Contingency Table

Empty cells and sparse tables can cause problems with existence of estimates for loglinear model parameters, estimation of odds ratios, performance of computational algorithms, and asymptotic approximations of chi-squared statistics. However, they need not be problematic. The likelihood can still be maximized, a point estimate of ∞ for an effect still usually has a finite lower bound for a likelihood-based confidence interval, and one can use small-sample inferential methods rather than asymptotic ones.

One way to obtain finite estimates of all effects and ensure convergence of fitting algorithms is to add a small constant to cell counts. Some algorithms add $\frac{1}{2}$ to each cell, as Goodman (1964b, 1970, 1971a) recommended for saturated models. An example of the beneficial effect of this for a saturated model is bias reduction for estimating an odds ratio in a 2×2 table (Gart 1966; Gart and Zweiful 1967). Adding $\frac{1}{2}$ to each cell before fitting an unsaturated model smooths the data too much, however, causing havoc with sampling distributions. This operation has too conservative an influence on estimated effects and test statistics. The effect is very severe with a large number of cells.

Even for a saturated model, adding $\frac{1}{2}$ to each cell is not a panacea for all purposes. When the ordinary ML estimate of an odds ratio is infinite, the estimate after adding $\frac{1}{2}$ to each cell is finite, as are the endpoints of any confidence interval. However, it is more sensible to use an upper bound of ∞ for the odds ratio, since no sample evidence suggests that the odds ratio falls below any given value.

When in doubt about the effect of sparse data, one should perform a sensitivity analysis. For example, for each possibly influential observation, delete it or move it to another cell to see how results vary with small perturbations to the data. Influence diagnostics for GLMs (Williams 1987) are also useful for this purpose. Often, some associations are not affected by empty cells and give stable results for the various analyses, whereas others

that are affected are highly unstable. Use caution in making conclusions about an association if small changes in the data are influential.

Later chapters show ways to smooth data in a less *ad hoc* manner than adding arbitrary constants to cells. These include random effects models (Section 12.3) and Bayesian methods (Section 15.2).

NOTES

Section 9.1: Association Graphs and Collapsibility

9.1. Darroch et al. (1980) defined a class of *graphical models* that contains the family of decomposable models (see Note 8.2). For expositions on graphical models and their relevant *independence graphs*, which show the conditional independence structure, see also Anderson and Böckenholt (2000), Edwards (2000), Edwards and Kreiner (1983), Kreiner (1998), Lauritzen (1996), and Whittaker (1990). Whittaker (1990, Sec. 12.5) summarized connections with various definitions of collapsibility.

9.2 For $I \times J \times 2$ tables, the collapsibility conditions (Section 9.1.2) are necessary as well as sufficient (Simpson 1951; Whittemore 1978). For $I \times J \times K$ tables, Ducharme and Lepage (1986) showed the conditions are necessary and sufficient for the odds ratios to remain the same no matter how the levels of Z are pooled (i.e., no matter how Z is partially collapsed).

Darroch (1962) defined a *perfect* table as one for which for all i, j, k,

$$\sum_i \frac{\pi_{ij+}\,\pi_{i+k}}{\pi_{i++}} = \pi_{+j+}\,\pi_{++k}, \qquad \sum_j \frac{\pi_{+jk}\,\pi_{ij+}}{\pi_{+j+}} = \pi_{i++}\,\pi_{++k},$$

$$\sum_k \frac{\pi_{i+k}\,\pi_{+jk}}{\pi_{++k}} = \pi_{i++}\,\pi_{+j+}.$$

For perfect tables, homogeneous association implies that

$$\{\pi_{ijk} = \pi_{ij+}\,\pi_{i+k}\,\pi_{+jk}/\pi_{i++}\,\pi_{+j+}\,\pi_{++k}\}$$

and conditional odds ratios are identical to marginal odds ratios. Whittemore (1978) used perfect tables to illustrate that for $I \times J \times K$ tables with $K > 2$, conditional and marginal odds ratios can be identical even when no pair of variables is conditionally independent. See also Davis (1986b).

Suppose that the difference of proportions or relative risk, computed for a binary response Y and predictor X, is the same at every level of Z. If Z is independent of X in the marginal XZ table or if Z is conditionally independent of Y given X, the measure has the same value in the marginal XY table (Shapiro 1982). Thus, for factorial designs with the same number of observations at each combination of levels, the difference of proportions and relative risk are collapsible. See also Wermuth (1987).

Section 9.2: Model Selection and Comparison

9.3. Articles on loglinear model selection include Aitkin (1979, 1980), Benedetti and Brown (1978), Brown (1976), Goodman (1970, 1971a), Wermuth (1976), and Whittaker and Aitkin (1978). When a certain model holds, G^2/df has an asymptotic mean of 1. Goodman (1971a) recommended this index for comparing fits. Smaller values represent better fits.

9.4. Kullback et al. (1962) and Lancaster (1951) were among the first to partition chi-squared statistics in multiway tables. Goodman (1970) and Plackett (1962) noted difficulties with their approaches. When observations have distribution in the natural exponential family, Simon (1973) showed $G^2(M_0 \mid M_1) = 2\Sigma_i \hat{\mu}_{1i} \log(\hat{\mu}_{1i}/\hat{\mu}_{0i})$ whenever models are linear in the natural parameters. See Lang (1996b) for partitionings for more complex models.

Section 9.4: Modeling Ordinal Associations

9.5. Goodman (1979a) stimulated research on loglinear models for ordinal data. His work extended Haberman (1974b), who expressed the λ^{XY} association term with an expansion in orthogonal polynomials. For more general ordinal models for multiway tables, see Agresti (1984), Becker (1989a), Becker and Clogg (1989), and Goodman (1986).

Section 9.6: Association Models, Correlation Models, and Correspondence Analysis

9.6. Early articles on the *RC* model include Goodman (1979a, 1981a, b) and Andersen (1980, pp. 210–216), apparently partly motivated by earlier work of G. Rasch (see Andersen 1995). Anderson and Böckenholt (2000), Becker (1989a, b, 1990), Becker and Clogg (1989), Chuang et al. (1985), and Goodman (1985, 1986, 1996) discussed generalizations for multiway tables. Anderson (1984) discussed a related model. Anderson and Vermunt (2000) showed that *RC* and related association models arise when observed variables are conditionally independent given a latent variable that is conditionally normal, given the observed variables. Their work generalizes results in Lauritzen and Wermuth (1989) and discussion by Whittaker of van der Heijden et al. (1989). See also de Falguerolles et al. (1995). Clogg and Shihadeh (1994) surveyed association models and related correlation models.

9.7. Kendall and Stuart (1979, Chap. 33) surveyed basic canonical correlation methods for contingency tables. See also Williams (1952), who discussed earlier work by R. A. Fisher and others. Karl Pearson often analyzed tables by assuming an underlying bivariate normal distribution (Section 16.1). For estimating that distribution's correlation, see Becker (1989b), Goodman (1981b), Kendall and Stuart (1979, Chaps. 26 and 33), Lancaster (1969, Chap. X), the Pearson (1904) tetrachoric correlation for 2×2 tables, and the Lancaster and Hamdan (1964) polychoric correlation for $I \times J$ tables.

9.8. Correspondence analysis gained popularity in France under the influence of Benzécri (see, e.g., 1973). Goodman (1996) attributed its origins to H. O. Hartley, publishing under his original German name (Hirschfeld, 1935). Greenacre (1993) related it to the singular value decomposition of a matrix. For other discussion, see Escoufier (1982), Friendly (2000, Chap. 5), Goodman (1986, 1996, 2000), Michailidis and de Leeuw (1998), van der Heijden and de Leeuw (1985), and van der Heijden et al. (1989). Gabriel (1971) discussed related work on biplots.

Section 9.7: Poisson Regression for Rates

9.9. Another application using offsets is table standardization (Section 8.7.4). For analyses of rate data, see Breslow and Day (1987, Sec. 4.5), Freeman and Holford (1980), Frome (1983), and Hoem (1987). Articles dealing with grouped survival data, particularly loglinear and logit models for survival probabilities, include Aranda-Ordaz (1983), Larson (1984), Prentice and Gloeckler (1978), Schluchter and Jackson (1989), Stokes et al. (2000, Chap. 17), and Thompson (1977). Aitkin and Clayton (1980) discussed exponential survival models and also presented similar models having hazard functions

for Weibull or extreme-value survival distributions. Log likelihood (9.19) actually applies only for *noninformative* censoring mechanisms. It does not make sense if subjects tend to withdraw from the study because of factors related to it, perhaps because of health effects related to one of the treatments.

9.10. Lindsey and Mersch (1992) showed a clever way to use loglinear models to fit exponential family distributions $f(y;\theta)$ of form (4.14) with ϕ known. One breaks the response scale into intervals $\{(y_k - \Delta_k/2, y_k + \Delta_k/2)\}$. Counts in those intervals follow a multinomial with probabilities approximated by $\{f(y_k, \theta)\Delta_k\}$. The log expected count approximations are linear in θ with an offset.

PROBLEMS

Applications

9.1 Use odds ratios in Table 8.3 to illustrate the collapsibility conditions.

a. For (A, C, M), all conditional odds ratios equal 1.0. Explain why all reported marginal odds ratios equal 1.0.

b. For (AC, M), explain why **(i)** all conditional odds ratios are the same as the marginal odds ratios, and **(ii)** all $\hat{\mu}_{ac+} = n_{ac+}$.

c. For (AM, CM), explain why **(i)** the AC conditional odds ratios of 1.0 need not be the same as the AC marginal odds ratio, **(ii)** the AM and CM conditional odds ratios are the same as the marginal odds ratios, and **(iii)** all $\hat{\mu}_{a+m} = n_{a+m}$ and $\hat{\mu}_{+cm} = n_{+cm}$.

d. For (AC, AM, CM), explain why **(i)** no conditional odds ratios need be the same as the related marginal odds ratios, and **(ii)** the fitted marginal odds ratios must equal the sample marginal odds ratios.

9.2 Table 9.17 summarizes a study with variables age of mother (A), length of gestation (G) in days, infant survival (I), and number of cigarettes smoked per day during the prenatal period (S). Treat G and I as response variables and A and S as explanatory.

a. Explain why a loglinear model should include the λ^{AS} term.

b. Fit the models $(AGIS)$, (AGI, AIS, AGS, GIS), (AG, AI, AS, GI, GS, IS), and (AS, G, I). Identify a subset of models nested between two of these that may fit well. Select one such model.

c. Use **(i)** forward selection, and **(ii)** backward elimination to build a model. Compare the results of the strategies, and interpret the models chosen.

9.3 Refer to Table 2.13. Consider the nested set $\{(DVP), (DP, VP, DV), (VP, DV), (P, DV), (D, V, P)\}$. Partition chi-squared to compare the four pairs, ensuring that the overall type I error probability for the four comparisons does not exceed $\alpha = 0.10$. Which model would you select, using a backward comparison starting with (DVP)? Show that the final

TABLE 9.17 Data for Problem 9.2

Age	Smoking	Gestation	Infant Survival	
			No	Yes
< 30	< 5	≤ 260	50	315
		> 260	24	4012
	5 +	≤ 260	9	40
		> 260	6	459
30 +	< 5	≤ 260	41	147
		> 260	14	1594
	5 +	≤ 260	4	11
		> 260	1	124

Source: N. Wermuth, pp. 279–295 in *Proc. 9th International Biometrics Conference*, Vol. 1 (1976). Reprinted with permission from the Biometric Society.

model selected depends on the choice of nested set, by repeating the analysis with (DP, VP, DV), (DP, DV), (P, DV), (D, V, P).

9.4 Consider the loglinear model selection for Table 6.3.

a. Why is it not sensible to consider models omitting the λ^{GM} term?

b. Using forward selection starting with (GM, E, P), show that model (GM, GP, EG, EMP) seems reasonable.

c. Using backward elimination, show that (GM, GP, EMP) or (GM, GP, EG, EMP) seems reasonable.

d. The EMP interaction seems vital. To describe it, show that the effect of extramarital sex on divorce is greater for subjects who had no premarital sex.

e. Use residuals to describe the lack of fit of model (GM, EMP).

9.5 For model (AC, AM, CM) with Table 8.3, the standardized Pearson residual in each cell equals ± 0.63. Interpret, and explain why each one has the same absolute value. By contrast, model (AM, CM) has standardized Pearson residual ± 3.70 in each cell where $M =$ yes (e.g., $+3.70$ when $A = C =$ yes) and ± 12.80 in each cell where $M =$ no (e.g., $+12.80$ when $A = C =$ yes). Interpret.

9.6 Refer to Table 8.8. Conduct a residual analysis with the model of no three-factor interaction to describe the nature of the interaction.

9.7 Perform a residual analysis for the independence model with Table 3.2. Explain why it suggests that the linear-by-linear association model may fit better. Fit it, compare to the independence model, and interpret.

9.8 Refer to Problem 9.7.

a. Using standardized scores, find $\hat{\beta}$. Comment on the strength of association.

b. Fit a model in which job satisfaction scores are parameters. Interpret the estimated scores, and compare the fit to the $L \times L$ model.

9.9 Refer to Table 9.3.

a. For the linear-by-linear association model, construct a 95% confidence interval for the odds ratio using the four corner cells. Interpret.

b. Fit the column effects model. Compare estimated column scores to the equal-interval scores in part (a). Test that the true column scores are equal-interval, given that the model holds. Interpret. Construct a 95% confidence interval for the odds ratio using the four corner cells. Compare to part (a).

9.10 A weak local association may be substantively important for nonlocal categories. Illustrate with the $L \times L$ model for Table 9.9, showing how the estimated odd ratio for the four corner cells compares to the estimated local odds ratio.

9.11 Refer to Table 7.8. Fit the homogeneous linear-by-linear association model, and interpret. Test conditional independence between income (I) and job satisfaction (S), controlling for gender (G), using (**a**) that model, and (**b**) model (IS, IG, SG). Explain why the results are so different.

9.12 Fit the *RC* model to Table 9.3. Interpret the estimated scores. Does it fit better than the uniform association model?

9.13 Replicate the results in Section 9.6 for the correlation and correspondence models with Table 9.9.

9.14 One hundred leukemia patients were randomly assigned to two treatments. During the study, 10 subjects on treatment A died and 18 subjects on treatment B died. The total time at risk was 170.4 years for treatment A and 147.3 years for treatment B. Test whether the two treatments have the same death rates. Compare the rates with a confidence interval.

9.15 For Table 9.11, fit a model in which death rate depends only on age. Interpret the age effect.

9.16 Consider model (9.18). What is the effect on the model parameter estimates, their standard errors, and the goodness-of-fit statistics when (**a**) the times at risk are doubled, but the numbers of deaths stay the

same; (**b**) the times at risk stay the same, but the numbers of deaths double; and (**c**) the times at risk and the numbers of deaths both double.

9.17 Consider Table 9.13. Explain how one could analyze whether the hazard depends on time.

9.18 An article by W. A. Ray et al. (*Amer. J. Epidemiol.* **132**: 873–884, 1992) dealt with motor vehicle accident rates for 16,262 subjects aged 65–84 years, with data on each for up to 4 years. In 17.3 thousand years of observation, the women had 175 accidents in which an injury occurred. In 21.4 thousand years, men had 320 injurious accidents.

a. Find a 95% confidence interval for the true overall rate of injurious accidents.

b. Using a model, compare the rates for men and women.

9.19 A table at the text's Web site (*www.stat.ufl.edu/~aa/cda/cda.html*) shows the number of train miles (in millions) and the number of collisions involving British Rail passenger trains between 1970 and 1984. A Poisson model assuming a constant log rate α over the 14-year period has $\hat{\alpha} = -4.177$ (SE = 0.1325) and $X^2 = 14.8$ (df = 13). Interpret.

9.20 Table 9.18 lists total attendance (in thousands) and the total number of arrests in the 1987–1988 season for soccer teams in the Second Division of the British football league. Let Y = number of arrests for a team, and let t = total attendance. Explain why the model $E(Y) = \mu t$

TABLE 9.18 Data for Problem 9.20

Team	Attendance (thousands)	Arrests	Team	Attendance (thousands)	Arrests
Aston Villa	404	308	Shrewsbury	108	68
Bradford City	286	197	Swindon Town	210	67
Leeds United	443	184	Sheffield Utd.	224	60
Bournemouth	169	149	Stoke City	211	57
West Brom	222	132	Barnsley	168	55
Hudderfield	150	126	Millwall	185	44
Middlesbro	321	110	Hull City	158	38
Birmingham	189	101	Manchester City	429	35
Ipswich Town	258	99	Plymouth	226	29
Leicester City	223	81	Reading	150	20
Blackburn	211	79	Oldham	148	19
Crystal Palace	215	78			

Source: The *Independent* (London), Dec. 21, 1988. Thanks to P. M. E. Altham for showing me these data.

might be plausible. Assuming Poisson sampling, fit it and interpret. Plot arrests against attendance, and overlay the prediction equation. Use residuals to identify teams that had arrest counts much different than expected.

TABLE 9.19 Data for Problem 9.21

	Person-Years		Coronary Deaths	
Age	Nonsmokers	Smokers	Nonsmokers	Smokers
35–44	18,793	52,407	2	32
45–54	10,673	43,248	12	104
55–64	5710	28,612	28	206
65–74	2585	12,663	28	186
75–84	1462	5317	31	102

Source: R. Doll and A. B. Hill, *Natl. Cancer Inst. Monogr.* **19**: 205–268 (1966). See also N. R. Breslow in *A Celebration of Statistics*, ed. A. C. Atkinson and S. E. Fienberg, (New York: Springer-Verlag, 1985).

9.21 Table 9.19 is based on a study with British doctors.

a. For each age, find the sample coronary death rates per 1000 person-years for nonsmokers and smokers. To compare them, take their ratio and describe its dependence on age.

b. Fit a main-effects model for the log rates having four parameters for age and one for smoking. In discussing lack of fit, show that this model assumes a constant ratio of nonsmokers' to smokers' coronary death rates over age.

c. From part (a), explain why it is sensible to add a quantitative interaction of age and smoking. For this model, show that the log ratio of coronary death rates changes linearly with age. Assign scores to age, fit the model, and interpret.

9.22 Analyze Table 9.9 using ordinal logit models. Interpret, and discuss advantages/disadvantages compared to loglinear analyses.

9.23 Refer to Problem 8.6. Analyze these data, using methods of this chapter.

Theory and Methods

9.24 In a $2 \times 2 \times K$ table, the true XY conditional odds ratios are identical, but different from the XY marginal odds ratio. Is there three-factor interaction? Is Z conditionally independent of X or Y? Explain.

9.25 Consider loglinear model (WX, XY, YZ). Explain why W and Z are independent given X alone or given Y alone or given both X and Y. When are W and Y conditionally independent? When are X and Z conditionally independent?

9.26 Suppose that loglinear model (XY, XZ) holds.

a. Find μ_{ij+} and $\log \mu_{ij+}$. Show the loglinear model for the XY marginal table has the same association parameters as $\{\lambda_{ij}^{XY}\}$ in (XY, XZ). Deduce that odds ratios are the same in the XY marginal table as in the partial tables. Using an analogous result for model (XY, YZ), deduce the collapsibility conditions in Section 9.1.2.

b. Calculate $\log \mu_{ij+}$ for model (XY, XZ, YZ), and explain why marginal associations need not equal conditional associations.

9.27 For a four-way table, is the WX conditional association the same as the WX marginal association for the loglinear model **(a)** (WX, XYZ)? and **(b)** (WX, WZ, XY, YZ)? Why?

9.28 Loglinear model M_0 is a special case of loglinear model M_1.

a. Explain why the fitted values for the two models are identical in the sufficient marginal distributions for M_0.

b. Haberman (1974a) showed that when $\{\hat{\mu}_i\}$ satisfy any model that is a special case of M_0, $\Sigma_i \hat{\mu}_{1i} \log \hat{\mu}_i = \Sigma_i \hat{\mu}_{0i} \log \hat{\mu}_i$. Thus, $\hat{\boldsymbol{\mu}}_0$ is the orthogonal projection of $\hat{\boldsymbol{\mu}}_1$ onto the linear manifold of $\{\log \boldsymbol{\mu}\}$ satisfying M_0. Using this, show that $G^2(M_0) - G^2(M_1) = 2\Sigma_i \hat{\mu}_{1i} \log(\hat{\mu}_{1i}/\hat{\mu}_{0i})$.

9.29 Refer to Section 9.2.4. Show that $G^2(M_j | M_{j-1})$ equals G^2 for independence in the 2×2 table comparing columns 1 through $j-1$ with column j.

9.30 For T categorical variables $X_1, \ldots, X_T$, explain why:

a. $G^2(X_1, X_2, \ldots, X_T) = G^2(X_1, X_2) + G^2(X_1X_2, X_3) + \cdots + G^2(X_1X_2 \cdots X_{T-1}, X_T)$.

b. $G^2(X_1 \cdots X_{T-1}, X_T) = G^2(X_1, X_T) + G^2(X_1X_T, X_1X_2) + \cdots + G^2(X_1X_2 \cdots X_{T-1}, X_1X_2 \cdots X_{T-2}X_T)$.

9.31 For $I \times 2$ contingency tables, explain why the linear-by-linear association model is equivalent to the linear logit model (5.5).

9.32 Consider the $L \times L$ model (9.6) with $\{v_j = j\}$ replaced by $\{v_j = 2j\}$. Explain why $\hat{\beta}$ is halved but $\{\hat{\mu}_{ij}\}$, $\{\hat{\theta}_{ij}\}$, and G^2 are unchanged.

9.33 Lehmann (1966) defined (X, Y) to be *positively likelihood-ratio dependent* if their joint density satisfies $f(x_1, y_1)f(x_2, y_2) \geq f(x_1, y_2)f(x_2, y_1)$ whenever $x_1 < x_2$ and $y_1 < y_2$. Then, the conditional distribution of Y (X) stochastically increases as X (Y) increases (Goodman 1981a).

a. For the $L \times L$ model, show that the conditional distributions of Y and of X are stochastically ordered. What is its nature if $\beta > 0$?

b. In row effects model (9.8), if $\mu_i > \mu_h$, show that the conditional distribution of Y is stochastically higher in row i than in row h. Explain why $\mu_1 = \cdots = \mu_I$ is equivalent to the equality of the I conditional distributions within rows.

9.34 Yule (1906) defined a table to be *isotropic* if an ordering of rows and of columns exists such that the local log odds ratios are all nonnegative [see also Goodman (1981a)].

a. Show that a table is isotropic if it satisfies **(i)** the linear-by-linear association model, **(ii)** the row effects model, and **(iii)** the *RC* model.

b. Explain why a table that is isotropic for a certain ordering is still isotropic when adjacent rows or columns are combined.

9.35 Consider the log likelihood for the linear-by-linear association model.

a. Differentiating with respect to β and evaluating at $\beta = 0$ and null estimates of parameters, show that the score function is proportional to

$$\sum_i \sum_j u_i v_j (p_{ij} - p_{i+}p_{+j}).$$

b. Use the delta method to show that its null SE is

$$\left\{ \left[\sum u_i^2 p_{i+} - \left(\sum u_i p_{i+} \right)^2 \right] \left[\sum v_j^2 p_{+j} - \left(\sum v_j p_{+j} \right)^2 \right] \Big/ n \right\}^{1/2}.$$

c. Construct a score statistic for testing independence. Show that it is essentially the correlation test (3.15). [Hirotsu (1982) discussed a family of score tests for ordered alternatives.]

9.36 Given the parenthetical result in Problem 7.33, show that if cumulative logit model (7.24) holds and $|\beta|$ is small, the linear-by-linear association model should fit well with row scores $\{x_i\}$ and "ridit" column scores $\{v_j = [P(Y \leq j - 1) + P(Y \leq j)]/2\}$, with its β parameter about twice β for model (7.24).

9.37 Consider the row effects model (9.8).

a. Show that no loss of generality occurs in letting $\lambda_I^X = \lambda_J^Y = \mu_I = 0$.

b. Show that minimal sufficient statistics are $\{n_{i+}\}$, $\{n_{+j}\}$, and $\{\sum_j v_j n_{ij}, i = 1, \ldots, I\}$, and derive the likelihood equations.

9.38 Show that the column effects model corresponds to a baseline-category logit model for Y that is linear in scores for X, with slope depending on the paired response categories.

9.39 Refer to the homogeneous linear-by-linear association model (9.10).

a. Show that the likelihood equations are, for all i, j, and k,

$$\hat{\mu}_{i+k} = n_{i+k}, \quad \hat{\mu}_{+jk} = n_{+jk}, \quad \sum_i \sum_j u_i v_j \hat{\mu}_{ij+} = \sum_i \sum_j u_i v_j n_{ij+}.$$

b. Show that residual df $= K(I-1)(J-1) - 1$.

c. When $I = J = 2$, explain why it is equivalent to (XY, XZ, YZ).

d. Show how the last likelihood equation above changes for heterogeneous linear-by-linear XY association (9.11). Explain why, in each stratum, the fitted XY correlation equals the sample correlation.

9.40 When model (XY, XZ, YZ) is inadequate and variables are ordinal, useful models are nested between it and (XYZ). For ordered scores $\{u_i\}$, $\{v_j\}$, and $\{w_k\}$, consider

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \beta u_i v_j w_k. \quad (9.22)$$

a. Define $\theta_{ijk} = \theta_{ij(k+1)}/\theta_{ij(k)} = \theta_{i(j+1)k}/\theta_{i(j)k} = \theta_{(i+1)jk}/\theta_{(i)jk}$. For unit-spaced scores, show that $\log \theta_{ijk} = \beta$. Goodman (1979a) called this the *uniform interaction model*.

b. Show that log odds ratios for any two variables change linearly across levels of the third variable.

c. Show that the likelihood equations are those for model (XY, XZ, YZ) plus

$$\sum_i \sum_j \sum_k u_i v_j w_k \hat{\mu}_{ijk} = \sum_i \sum_j \sum_k u_i v_j w_k n_{ijk}.$$

d. Explain why model (9.12) is a special case of model (9.22).

9.41 Construct a model having general XZ and YZ associations, but row effects for the XY association that are **(a)** homogeneous, and **(b)** heterogeneous across levels of Z. Interpret.

9.42 Explain why the *RC* model requires scale constraints for the scores. Show the residual df $= (I-2)(J-2)$. Find and interpret the likelihood equations. Explain why the fit is invariant to category orderings.

9.43 Refer to correlation model (9.16) (Goodman 1985, 1986).

a. Show that λ is the correlation between the scores.

b. If this model holds, show that $\sum_i \mu_i(\pi_{ij}/\pi_{+j}) = \lambda\nu_j$ and $\sum_j \nu_j(\pi_{ij}/\pi_{i+}) = \lambda\mu_i$. Interpret.

c. With λ close to zero, show that $\log(\pi_{ij})$ has form $\gamma_i + \delta_i + \lambda\mu_i\nu_j + o(\lambda)$, where $o(\lambda)/\lambda \to 0$ as $\lambda \to 0$. Thus, when the association is weak, the correlation model is similar to the linear-by-linear association model with $\beta = \lambda$ and scores $\{u_i = \mu_i\}$ and $\{v_j = \nu_j\}$.

9.44 For the general canonical correlation model, show that $\sum\lambda_k^2 = \sum_i\sum_j(\pi_{ij} - \pi_{i+}\pi_{+j})^2/\pi_{i+}\pi_{+j}$. Thus, the squared correlations partition a dependence measure that is the noncentrality (6.8) of X^2 for the independence model with $n = 1$. [Goodman (1986) stated other partitionings.]

9.45 Refer to model (9.18). Given the times at risk $\{t_{ij}\}$, show that sufficient statistics are $\{n_{i+}\}$ and $\{n_{+j}\}$.

9.46 Refer to Section 9.7.3. Let $T = \sum t_i$ and $W = \sum w_i$. Suppose that survival times have a negative exponential distribution with parameter λ.

a. Using log likelihood (9.19), show that $\hat{\lambda} = W/T$.

b. Conditional on T, show that W has a Poisson distribution with mean $T\lambda$. Using the Poisson likelihood, show that $\hat{\lambda} = W/T$.

9.47 Show that ML estimates do not exist for Table 9.15. [*Hint:* Haberman (1973b, 1974a, p. 398): If $\hat{\mu}_{111} = c > 0$, then marginal constraints the model satisfy imply that $\hat{\mu}_{222} = -c$.]

9.48 For a loglinear model, explain heuristically why the ML estimate of a parameter is infinite when its sufficient statistic takes its maximum or minimum possible value, for given values of other sufficient statistics.

Categorical Data Analysis, Second Edition. Alan Agresti

ISBN: 0-471-36093-7

CHAPTER 10

Models for Matched Pairs

We next introduce methods for comparing categorical responses for two samples when each observation in one sample pairs with an observation in the other. Such *matched-pairs* data commonly occur in studies with repeated measurement of subjects, such as *longitudinal studies* that observe subjects over time. Because of the matching, the responses in the two samples are statistically dependent. This is the first of four chapters on special methods for handling such dependence.

Table 10.1 illustrates matched-pairs data. For a poll of a random sample of 1600 voting-age British citizens, 944 indicated approval of the Prime Minister's performance in office. Six months later, of these same 1600 people, 880 indicated approval. The two cells with identical row and column response form the main diagonal of the table. These subjects had the same opinion at both surveys. They compose most of the sample, since relatively few people changed opinion. A strong association exists between opinions six months apart, the sample odds ratio being $(794 \times 570)/(150 \times 86) = 35.1$.

For matched pairs with a categorical response, a two-way contingency table with the same row and column categories summarizes the data. The table is *square*. In this chapter we present analyses of square tables. In Section 10.1 we describe methods for comparing proportions with a binary response. In Section 10.2 we discuss logistic regression analyses of such data. For multicategory responses, Section 10.3 covers nominal and ordinal logit

TABLE 10.1 Rating of Performance of Prime Minister

First Survey	Second Survey		
	Approve	Disapprove	Total
Approve	794	150	944
Disapprove	86	570	656
Total	880	720	1600

models for comparing the response distributions. In Section 10.4 we introduce loglinear models for square tables. In Sections 10.5 and 10.6 we discuss two matched-pairs applications for which models for square tables are useful: analyzing agreement between two observers who rate a common set of subjects, and evaluating preferences of treatments based on their pairwise evaluation.

Section 10.7 extends the models of Sections 10.2 through 10.4 to multiway tables that result from matched sets of observations. In Chapter 11 we extend them further to incorporate explanatory variables.

10.1 COMPARING DEPENDENT PROPORTIONS

For each of n matched pairs, let π_{ab} denote the probability of outcome a for the first observation and outcome b for the second. Let n_{ab} count the number of such pairs, with $p_{ab} = n_{ab}/n$ the sample proportion. We treat $\{n_{ab}\}$ as a sample from a multinomial $(n; \{\pi_{ab}\})$ distribution. Then p_{a+} is the proportion in category a for observation 1, and p_{+a} is the corresponding proportion for observation 2. We compare samples by comparing marginal proportions $\{p_{a+}\}$ with $\{p_{+a}\}$. With matched samples, these proportions are correlated, and methods for independent samples are inappropriate.

In this section we consider binary outcomes. When $\pi_{1+} = \pi_{+1}$, then $\pi_{2+} = \pi_{+2}$ also, and there is *marginal homogeneity*. Since

$$\pi_{1+} - \pi_{+1} = (\pi_{11} + \pi_{12}) - (\pi_{11} + \pi_{21}) = \pi_{12} - \pi_{21},$$

marginal homogeneity in 2×2 tables is equivalent to $\pi_{12} = \pi_{21}$. The table then shows *symmetry* across the main diagonal.

10.1.1 Inference for Dependent Proportions

One comparison of the marginal distributions uses $\delta = \pi_{+1} - \pi_{1+}$. Let

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

From formula (1.3) for multinomial covariances, $\text{cov}(p_{+1}, p_{1+}) = \text{cov}(p_{11} + p_{21}, p_{11} + p_{12})$ simplifies to $(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})/n$. Thus,

$$\text{var}(\sqrt{n}\, d) = \pi_{1+}(1 - \pi_{1+}) + \pi_{+1}(1 - \pi_{+1}) - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21}). \tag{10.1}$$

For large samples, d has approximately a normal sampling distribution. A confidence interval for $\delta = \pi_{+1} - \pi_{1+}$ is then

$$d \pm z_{\alpha/2}\, \hat{\sigma}(d),$$

where

$$\hat{\sigma}^2(d) = \left[p_{1+}(1 - p_{1+}) + p_{+1}(1 - p_{+1}) - 2(p_{11}p_{22} - p_{12}p_{21}) \right]/n$$
$$= \left[(p_{12} + p_{21}) - (p_{12} - p_{21})^2 \right]/n, \tag{10.2}$$

with the second formula following after substitution and some algebra. Inverting the score test of H_0: $\delta = \delta_0$ is more complex but provides coverage probabilities closer to the nominal values (Tango 1998), as does adding 1 to each cell before computing d and $\hat{\sigma}(d)$.

The hypothesis of marginal homogeneity is H_0: $\pi_{1+} = \pi_{+1}$ (i.e., $\delta = 0$). The ratio $z = d/\hat{\sigma}(d)$ or its square is a Wald test statistic. Under H_0, an alternative estimated variance is

$$\hat{\sigma}_0^2(d) = \frac{p_{12} + p_{21}}{n} = \frac{n_{12} + n_{21}}{n^2}. \tag{10.3}$$

The score test statistic $z_0 = d/\hat{\sigma}_0(d)$ simplifies to

$$z_0 = \frac{n_{21} - n_{12}}{(n_{21} + n_{12})^{1/2}}. \tag{10.4}$$

The square of z_0 is a chi-squared statistic with df = 1. The test using it is called *McNemar's test* (McNemar 1947).

The McNemar statistic depends only on cases classified in *different* categories for the two observations. The $n_{11} + n_{22}$ on the main diagonal are irrelevant to inference about whether π_{1+} and π_{+1} differ. This may seem surprising, but *all* cases contribute to inference about *how much* π_{1+} and π_{+1} differ: for instance, to estimating δ and the standard error.

10.1.2 Prime Minister Approval Rating Example

For Table 10.1, the sample proportions of approval of the prime minister's performance are $p_{1+} = 944/1600 = 0.59$ for the first survey and $p_{+1} = 880/1600 = 0.55$ for the second. Using (10.2), a 95% confidence interval for $\pi_{+1} - \pi_{1+}$ is $(0.55 - 0.59) \pm 1.96(0.0095)$, or $(-0.06, -0.02)$. The approval rating appears to have dropped between 2 and 6%.

For testing marginal homogeneity, the test statistic (10.4) using the null variance is

$$z_0 = \frac{86 - 150}{(86 + 150)^{1/2}} = -4.17.$$

It shows strong evidence of a drop in the approval rating.

10.1.3 Increased Precision with Dependent Samples

The final term of formula (10.1), based on $\text{cov}(p_{+1}, p_{1+})$, reflects the dependence between the marginal proportions. By contrast, for *independent* samples of size n each to estimate binomial probabilities π_1 and π_2, the covariance for the sample proportions is zero, and

$$\text{var}\left[\sqrt{n}\,(\text{difference of sample proportions})\right] = \pi_1(1-\pi_1) + \pi_2(1-\pi_2).$$

Dependent samples usually exhibit a positive dependence, with $\log\theta = \log[\pi_{11}\pi_{22}/\pi_{12}\pi_{21}] > 0$; that is, $\pi_{11}\pi_{22} > \pi_{12}\pi_{21}$. From (10.1), positive dependence implies that $\text{var}(d)$ is smaller than when the samples are independent.

A study design using dependent samples can help improve the precision of statistical inferences for within-subject effects. (By contrast, standard errors tend to be larger, per given number of observations, for between-subject group comparisons.) The improvement is substantial when samples are highly correlated. To illustrate, Table 10.1 with dependent samples of size 1600 each has a standard error of 0.0095 for $d = 0.55 - 0.59$. The two observations have strong association, the sample odds ratio being 35.1. *Independent* samples of size 1600 each with $\hat{\pi}_1 - \hat{\pi}_2 = 0.55 - 0.59$ have a standard error of 0.0175 for the difference, nearly twice as large.

10.1.4 Small-Sample Test Comparing Matched Proportions

The null hypothesis of marginal homogeneity for binary matched pairs is, equivalently, H_0: $\pi_{12} = \pi_{21}$ or $\pi_{21}/(\pi_{21} + \pi_{12}) = 0.5$. For small samples, an exact test conditions on $n^* = n_{21} + n_{12}$ (Mosteller 1952). Under H_0, n_{21} has a binomial $(n^*, \frac{1}{2})$ distribution, for which $E(n_{21}) = \frac{1}{2}n^*$. The P-value for the test is a binomial tail probability.

For instance, for Table 10.1, consider H_a: $\pi_{+1} < \pi_{1+}$, or equivalently, H_a: $\pi_{21} < \pi_{12}$. Since $n^* = 86 + 150 = 236$, the reference distribution is bin(236, $\frac{1}{2}$). The P-value is the probability of at least 150 successes out of 236 trials, which equals 0.00002. The P-value for H_a: $\pi_{+1} \neq \pi_{1+}$ doubles this.

When $n^* > 10$, the reference binomial distribution is approximately normal with mean $\frac{1}{2}n^*$ and variance $n^*(\frac{1}{2})(\frac{1}{2})$. The standardized normal test statistic equals

$$z = \frac{n_{21} - \frac{1}{2}n^*}{\left[n^*\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\right]^{1/2}} = \frac{n_{21} - n_{12}}{(n_{21} + n_{12})^{1/2}}.$$

This is identical to the McNemar statistic (10.4).

10.1.5 Connection between McNemar and Cochran–Mantel–Haenszel Tests

An alternative representation of binary responses for n matched pairs presents the data in n partial tables, one 2×2 table for each pair. It has columns that are the two possible outcomes for each measurement. Row 1 shows the outcome of the first observation, and row 2 shows the outcome of the second.

Table 10.2 shows the four possible partial tables in this representation. For Table 10.1, the full three-way table has 1600 partial tables; 794 look like the one for subject 1 (i.e., "approve" at both surveys), 570 who disapproved at each survey have tables like the one for subject 2, 86 have tables like the one for subject 3, and 150 have tables like the one for subject 4. The 1600 subjects from Table 10.1 provide 3200 observations in a $2 \times 2 \times 1600$ contingency table. Collapsing this table over the 1600 partial tables yields a 2×2 table with first row equal to $(944, 656)$ and second row equal to $(880, 720)$. These are the total number of (approve, disapprove) responses for the two surveys. They form the marginal counts in Table 10.1.

For each subject, suppose that the probability of approval is identical in each survey. Then, conditional independence exists between the opinion outcome and the survey time, controlling for subject. The probability of approval is then also the same for each survey in the marginal table collapsed over the subjects. But this implies that the true probabilities for Table 10.1 satisfy marginal homogeneity. Thus, a test of conditional independence in the $2 \times 2 \times 1600$ table provides a test of marginal homogeneity for Table 10.1.

To test conditional independence in this three-way table, one can use the Cochran–Mantel–Haenszel (CMH) statistic (6.6). The result of that chi-squared statistic is algebraically identical to the squared McNemar's statistic, namely $(n_{21} - n_{12})^2/(n_{12} + n_{21})$ for tables of form (10.1). McNemar's test is a special case of the CMH test applied to the binary responses of n matched pairs displayed in n partial tables. This connection is not helpful for computational purposes, since the McNemar statistic is simple. But it does suggest

TABLE 10.2 Representation of Four Types of Matched Pairs Contributing to Counts in Table 10.1

		Response	
Subject	Survey	Approve	Disapprove
1	First	1	0
	Second	1	0
2	First	0	1
	Second	0	1
3	First	0	1
	Second	1	0
4	First	1	0
	Second	0	1

ways of handling more complex matched data. With several outcome categories or several observations, one can test marginal homogeneity by applying the generalized CMH tests (Section 7.5) using a single stratum for each subject, with each row representing a particular observation (Darroch 1981; Mantel and Byar 1978).

Coming sections refer to the $2 \times 2 \times n$ table representation of matched-pairs data as the *subject-specific* table. They refer to the 2×2 table of form of Table 10.1 as the *population-averaged* table, since its margins provide direct estimates of population marginal proportions.

10.2 CONDITIONAL LOGISTIC REGRESSION FOR BINARY MATCHED PAIRS

In Section 6.7 we introduced *conditional logistic regression* for eliminating nuisance parameters from an analysis. We now study this for binary matched-pairs data. The models refer to subject-specific tables.

10.2.1 Marginal versus Conditional Models for Matched Pairs

The analyses of Section 10.1 occur in the context of models. Let (Y_1, Y_2) denote the pair of observations for a randomly selected subject, where a "1" outcome denotes category 1 (success) and "0" denotes category 2. The difference $\delta = P(Y_2 = 1) - P(Y_1 = 1)$ between marginal probabilities occurs as a parameter in

$$P(Y_t = 1) = \alpha + \delta x_t, \tag{10.5}$$

where $x_1 = 0$ and $x_2 = 1$; then, $P(Y_1 = 1) = \alpha$ and $P(Y_2 = 1) = \alpha + \delta$. Alternatively, the logit link yields

$$\text{logit}\big[P(Y_t = 1)\big] = \alpha + \beta x_t. \tag{10.6}$$

The parameter β is a log odds ratio with the marginal distributions.

Models (10.5) and (10.6) are *marginal models*: They focus on the marginal distributions of responses for the two observations. For instance, in terms of the population-averaged table, the ML estimate of β in (10.6) is the log odds ratio of marginal proportions, $\hat{\beta} = \log[\, p_{+1} p_{2+} / p_{+2} p_{1+}\,]$. See Problem 10.26 for its asymptotic variance.

By contrast, the subject-specific table having strata like Table 10.2 implicitly allows probabilities to vary by subject. Let (Y_{i1}, Y_{i2}) denote the ith pair of observations, $i = 1, \ldots, n$. A model then has the form

$$\text{link}\big[P(Y_{it} = 1)\big] = \alpha_i + \beta x_t. \tag{10.7}$$

This is called a *conditional model*, since the effect β is defined conditional on the subject. Its estimate describes conditional association for the three-way table stratified by subject. The effect is *subject-specific*, since it is defined at

the subject level. By contrast, the effects in marginal models (10.5) and (10.6) are *population-averaged*, since they refer to averaging over the entire population rather than to individual subjects.

For the identity link, subject-specific and population-averaged effects are identical. For instance, for the conditional model (10.7) with identity link, $\beta = P(Y_{i2} = 1) - P(Y_{i1} = 1)$ for all i, and averaging this over subjects in the population equates β to the δ parameter in model (10.5). For nonlinear links, however, the effects differ. For model (10.7) with the logit link, for instance,

$$P(Y_{it} = 1) = \exp(\alpha_i + \beta x_t)/\big[1 + \exp(\alpha_i + \beta x_t)\big].$$

The average of this for the population does not have the form $\exp(\alpha + \beta x_t)/[1 + \exp(\alpha + \beta x_t)]$ corresponding to the marginal logit model (10.6). We now take a closer look at the conditional model with logit link.

10.2.2 A Logit Model with Subject-Specific Probabilities

Model (10.7) differs from models in earlier chapters by permitting subjects to have their own probability distributions. Cox (1958b, 1970) and Rasch (1961) presented this model with logit link. This model for Y_{it}, observation t for subject i, is

$$\text{logit}\big[P(Y_{it} = 1)\big] = \alpha_i + \beta x_t, \tag{10.8}$$

where $x_1 = 0$ and $x_2 = 1$. Although permitting subject-specific distributions, it assumes a common effect β. For subject i,

$$P(Y_{i1} = 1) = \frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)}, \qquad P(Y_{i2} = 1) = \frac{\exp(\alpha_i + \beta)}{1 + \exp(\alpha_i + \beta)}.$$

The parameter β compares the response distributions. For each subject, the odds of success for observation 2 are $\exp(\beta)$ times the odds for observation 1.

Given the parameters, with model (10.8) one normally assumes independence of responses for different subjects and for the two observations on the same subject. However, averaged over all subjects, the responses are nonnegatively associated. Suppose that $|\beta|$ is small compared to $|\alpha_i|$. A subject with a large positive α_i has high $P(Y_{it} = 1)$ for each t and is likely to have a success each time; a subject with a large negative α_i has low $P(Y_{it} = 1)$ for each t and is likely to have a failure each time. The greater the variability in $\{\alpha_i\}$, the greater the overall positive association between responses, successes (failures) for observation 1 tending to occur with successes (failures) for observation 2. This is true for any β. The positive association reflects the shared value of α_i for each observation in a pair. No association occurs only when $\{\alpha_i\}$ are identical. Thus, the model does account for the dependence in matched pairs. Fitting it takes into account nonnegative association through the structure of the model.

For this model, the large number of $\{\alpha_i\}$ causes difficulties with the fitting process and with the properties of ordinary ML estimators (Problem 10.24). The remedy of conditional ML treats them as nuisance parameters and maximizes the likelihood function for a conditional distribution that eliminates them. A note on terminology: We've referred to model (10.8) as a *conditional* model, meaning that its effect β is subject-specific, conditional on the subject. The analyses described below for such models are examples of *conditional* logistic regression; but here the term *conditional* refers to the ML analysis that is performed conditional on sufficient statistics for nuisance parameters, to eliminate those parameters from the likelihood.

10.2.3 Conditional ML Inference for Binary Matched Pairs

For model (10.8), assuming independence of responses for different subjects and for the two observations on the same subject, the joint mass function for $\{(y_{11}, y_{12}), \ldots, (y_{n1}, y_{n2})\}$ is

$$\prod_{i=1}^{n} \left(\frac{\exp(\alpha_i)}{1+\exp(\alpha_i)}\right)^{y_{i1}} \left(\frac{1}{1+\exp(\alpha_i)}\right)^{1-y_{i1}} \times \left(\frac{\exp(\alpha_i+\beta)}{1+\exp(\alpha_i+\beta)}\right)^{y_{i2}} \left(\frac{1}{1+\exp(\alpha_i+\beta)}\right)^{1-y_{i2}}.$$

In terms of the data, this is proportional to

$$\exp\left[\sum_i \alpha_i(y_{i1}+y_{i2}) + \beta\left(\sum_i y_{i2}\right)\right].$$

To eliminate $\{\alpha_i\}$, we condition on their sufficient statistics, the pairwise success totals $\{S_i = y_{i1} + y_{i2}\}$. Given $S_i = 0$, $P(Y_{i1} = Y_{i2} = 0) = 1$, and given $S_i = 2$, $P(Y_{i1} = Y_{i2} = 1) = 1$. The distribution of (Y_{i1}, Y_{i2}) depends on β only when $S_i = 1$; that is, only when outcomes differ for the two responses. Given $y_{i1} + y_{i2} = 1$, the conditional distribution is

$$\begin{aligned}
&P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2} \,|\, S_i = 1) \\
&\quad = P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}) \big/ \big[P(Y_{i1} = 1, Y_{i2} = 0) + P(Y_{i1} = 0, Y_{i2} = 1)\big] \\
&\quad = \frac{\left(\frac{\exp(\alpha_i)}{1+\exp(\alpha_i)}\right)^{y_{i1}} \left(\frac{1}{1+\exp(\alpha_i)}\right)^{1-y_{i1}} \left(\frac{\exp(\alpha_i+\beta)}{1+\exp(\alpha_i+\beta)}\right)^{y_{i2}} \left(\frac{1}{1+\exp(\alpha_i+\beta)}\right)^{1-y_{i2}}}{\frac{\exp(\alpha_i)}{1+\exp(\alpha_i)}\frac{1}{1+\exp(\alpha_i+\beta)} + \frac{1}{1+\exp(\alpha_i)}\frac{\exp(\alpha_i+\beta)}{1+\exp(\alpha_i+\beta)}} \\
&\quad = \exp(\beta)/[1+\exp(\beta)], \quad y_{i1} = 0, \quad y_{i2} = 1 \\
&\quad = 1/[1+\exp(\beta)], \quad y_{i1} = 1, \quad y_{i2} = 0.
\end{aligned}$$

Again, let $\{n_{ab}\}$ denote the counts for the four possible sequences. For subjects having $S_i = 1$, $\Sigma_i y_{i1} = n_{12}$, the number of subjects having success for observation 1 and failure for observation 2. Similarily, for those subjects, $\Sigma_i y_{i2} = n_{21}$ and $\Sigma_i S_i = n^* = n_{12} + n_{21}$. Since n_{21} is the sum of n^* independent, identical Bernoulli variates, its conditional distribution is binomial with parameter $\exp(\beta)/[1 + \exp(\beta)]$. For testing marginal homogeneity ($\beta = 0$), the parameter equals $\frac{1}{2}$. In summary, the conditional analysis for the logit model implies that pairs in which $y_{i1} = y_{i2}$ are irrelevant to inference about β. When this model is realistic, it provides justification for comparing marginal distributions using only the $n_{12} + n_{21}$ pairings having outcomes in different categories at the two observations.

Conditional on $S_i = 1$, the joint distribution of the matched pairs is

$$\prod_{S_i=1} \left(\frac{1}{1 + \exp(\beta)} \right)^{y_{i1}} \left(\frac{\exp(\beta)}{1 + \exp(\beta)} \right)^{y_{i2}} = \frac{[\exp(\beta)]^{n_{21}}}{[1 + \exp(\beta)]^{n^*}} \quad (10.9)$$

where the product refers to all pairs having $S_i = 1$. Differentiating the log of this conditional likelihood and equating to 0 and solving yields the conditional ML estimator of β in model (10.8). You can check that it and its standard error are

$$\hat{\beta} = \log(n_{21}/n_{12}), \qquad \text{SE} = \sqrt{1/n_{21} + 1/n_{12}}\,. \quad (10.10)$$

10.2.4 Random Effects in Binary Matched-Pairs Model

An alternative remedy to handling the huge number of nuisance parameters in logit model (10.8) treats $\{\alpha_i\}$ as *random effects*. This regards $\{\alpha_i\}$ as an unobserved random sample from a probability distribution, usually assumed to be $N(\mu, \sigma^2)$ with unknown μ and σ. It eliminates $\{\alpha_i\}$ by averaging with respect to their distribution, yielding a marginal distribution. The likelihood function then depends on β as well as the $N(\mu, \sigma^2)$ parameters. It has only three parameters and is more manageable. For matched pairs with nonnegative sample log odds ratio, this approach also yields $\hat{\beta} = \log(n_{21}/n_{12})$ (Neuhaus et al. 1994). This model is an example of a *generalized linear mixed model*, containing both random effects and the fixed effect β. Its analysis is presented in Chapter 12.

Model (10.8) implies that the true odds ratio for each of the n subject-specific partial tables equals $\exp(\beta)$. In Section 6.3.5 we presented the Mantel–Haenszel estimate of a common odds ratio for several 2×2 tables. In fact, that estimator applied to subject-specific tables of the form shown in Table 10.2 is algebraically identical to n_{21}/n_{12} for tables of the form shown in Table 10.1. (Recall that partial tables with responses in only one column do not contribute to the CMH test or Mantel–Haenszel estimate.) In summary, the Mantel–Haenszel estimate, the conditional ML estimate, and

(with nonnegative log odds ratio) the ML estimate for the random effects version of logit model (10.8) yield $\exp(\hat{\beta}) = n_{21}/n_{12}$.

10.2.5 Logistic Regression for Matched Case–Control Studies

The two observations (y_{i1}, y_{i2}) in a matched pair need not refer to the same subject. For instance, case–control studies that match a single control with each case yield matched-pairs data. For a binary response Y, each case $(Y = 1)$ is matched with a control $(Y = 0)$ according to criteria that could affect the response. Subjects in the matched pairs are measured on the predictor variable(s) of interest, X, and the XY association is analyzed.

Table 10.3 illustrates. A case–control study of acute myocardial infarction (MI) among Navajo Indians matched 144 victims of MI according to age and gender with 144 people free of heart disease. Subjects were asked whether they had ever been diagnosed as having diabetes ($x = 0$, no; $x = 1$, yes). Table 10.3 has the same form as Table 10.1 except that the levels of X rather than the levels of Y form the rows and the columns.

One can display the data for each matched case–control pair using a partial table of the form shown in Table 10.2, but reversing the roles of X and Y. The X values have four possible patterns, shown in Table 10.4. There are 37 partial tables of type a, since for 37 pairs the case had diabetes and the control did not, 16 partial tables of type b, 9 of type c, and 82 of type d.

Now, for subject t in matched pair i, consider the model

$$\text{logit}\big[P(Y_{it} = 1)\big] = \alpha_i + \beta x_{it}. \tag{10.11}$$

TABLE 10.3 Previous Diagnoses of Diabetes for Myocardial Infarction (MI) Case–Control Pairs

	MI Cases		
MI Controls	Diabetes	No Diabetes	Total
Diabetes	9	16	25
No diabetes	37	82	119
Total	46	98	144

Source: J. L. Coulehan et al., *Amer. J. Public Health* **76**: 412–414 (1986), reprinted with permission from the American Public Health Association.

TABLE 10.4 Possible Case–Control Pairs for Table 10.3

	a		b		c		d	
Diabetes	Case	Control	Case	Control	Case	Control	Case	Control
Yes	1	0	0	1	1	1	0	0
No	0	1	1	0	0	0	1	1

The probabilities modeled refer to the distribution of Y given X, but the retrospective study provides information only about the distribution of X given Y. One can estimate the odds ratio $\exp(\beta)$, however, since it refers to the XY odds ratio, which relates to both conditional distributions (Sections 2.2.4, 5.1.4). Even though this study reverses the roles of X and Y in terms of which is fixed and which is random, the conditional ML estimate of $\exp(\beta)$ is simply $n_{21}/n_{12} = 37/16 = 2.3$.

10.2.6 Conditional ML for Matched Pairs with Multiple Predictors

When the binary response has p predictors for case–control or subject-specific matched pairs, the model generalizes to

$$\text{logit}\big[P(Y_{it} = 1)\big] = \alpha_i + \beta_1 x_{1it} + \beta_2 x_{2it} + \cdots + \beta_p x_{pit}, \quad (10.12)$$

where x_{hit} denotes the value of predictor h for observation t in pair i, $t = 1, 2$. Typically, one predictor is an explanatory variable of interest, such as diabetes status. The others are covariates being controlled, in addition to those already controlled by virtue of using them to form the matched pairs. The conditional ML approach to estimating $\{\beta_j\}$ conditions on sufficient statistics for α_i to eliminate them from the likelihood.

Let $\mathbf{x}_{it} = (x_{1it}, \ldots, x_{pit})'$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. A generalization of the derivation in Section 10.2.3 shows that

$$P(Y_{i1} = 0, Y_{i2} = 1 \,|\, S_i = 1) = \exp(\mathbf{x}_{i2}'\boldsymbol{\beta}) \big/ \big[\exp(\mathbf{x}_{i1}'\boldsymbol{\beta}) + \exp(\mathbf{x}_{i2}'\boldsymbol{\beta})\big],$$

$$P(Y_{i1} = 1, Y_{i2} = 0 \,|\, S_i = 1) = \exp(\mathbf{x}_{i1}'\boldsymbol{\beta}) \big/ \big[\exp(\mathbf{x}_{i1}'\boldsymbol{\beta}) + \exp(\mathbf{x}_{i2}'\boldsymbol{\beta})\big]. \quad (10.13)$$

Dividing numerator and denominator by $\exp(\mathbf{x}_{i1}'\boldsymbol{\beta})$ shows that the first equation has the form of logistic regression with no intercept and with predictor values $\mathbf{x}_i^* = \mathbf{x}_{i2} - \mathbf{x}_{i1}$. In fact, one can obtain conditional ML estimates for model (10.12) by fitting a logistic regression model to those pairs alone, using artificial response $y^* = 1$ when $(y_{i1} = 0,\ y_{i2} = 1)$, $y^* = 0$ when $(y_{i1} = 1,\ y_{i2} = 0)$, no intercept, and predictor values $\mathbf{x}_i^*$. This addresses the same likelihood as the conditional likelihood (Breslow et al. 1978; Chamberlain 1980).

To illustrate, for model (10.11) with Table 10.3, let $y_i^* = y_{i2} - y_{i1}$ and $x_i^* = x_{i2} - x_{i1}$. If $t = 1$ refers to the control and $t = 2$ to the case, then $y_i^* = 1$ always. Since $x_{it} = 1$ represents "yes" for diabetes and $x_{it} = 0$ represents "no," $(y_i^* = 1, x_i^* = -1)$ for 16 observations, $(y_i^* = 1, x_i^* = 0)$ for $9 + 82 = 91$ observations, and $y_i^* = 1, x_i^* = +1)$ for 37 observations. The logit model that forces $\hat{\alpha} = 0$ has $\hat{\beta} = 0.84$. With a single binary predictor, the estimate is identical to $\log(n_{21}/n_{12})$.

10.2.7 Marginal Models and Conditional Models: Extensions

For binary matched-pairs data, Section 10.1 presented analyses for a marginal (i.e., population-averaged) model, and this section presented analyses for a conditional (i.e., subject-specific) model. These models generalize to multinomial responses and to matched sets. For instance, Chamberlain (1980) discussed conditional ML for matched pairs on a multinomial response. For binary responses, model (10.12) applies when α_i refers to a set of repeated measurements on subject i. Or, it could refer to a matched set that is a cluster of subjects, such as children from family i or fetuses from litter i.

With extensions of the conditional model to matched-set clusters, the conditional ML approach is restricted to estimating β_j that are within-cluster effects, such as occur in case–control and crossover studies. For these, the explanatory variable varies in t for each i. Conditional ML cannot estimate a between-cluster effect. Statistics providing information about such an effect use subject totals at different levels of the relevant explanatory variable; however, those totals sum the sufficient statistics for $\{\alpha_i\}$, so they are themselves fixed and have degenerate distributions after conditioning on the sufficient statistics. An explanatory variable that is constant in t for each i cancels out of the conditional likelihood. [You can observe this for matched pairs with (10.13) for any j for which $x_{ji1} = x_{ji2}$ all i.] For it, at best one can stratify by its levels and fit a model estimating within-cluster effects separately at each level. An advantage of using the random effects approach instead of conditional ML with the conditional model is that it is not restricted to estimating within-cluster effects.

In the remainder of this chapter we emphasize marginal models for matched pairs with multinomial responses. In the following chapter we deal with marginal model extensions allowing matched sets and explanatory variables. Conditional models using a random effects approach have extra computational complexities. We mention briefly some multinomial conditional models in this chapter, but we defer most discussion to Chapter 12.

10.3 MARGINAL MODELS FOR SQUARE CONTINGENCY TABLES

Matched pairs analyses generalize from binary to $I > 2$ outcome categories. A square $I \times I$ table $\{n_{ab}\}$ shows counts of possible sequences (a, b) of outcomes for (Y_1, Y_2). Let $\pi_{ab} = P(Y_1 = a, Y_2 = b)$. Marginal homogeneity is $P(Y_1 = a) = P(Y_2 = a)$ for $a = 1, \ldots, I$. Marginal models compare $\{P(Y_1 = a)\}$ and $\{P(Y_2 = a)\}$.

10.3.1 Marginal Models for Ordinal Classifications

For ordered categories, marginal model (10.6) for binary matched pairs extends using ordinal logits. With cumulative logits,

$$\text{logit}\big[P(Y_t \le j)\big] = \alpha_j + \beta x_t, \quad t = 1, 2, \quad j = 1, \ldots, I - 1, \quad (10.14)$$

where $x_1 = 0$ and $x_2 = 1$. This model has proportional odds structure (Section 7.2.2). The odds of outcome $Y_2 \leq j$ equal $\exp(\beta)$ times the odds of outcome $Y_1 \leq j$. The model implies stochastically ordered marginal distributions, with $\beta > 0$ meaning that Y_1 tends to be higher than Y_2. Marginal homogeneity corresponds to $\beta = 0$.

Model fitting treats (Y_1, Y_2) as dependent. The ML approach maximizes the multinomial likelihood for $\{\pi_{ab}\}$. This is not simple. Since the model refers to marginal probabilities $\{P(Y_1 = a) = \pi_{a+}\}$ and $\{P(Y_2 = b) = \pi_{+b}\}$, one cannot substitute the model formula in the kernel $\sum_a \sum_b n_{ab} \log \pi_{ab}$ of the log likelihood, which has joint probabilities. We defer discussion of ML model fitting of marginal models to Section 11.2.5. Model (10.14) describes the $2(I - 1)$ marginal probabilities by I parameters, so df $= I - 2$ for testing fit. Alternatively, one can compare margins using summaries such as a difference in means for chosen category scores (Problem 10.38).

10.3.2 Premarital and Extramarital Sex Example

Refer to Table 10.5. For a General Social Survey, subjects gave their opinion about premarital sex (a couple having sex before marriage) and extramarital sex (a married person having sex with someone other than the marriage partner). The response categories are 1 = always wrong, 2 = almost always wrong, 3 = wrong only sometimes, 4 = not wrong at all.

The sample cumulative marginal proportions are $(0.307, 0.389, 0.611)$ for premarital sex and $(0.815, 0.918, 0.987)$ for extramarital sex. This suggests that responses on premarital sex tended to be higher on the ordinal scale than those on extramarital sex. With scores $(1, 2, 3, 4)$, the mean for premarital sex is 2.69, closest to the "wrong only sometimes" score, and the mean response for extramarital sex is 1.28, closest to the "always wrong" score.

The cumulative logit model (10.14) has $\hat{\beta} = 2.51$ (SE = 0.13). There is strong evidence that population responses are more positive on premarital than on extramarital sex. The fit of the marginal homogeneity model has $G^2 = 348.1$ (df = 3), and the fit of model (10.14) has $G^2 = 35.1$ (df = 2). The ordinal model does not fit well, but it fits much better than the marginal homogeneity model. Models to be considered in Section 10.4.7 fit better yet.

TABLE 10.5 Opinions on Premarital Sex and Extramarital Sex

Premarital Sex	Extramarital Sex				Total
	1	2	3	4	
1	144	2	0	0	146
2	33	4	2	0	39
3	84	14	6	1	105
4	126	29	25	5	185
Total	387	49	33	6	475

Source: 1989 General Social Survey, National Opinion Research Center.

10.3.3 Marginal Models for Nominal Classifications

With nominal responses, it is not sensible to assume the same effect for each logit. A baseline-category logit model has form

$$\log\big[P(Y_t = j)/P(Y_t = I)\big] = \alpha_j + \beta_j x_t, \qquad t = 1, 2, \quad j = 1, \ldots, I-1, \tag{10.15}$$

where $x_1 = 0$ and $x_2 = 1$. This model has $2(I-1)$ parameters for the $2(I-1)$ marginal probabilities. It is saturated.

Marginal homogeneity is the special case $\beta_1 = \cdots = \beta_{I-1} = 0$. To fit it, Lipsitz et al. (1990) and Madanksy (1963) maximized the multinomial likelihood for $\{n_{ab}\}$ subject to these constraints. Iterative methods produce fitted values $\{\hat{\mu}_{ab}\}$. Comparing these to $\{n_{ab}\}$ using G^2 or X^2 tests marginal homogeneity, with df $= I-1$.

Bhapkar (1966) tested marginal homogeneity by exploiting the asymptotic normality of marginal proportions. Let $d_a = p_{+a} - p_{a+}$, and let $\mathbf{d}' = (d_1, \ldots, d_{I-1})$. It is redundant to include d_I, since $\sum d_a = 0$. The sample covariance matrix $\hat{\mathbf{V}}$ of $\sqrt{n}\,\mathbf{d}$ has elements

$$\hat{v}_{ab} = -(p_{ab} + p_{ba}) - (p_{+a} - p_{a+})(p_{+b} - p_{b+}) \qquad \text{for } a \neq b,$$

$$\hat{v}_{aa} = p_{+a} + p_{a+} - 2p_{aa} - (p_{+a} - p_{a+})^2.$$

Now $\sqrt{n}\,[\mathbf{d} - E(\mathbf{d})]$ has an asymptotic multivariate normal distribution with estimated covariance matrix $\hat{\mathbf{V}}$. Under marginal homogeneity, $E(\mathbf{d}) = \mathbf{0}$, and

$$W = n\mathbf{d}'\hat{\mathbf{V}}^{-1}\mathbf{d} \tag{10.16}$$

is asymptotically chi-squared with df $= I-1$. This is a Wald test for parameters in the analog of model (10.15) using the identity link. Stuart (1955) proposed $W_0 = n\mathbf{d}'\hat{\mathbf{V}}_0^{-1}\mathbf{d}$, which uses the sample *null* covariance matrix $\hat{\mathbf{V}}_0$ and is the score test. This has

$$\hat{v}_{ab0} = -(p_{ab} + p_{ba}) \qquad \text{for } a \neq b,$$

$$\hat{v}_{aa0} = p_{+a} + p_{a+} - 2p_{aa}.$$

Ireland et al. (1969) noted that $W = W_0/(1 - W_0/n)$. For $I = 2$, W_0 is McNemar's statistic, the square of (10.4).

These tests use all $I-1$ degrees of freedom available for comparisons of I pairs of marginal proportions. With ordered categories, when I is large and the dependence between classifications is strong, ordinal tests (with df = 1) can be much more powerful (Agresti 1984, p. 209).

TABLE 10.6 Migration from 1980 to 1985, with Fit of Marginal Homogeneity Model

Residence in 1980	Residence in 1985				Total
	Northeast	Midwest	South	West	
Northeast	11,607 (11,607)	100 (98.1)	366 (265.7)	124 (94.0)	12,197 (12,064.7)
Midwest	87 (88.7)	13,677 (13,677)	515 (379.1)	302 (323.3)	14,581 (14,377.1)
South	172 (276.5)	255 (350.8)	17,819 (17,819)	270 (287.3)	18,486 (18,733.5)
West	63 (92.5)	176 (251.3)	286 (269.8)	10,192 (10,192)	10,717 (10,805.6)
Total	11,929 (12,064.7)	14,178 (14,377.1)	18,986 (18,733.5)	10,888 (10,805.6)	55,981

Source: Data based on Table 12 of U.S. Bureau of the Census, Current Population Reports, Series P-20, No. 420, *Geographical Mobility: 1985* (Washington, DC: U.S. Government Printing Office), 1987.

10.3.4 Migration Example

For a sample of U.S. residents, Table 10.6 compares region of residence in 1985 with 1980. Relatively few people changed region, 95% of the observations falling on the main diagonal. The ML fit of marginal homogeneity, shown in Table 10.6, gives $G^2 = 240.8$ (df = 3). Statistics using differences in sample marginal proportions give similar results. For instance, Bhapkar's statistic (10.16) is $W = 236.5$ (df = 3).

The sample marginal proportions for the four regions were (0.218, 0.260, 0.330, 0.191) in 1980 and (0.213, 0.253, 0.339, 0.194) in 1985. Little change occurred over such a short time period. The large test statistics reflect the huge sample size. To estimate the change for a given region, we apply (10.2) to the collapsed 2×2 table that combines the other regions. A 95% confidence interval for $\pi_{+1} - \pi_{1+}$ is $(0.2131 - 0.2179) \pm 1.96(0.00054)$, or -0.005 ± 0.001. Similarly, a 95% confidence interval for $\pi_{+2} - \pi_{2+}$ is -0.007 ± 0.001, for $\pi_{+3} - \pi_{3+}$ is 0.009 ± 0.001, and for $\pi_{+4} - \pi_{4+}$ is 0.003 ± 0.001. Although strong evidence of change occurs for all four regions, the changes were small.

10.4 SYMMETRY, QUASI-SYMMETRY, AND QUASI-INDEPENDENCE

An alternative analysis of square contingency tables directly models the joint distribution using logit or loglinear models. Some models have marginal homogeneity as a special case.

An $I \times I$ joint distribution $\{\pi_{ab}\}$ satisfies *symmetry* if

$$\pi_{ab} = \pi_{ba} \qquad \text{whenever } a \neq b. \tag{10.17}$$

Under symmetry, $\pi_{a+} = \sum_b \pi_{ab} = \sum_b \pi_{ba} = \pi_{+a}$ for all a, so marginal homogeneity occurs. For $I = 2$, symmetry is equivalent to marginal homogeneity, but for $I > 2$, marginal homogeneity can occur without symmetry.

10.4.1 Symmetry as Logit and Loglinear Models

When all $\pi_{ab} > 0$, symmetry is a logit and a loglinear model. In logit form, it is trivially

$$\log(\pi_{ab}/\pi_{ba}) = 0 \qquad \text{for all } a < b.$$

For expected frequencies $\{\mu_{ab} = n\pi_{ab}\}$, it has the loglinear form

$$\log \mu_{ab} = \lambda + \lambda_a + \lambda_b + \lambda_{ab} \tag{10.18}$$

where all $\lambda_{ab} = \lambda_{ba}$. Both classifications have the same single-factor parameters $\{\lambda_a\}$, so $\log \mu_{ab} = \log \mu_{ba}$. Identifiability requires constraints. A simpler expression is $\log \mu_{ab} = \lambda_{ab}$, with all $\lambda_{ab} = \lambda_{ba}$.

For Poisson or multinomial cell counts $\{n_{ab}\}$, the likelihood equations are

$$\hat{\mu}_{ab} + \hat{\mu}_{ba} = n_{ab} + n_{ba} \quad \text{for all } a < b \quad \text{and} \quad \hat{\mu}_{aa} = n_{aa} \quad \text{for all } a.$$

The main diagonal has perfect fit. The solution that satisfies symmetry is

$$\hat{\mu}_{ab} = \frac{n_{ab} + n_{ba}}{2} \quad \text{for all } a, b.$$

The logit symmetry model has no parameters for the $\binom{I}{2}$ binomial pairs $\{(n_{ab}, n_{ba})\}$ with $a < b$, so its residual df $= I(I-1)/2$. Equivalently, the loglinear symmetry model $\log \mu_{ab} = \lambda_{ab}$ $(\lambda_{ab} = \lambda_{ba})$ for I^2 Poisson counts $\{n_{ab}\}$ has $\binom{I}{2}$ $\{\lambda_{ab}\}$ with $a < b$ and I $\{\lambda_{aa}\}$, so df $= I^2 - [I + I(I-1)/2] = I(I-1)/2$. For testing symmetry, Bowker (1948) showed that X^2 simplifies to

$$X^2 = \sum\sum_{a<b} \frac{(n_{ab} - n_{ba})^2}{n_{ab} + n_{ba}}.$$

For $I = 2$ this is McNemar's statistic, the square of (10.4). The standardized Pearson residuals equal

$$r_{ab} = (n_{ab} - n_{ba}) \Big/ (n_{ab} + n_{ba})^{1/2}.$$

Only one residual for each pair of categories is nonredundant, since $r_{ab} = -r_{ba}$. They satisfy $\sum_{a<b} r_{ab}^2 = X^2$.

The symmetry model is very simple. Except for a few specialized applications, such as describing intraobserver agreement for pairs of measurements by an observer, it rarely fits well. When the marginal distributions differ substantially, it fits poorly.

10.4.2 Quasi-symmetry

One can accommodate marginal heterogeneity by permitting the main-effect terms in the symmetry model (10.18) to differ. The resulting loglinear model, called *quasi-symmetry*, is

$$\log \mu_{ab} = \lambda + \lambda_a^X + \lambda_b^Y + \lambda_{ab}, \tag{10.19}$$

where $\lambda_{ab} = \lambda_{ba}$ for all $a < b$ (Caussinus 1966). Symmetry is the special case $\lambda_a^X = \lambda_a^Y$ for $a = 1, \ldots, I$, and independence is the special case in which all $\lambda_{ab} = 0$.

The likelihood equations for quasi-symmetry are

$$\begin{aligned} \hat{\mu}_{a+} &= n_{a+}, \qquad a = 1, \ldots, I \\ \hat{\mu}_{+b} &= n_{+b}, \qquad b = 1, \ldots, I \\ \hat{\mu}_{ab} + \hat{\mu}_{ba} &= n_{ab} + n_{ba} \qquad \text{for } a \le b. \end{aligned} \tag{10.20}$$

Only one of the first two sets of equations is needed. The other is redundant, given the other two. The residual df $= (I-1)(I-2)/2$. From (10.20), $\hat{\mu}_{aa} = n_{aa}$ for $a = 1, \ldots, I$. Otherwise, the likelihood equations do not have a direct solution. They are solved using iterative methods such as Newton–Raphson and IPF (Caussinus 1966).

The quasi-symmetry model has multiplicative form

$$\pi_{ab} = \alpha_a \beta_b \gamma_{ab}, \quad \text{where } \gamma_{ab} = \gamma_{ba} \quad \text{all } a < b \tag{10.21}$$

and all parameters are positive. The symmetry model is (10.21) with $\alpha_a = \beta_a$ for all a. This equation indicates that a table satisfying quasi-symmetry is the cellwise product of a table satisfying independence with one satisfying symmetry. The association symmetry implies that odds ratios on one side of the main diagonal are identical to corresponding odds ratios on the other side. In fact, the model can be defined by properties such as

$$\frac{\mu_{ab}\,\mu_{II}}{\mu_{aI}\,\mu_{Ib}} = \frac{\mu_{ba}\,\mu_{II}}{\mu_{bI}\,\mu_{Ia}} \quad \text{for all } a < b \tag{10.22}$$

or $\theta_{ab} = \theta_{ba}$ for local odds ratios. Goodman (1979a) referred to it as the *symmetric association* model.

The meaning of quasi-symmetry is less obvious than symmetry. However, it usually fits much better and has greater scope. One way to interpret its parameters relates to subject-specific logit models. For such models having additivity of subject terms and occasion terms, of which model (10.8) is the simplest case, the joint distribution in the corresponding population-averaged table necessarily satisfies quasi-symmetry (see Darroch 1981; Section 13.2.7 shows this). Consider the generalization of baseline-category logit model (10.15) to a subject-specific model

$$\log\big[P(Y_{it}=j)/P(Y_{it}=I)\big] = \alpha_{ij} + \beta_j x_t, \quad t=1,2, \quad j=1,\ldots,I-1.$$

This has the additive form of (10.8) for each j. The model implies, averaging over subjects, that the quasi-symmetry model (10.19) holds for the $I \times I$ population-averaged table with $\{\beta_j = \lambda_j^Y - \lambda_j^X\}$, when one constrains $\lambda_I^X = \lambda_I^Y = 0$. In fact, for the conditional ML analysis that conditions out $\{\alpha_{ij}\}$, the conditional ML estimates of $\{\hat{\beta}_j\}$ relate to the ordinary ML fit of quasi-symmetry by $\{\hat{\beta}_j = \hat{\lambda}_j^Y - \hat{\lambda}_j^X\}$ (Conaway 1989). This provides an interpretation for the main-effect terms in quasi-symmetry.

Related results hold for multiple occasions using a multivariate form (10.33) of quasi-symmetry (e.g., Agresti 1997; Conaway 1989; Darroch 1981; Tjur 1982; see also Section 13.2.7). In addition, quasi-symmetry contains as a special case other useful models. These include the ones in Sections 10.4.3 and 10.6.3.

10.4.3 Quasi-independence

Square tables usually exhibit positive dependence, manifested by larger counts on the main diagonal than the independence model predicts. Conditional on the event that a matched pair falls off the main diagonal, though, the relationship may have a simple structure.

A square contingency table satisfies *quasi-independence* when the variables are independent, given that the row and column outcomes differ. This has the loglinear form

$$\log \mu_{ab} = \lambda + \lambda_a^X + \lambda_b^Y + \delta_a I(a=b), \tag{10.23}$$

where $I(\cdot)$ is the indicator function,

$$I(a=b) = \begin{cases} 1, & a=b \\ 0, & a \neq b. \end{cases}$$

This adds a parameter to the independence model for each cell on the main diagonal. The first three terms in (10.23) specify independence, and $\{\delta_a\}$ permit $\{\mu_{aa}\}$ to depart from this pattern and have arbitrary positive values. When $\delta_a > 0$, μ_{aa} is larger than under independence.

The likelihood equations for quasi-independence are

$$\hat{\mu}_{a+} = n_{a+}, \qquad \hat{\mu}_{+a} = n_{+a}, \qquad \hat{\mu}_{aa} = n_{aa}, \quad a=1,\ldots,I.$$

A perfect fit occurs on the main diagonal, but independence holds for the remaining cells. The model implies that odds ratios equal 1.0 for all rectangularly formed 2×2 tables in which all cells fall off the main diagonal. One can fit the model using Newton–Raphson or IPF. The model has I more parameters than the independence model, so its residual df $= (I-1)^2 - I$. It applies to tables with $I \geq 3$.

Quasi-independence is the special case of quasi-symmetry (10.21) in which $\{\gamma_{ab}$ for $a \neq b\}$ are identical. Caussinus (1966, p. 146) showed that they are equivalent when $I = 3$.

10.4.4 Migration Revisited

We now return to Table 10.6 on migration patterns. Not surprisingly, the independence model fits terribly, with $G^2 = 125{,}923$ and $X^2 = 146{,}929$. (The maximum possible value of X^2 is $3n = 167{,}943$; see Problem 3.33.) The symmetry model is also unpromising. For instance, 124 people moved from the northeast to the west, but only 63 people made the reverse move. The deviance for testing symmetry is $G^2 = 243.6$ (df = 6).

Quasi-independence states that for people who moved, residence in 1985 is independent of region in 1980. Table 10.7 contains its fitted values, for which $G^2 = 69.5$ (df = 5). This model fits much better than the independence model, primarily because it forces a perfect fit on the main diagonal, where most observations occur. However, lack of fit is apparent off that diagonal. Many more people moved from the northeast to the south and many fewer moved from the west to the south than quasi-independence predicts.

TABLE 10.7 Fit of Models to Table 10.6

Residence in 1980	Residence in 1985[a]				
	Northeast	Midwest	South	West	Total
Northeast	11,607	100 (126.6)[1] (95.8)[2]	366 (312.9) (370.4)	124 (150.5) (123.8)	12,197
Midwest	87 (117.4) (91.2)	13,677	515 (531.1) (501.7)	302 (255.5) (311.1)	14,581
South	172 (133.2) (167.6)	255 (243.8) (238.3)	17,189	270 (290.0) (261.1)	18,486
West	63 (71.4) (63.2)	176 (130.6) (166.9)	286 (323.0) (294.9)	10,192	10,717
Total	11,929	14,178	18,986	10,888	55,981

[a][1] Quasi-independence fit; [2] quasi-symmetry fit; both models giving perfect fit on main diagonal.

The quasi-symmetry model has $G^2 = 3.0$, with df = 3. Table 10.7 displays its fit, which is much better than with quasi-independence. The lack of symmetry in cell probabilities reflects slight marginal heterogeneity. The subject-specific effects can be described using the model's parameter estimates, $\{\hat{\lambda}_1^Y - \hat{\lambda}_1^X = -0.672,\ \hat{\lambda}_2^Y - \hat{\lambda}_2^X = -0.623,\ \hat{\lambda}_3^Y - \hat{\lambda}_3^X = 0.122\}$. For instance, for a given subject the estimated odds of living in the south instead of the west in 1985 were $\exp(0.122) = 1.13$ times the odds in 1980. We'll see in Chapter 12 that such subject-specific effects tend to be stronger than those in corresponding marginal models, especially in tables like this with strong association.

A related application with matched samples is the study of occupational mobility. Each observation pairs parent's occupation with child's occupation (Goodman 1979b; Hout et al. 1987).

10.4.5 Marginal Homogeneity and Quasi-symmetry

Marginal homogeneity is not equivalent to a loglinear model. However, quasi-symmetry is a useful model for studying marginal homogeneity. Caussinus (1966) showed that symmetry is equivalent to quasi-symmetry and marginal homogeneity holding simultaneously. We have seen that symmetry implies both quasi-symmetry and marginal homogeneity. Now we give Caussinus's argument for the converse, that the joint occurrence of quasi-symmetry and marginal homogeneity implies symmetry.

From (10.21), if quasi-symmetry holds, $\pi_{ab} = \alpha_a \beta_b \gamma_{ab}$, where $\gamma_{ab} = \gamma_{ba} > 0$ for all $a < b$. Equivalently,

$$\pi_{ab} = \rho_a \delta_{ab},$$

where $\rho_a = \alpha_a / \beta_a$ and $\delta_{ab} = \beta_a \beta_b \gamma_{ab}$ also satisfies $\delta_{ab} = \delta_{ba} > 0$ for all $a < b$. If there is also marginal homogeneity, then

$$\pi_{j+} = \rho_j \sum_b \delta_{jb} = \sum_a \rho_a \delta_{aj} = \pi_{+j},$$

or

$$\rho_j = \left(\sum_a \rho_a \delta_{aj} \right) \Big/ \left(\sum_b \delta_{jb} \right) = \left(\sum_a \rho_a \delta_{aj} \right) \Big/ \left(\sum_b \delta_{bj} \right), \quad j = 1, \ldots, I.$$

Thus, each ρ_j is a weighted average of $\{\rho_a\}$, with weights $\{\delta_{aj} / \Sigma_b \delta_{bj} > 0, a = 1, \ldots, I\}$. Any set $\{\rho_a\}$ satisfying this must be identical. Otherwise, there would be a ρ_j that is no greater than any ρ_a but smaller than at least one, and hence it could not be a positive weighted average of all of them. But since $\{\rho_a\}$ are identical, $\pi_{ab} = \rho_a \delta_{ab} = \rho_b \delta_{ab} = \rho_b \delta_{ba} = \pi_{ba}$, so symmetry holds. Thus, a table that satisfies both quasi-symmetry and marginal homo-

geneity also satisfies symmetry. Since the converse holds,

$$\text{quasi-symmetry} + \text{marginal homogeneity} = \text{symmetry}. \qquad (10.24)$$

It follows that when quasi-symmetry (QS) holds, marginal homogeneity (MH) is equivalent to symmetry (S), which is $\{\lambda_a^X = \lambda_a^Y,\ a = 1, \ldots, I\}$ in the QS model. Thus, conditional on quasi-symmetry, testing marginal homogeneity is equivalent to testing symmetry. A test of marginal homogeneity compares fit statistics for the symmetry and quasi-symmetry models,

$$G^2(\mathrm{S} \mid \mathrm{QS}) = G^2(\mathrm{S}) - G^2(\mathrm{QS}), \qquad (10.25)$$

with df $= I - 1$. This is an alternative to approaches using marginal models discussed in Section 10.3.3.

Table 10.6 on migration from 1980 to 1985 has $G^2(\mathrm{S}) = 243.6$ and $G^2(\mathrm{QS}) = 3.0$. The difference $G^2(\mathrm{S} \mid \mathrm{QS}) = 240.6$ (df = 3) shows extremely strong evidence of marginal heterogeneity. Results are similar to those quoted in Section 10.3.4 for the likelihood-ratio test based on model (10.15), for which $G^2 = 240.8$, or the Wald test, for which $W = 236.5$ (both with df = 3).

10.4.6 Ordinal Quasi-symmetry Model

The loglinear models presented so far for square tables treat classifications as nominal. With ordered categories, more parsimonious models are useful. Let $u_1 \leq \cdots \leq u_I$ denote ordered scores for both the row and columns. An *ordinal quasi-symmetry model* is

$$\log \mu_{ab} = \lambda + \lambda_a + \lambda_b + \beta u_b + \lambda_{ab}, \qquad (10.26)$$

where $\lambda_{ab} = \lambda_{ba}$ for all $a < b$. It is the special case of the quasi-symmetry model (10.19) in which

$$\lambda_b^Y - \lambda_b^X = \beta u_b$$

has a linear trend. Symmetry is the special case $\beta = 0$.

This model has logit representation,

$$\log(\pi_{ab}/\pi_{ba}) = \beta(u_b - u_a) \qquad \text{for } a \leq b. \qquad (10.27)$$

This is the special case of the linear logit model, $\mathrm{logit}(\pi) = \alpha + \beta x$, with $\alpha = 0$, $x = u_b - u_a$ and π equal to the conditional probability of cell (a,b), given response sequence (a, b) or (b, a). The greater the value of $|\beta|$, the greater the difference between π_{ab} and π_{ba} and hence between the marginal distributions.

The likelihood equations for ordinal quasi-symmetry are

$$\sum_a u_a \hat{\mu}_{a+} = \sum_a u_a n_{a+}, \qquad \sum_b u_b \hat{\mu}_{+b} = \sum_b u_b n_{+b},$$

$$\hat{\mu}_{ab} + \hat{\mu}_{ba} = n_{ab} + n_{ba} \qquad \text{for} \quad a < b.$$

The fitted marginal counts need not equal the observed marginal counts. However, dividing the first two equations by n shows that they have the same means.

When $\beta \neq 0$, this model implies stochastically ordered margins. When $\beta > 0$ ($\beta < 0$), responses have a higher mean in the column (row) distribution. Like the ordinal marginal models (Section 10.3.1), this model concentrates the marginal effect on df = 1. A test of marginal homogeneity (H_0: $\beta = 0$) uses

ordinal quasi-symmetry + marginal homogeneity = symmetry.

The likelihood-ratio test statistic compares the deviance for symmetry and ordinal quasi-symmetry.

One can fit this model by fitting (10.27) with logit model software: Identify (n_{ab}, n_{ba}) as binomial with $n_{ab} + n_{ba}$ trials, and fit a logit model with no intercept and predictor $x = u_b - u_a$. One can also fit (10.26) using iterative methods for loglinear models.

10.4.7 Premarital and Extramarital Sex Revisited

For Table 10.5 on attitudes toward premarital and extramarital sex, a cursory glance at the data reveals that the symmetry model is inadequate ($G^2 = 402.2$, df = 6). By comparison, quasi-symmetry fits well ($G^2 = 1.4$, df = 3). The simpler model of ordinal quasi-symmetry also fits well: With scores $\{1, 2, 3, 4\}$, $G^2 = 2.1$ (df = 5).

The ML estimate $\hat{\beta} = -2.86$. From (10.27), the estimated probability that outcome on premarital sex is x categories more positive than the outcome on extramarital sex equals $\exp(2.86x)$ times the reverse probability. For instance, the estimated probability that premarital sex is judged almost always wrong and extramarital sex is always wrong equals $\exp(2.86) = 17.4$ times the estimated probability that premarital sex is always wrong and extramarital sex is almost always wrong.

10.4.8 Other Ordinal Models for Square Tables

For ordered classifications, when symmetry does not hold, often either $\pi_{ab} > \pi_{ba}$ for all $a < b$, or $\pi_{ab} < \pi_{ba}$ for all $a < b$. A generalization of

symmetry with this property is the logit model

$$\log(\pi_{ab}/\pi_{ba}) = \tau \qquad \text{for } a < b. \tag{10.28}$$

It implies that for all $a < b$,

$$P(Y_{i1} = a, Y_{i2} = b \,|\, Y_{i1} < Y_{i2}) = P(Y_{i1} = b, Y_{i2} = a \,|\, Y_{i1} > Y_{i2}).$$

The pattern of probabilities for cells above the main diagonal is a mirror image of the pattern for cells below it. This property is called *conditional symmetry* (McCullagh 1978). Problem 10.35 shows the corresponding loglinear model and its fit. Symmetry is the special case $\tau = 0$.

Another model generalizes quasi-independence. Let $\{u_a\}$ be ordered scores. The model

$$\log \mu_{ab} = \lambda + \lambda_a^X + \lambda_b^Y + \beta u_a u_b + \delta_a I(a = b) \tag{10.29}$$

permits linear-by-linear association [see (9.6)] off the main diagonal. It is a special case of quasi-symmetry, and quasi-independence is the special case $\beta = 0$. For equal-interval scores, it implies uniform local association, given that responses differ. Goodman (1979a) called it *quasi-uniform association*.

For Table 10.5 on opinions about premarital and extramarital sex, the conditional symmetry model has $\hat{\tau} = -4.130$ (SE = 0.451). The estimated probability that extramarital sex is considered more wrong are $\exp(4.13) = 62.2$ times the estimated probability that premarital sex is considered more wrong. The quasi-uniform association model has $\hat{\beta} = 0.632$ (SE = 0.106). Off the main diagonal, the estimated local odds ratio equals $\exp(0.632) = 1.88$.

10.5 MEASURING AGREEMENT BETWEEN OBSERVERS

We now discuss an application, analyzing agreement between two observers, that uses matched-pairs models. We illustrate with Table 10.8. This shows ratings by two pathologists, labeled A and B, who separately classified 118 slides regarding the presence and extent of carcinoma of the uterine cervix. The rating scale has the ordered categories (1) negative, (2) atypical squamous hyperplasia, (3) carcinoma *in situ*, (4) squamous or invasive carcinoma.

10.5.1 Agreement: Departures from Independence

Let π_{ab} denote the probability that observer A classifies a slide in category a and observer B classifies it in category b. Then π_{aa} is the probability that they both choose category a, and $\sum_a \pi_{aa}$ is the total probability of agreement. Perfect agreement occurs when $\sum_a \pi_{aa} = 1$.

With subjective scales, agreement is less than perfect. Analyses focus on describing strength of agreement and detecting patterns of disagreement.

TABLE 10.8 Diagnoses of Carcinoma

	Pathologist B[a]				
Pathologist A	1	2	3	4	Total
1	22 (8.5)	2 (−0.5)	2 (−5.9)	0 (−1.8)	26
2	5 (−0.5)	7 (3.2)	14 (−0.5)	0 (−1.8)	26
3	0 (−4.1)	2 (−1.2)	36 (5.5)	0 (−2.3)	38
4	0 (−3.3)	1 (−1.3)	17 (0.3)	10 (5.9)	28
Total	27	12	69	10	118

[a] Values in parentheses are standardized Pearson residuals for the independence model.

Source: N. S. Holmquist, C. A. McMahon, and O. D. Williams, *Arch. Pathol.* **84**: 334–345 (1967); reprinted with permission from the American Medical Association. See also Landis and Koch (1977).

Agreement and *association* are distinct facets of the joint distribution. Strong agreement requires strong association, but strong association can exist without strong agreement. If observer A consistently rates subjects one category higher than observer B, strength of agreement is poor even though the association is strong.

Evaluations of agreement compare $\{n_{ab}\}$ to the values $\{n_{a+}n_{+b}/n\}$ predicted under independence. That model is a baseline, showing the agreement expected if no association existed between ratings. Normally, it fits poorly if even mild agreement exists, but its cell standardized residuals (Section 3.3.1) show patterns of agreement and disagreement. Ideally, standardized residuals are large positive on the main diagonal and large negative off that diagonal. The sizes are influenced by sample size n, however, larger values tending to occur as n increases.

The independence model fits Table 10.8 poorly ($G^2 = 118.0$, df = 9). That table reports the standardized Pearson residuals in parentheses. The large positive residuals on the main diagonal indicate that agreement for each category is greater than expected by chance, especially for the first category. Off the main diagonal they are primarily negative. Disagreements occurred less than expected under independence, although the evidence of this is weaker for categories closer together. The most common disagreements were observer B choosing category 3 and observer A instead choosing category 2 or 4.

10.5.2 Using Quasi-independence to Analyze Agreement

More complex models add components that relate to agreement beyond that expected under independence. A useful generalization is quasi-independence

TABLE 10.9 Fitted Values for Carcinoma Diagnoses of Table 10.8

	Pathologist B^a			
Pathologist A	1	2	3	4
1	22	2	2	0
	$(22)^1$	(0.7)	(3.3)	(0.0)
	$(22)^2$	(2.4)	(1.6)	(0.0)
2	5	7	14	0
	(2.4)	(7)	(16.6)	(0.0)
	(4.6)	(7)	(14.4)	(0.0)
3	0		36	0
	(0.8)	(1.2)	(36)	(0.0)
	(0.4)	(1.6)	(36)	(0.0)
4	0	1	17	10
	(1.9)	(3.0)	(13.1)	(10)
	(0.0)	(1.0)	(17.0)	(10)

[a] [1]Quasi-independence model; [2]quasi-symmetry model.

(10.23), which adds main-diagonal parameters $\{\delta_a\}$. For Table 10.8, this model has $G^2 = 13.2$ (df = 5). It fits much better than independence, but some lack of fit remains. Table 10.9 shows the fit.

For two subjects, suppose that each observer classifies one in category a and one in category b. The odds that the observers agree rather than disagree on which is in category a and which is in category b equal

$$\tau_{ab} = \frac{\pi_{aa}\pi_{bb}}{\pi_{ab}\pi_{ba}} = \frac{\mu_{aa}\,\mu_{bb}}{\mu_{ab}\,\mu_{ba}}. \qquad (10.30)$$

As τ_{ab} increases, the observers are more likely to agree for that pair of categories. Under quasi-independence,

$$\tau_{ab} = \exp(\,\delta_a + \delta_b)\,.$$

Larger $\{\delta_a\}$ represent stronger agreement. For instance, for Table 10.8, $\hat{\delta}_2 = 0.6$ and $\hat{\delta}_3 = 1.9$, and $\hat{\tau}_{23} = 12.3$. The degree of agreement also seems fairly strong for other pairs of categories.

10.5.3 Quasi-symmetry and Agreement Modeling

For Table 10.8, the quasi-independence model shows some lack of fit. Given that the pathologists disagree, some association remains between ratings. For observer agreement tables, this is common. Quasi-symmetry (10.19) often fits much better, because it permits association. For Table 10.8, it has $G^2 = 1.0$ (df = 2). Table 10.9 displays the fit. It is not unusual for tables to have many

empty cells. When $n_{ab} + n_{ba} = 0$ for any pair (such as categories 1 and 4 in Table 10.8), the ML fitted values for quasi-symmetry in those cells must also be zero since one of its likelihood equations is $\hat{\mu}_{ab} + \hat{\mu}_{ba} = n_{ab} + n_{ba}$. One should eliminate those cells from the fitting process to get the proper residual df value.

Under quasi-symmetry, $\hat{\tau}_{ab} = \exp(\hat{\lambda}_{aa} + \hat{\lambda}_{bb} - \hat{\lambda}_{ab} - \hat{\lambda}_{ba})$, where $\hat{\lambda}_{ab} = \hat{\lambda}_{ba}$. For categories 2 and 3 of Table 10.8, for instance, $\hat{\tau}_{23} = 10.7$.

Loglinear models directly address the association component of agreement. The quasi-symmetry model also yields information about similarity of marginal distributions. The simpler symmetry model that forces the margins to be identical fits Table 10.8 poorly ($G^2 = 39.2$, df = 5). The statistic $G^2(S \mid QS) = 39.2 - 1.0 = 38.2$ (df = 3) provides strong evidence of marginal heterogeneity. In Table 10.8, differences in marginal proportions are substantial in each category but the first. The marginal heterogeneity is one reason that the agreement is not stronger.

Models for agreement can take ordering of categories into account. Conditional on observer disagreement, a tendency usually remains for high (low) ratings by one observer to occur with relatively high (low) ratings by the other observer (see Problem 10.41).

10.5.4 Kappa Measure of Agreement

An alternative approach summarizes agreement with a single index. For nominal scales, the most popular measure is *Cohen's kappa* (Cohen 1960). It compares the probability of agreement $\sum_a \pi_{aa}$ to that expected if the ratings were independent, $\sum_a \pi_{a+}\pi_{+a}$, by

$$\kappa = \frac{\sum_a \pi_{aa} - \sum_a \pi_{a+}\pi_{+a}}{1 - \sum_a \pi_{a+}\pi_{+a}}.$$

The denominator equals the numerator with $\sum_a \pi_{aa}$ replaced by its maximum possible value of 1, corresponding to perfect agreement. Kappa equals 0 when the agreement merely equals that expected under independence. It equals 1.0 when perfect agreement occurs. The stronger the agreement, the higher is κ, for given marginal distributions. Negative values occur when agreement is weaker than expected by chance, but this rarely happens.

For multinomial sampling, the sample value $\hat{\kappa}$ has a large-sample normal distribution. Its estimated asymptotic variance (Fleiss et al. 1969) is

$$\hat{\sigma}^2(\hat{\kappa}) = \frac{1}{n}\left\{\frac{P_o(1-P_o)}{(1-P_e)^2} + \frac{2(1-P_o)[2P_oP_e - \sum_a p_{aa}(p_{a+}+p_{+a})]}{(1-P_e)^3} + \frac{(1-P_o)^2\left[\sum_a\sum_b p_{ab}(p_{b+}+p_{+a})^2 - 4P_e^2\right]}{(1-P_e)^4}\right\},$$

where $P_o = \sum_a p_{aa}$ and $P_e = \sum_a p_{a+}p_{+a}$. It is rarely plausible that agreement is no better than expected by chance. Thus, rather than testing H_0: $\kappa = 0$, it is more relevant to estimate strength of agreement by interval estimation of κ.

For Table 10.8, $P_o = 0.636$ and $P_e = 0.281$. Sample kappa equals $(0.636 - 0.281)/(1 - 0.281) = 0.493$. The difference between observed agreement and that expected under independence is about 50% of the maximum possible difference. The estimated standard error is 0.057, so κ apparently falls roughly between 0.4 and 0.6, moderately strong agreement.

10.5.5 Weighted Kappa: Quantifying Disagreement

Kappa treats classifications as nominal. When categories are ordered, the seriousness of a disagreement depends on the difference between the ratings. For nominal classifications also, some disagreements may be considered more severe than others. The measure *weighted kappa* (Spitzer et al. 1967) uses weights $\{w_{ab}\}$ satisfying $0 \le w_{ab} \le 1$, with all $w_{aa} = 1$ and all $w_{ab} = w_{ba}$ to describe closeness of agreement. One possibility is $\{w_{ab} = 1 - |a - b|/(I - 1)\}$, for which agreement is greater for cells nearer the main diagonal. Fleiss and Cohen (1973) suggested $\{w_{ab} = 1 - (a - b)^2/(I - 1)^2\}$. The weighted agreement is $\sum_a \sum_b w_{ab}\pi_{ab}$ and weighted kappa is

$$\kappa_w = \frac{\sum_a \sum_b w_{ab}\pi_{ab} - \sum_a \sum_b w_{ab}\pi_{a+}\pi_{+b}}{1 - \sum_a \sum_b w_{ab}\pi_{a+}\pi_{+b}}.$$

Controversy surrounds the utility of kappa and weighted kappa, partly because their values depend strongly on the marginal distributions. The same diagnostic rating process can yield quite different values, depending on the proportions of cases of the various types (Problem 10.40). In summarizing a contingency table by a single number, the reduction in information can be severe. It is helpful to construct models providing more detailed investigation of the agreement and disagreement structure rather than to depend solely on a summary index.

10.5.6 Extensions to Multiple Observers

With several observers, ordinary loglinear models are not usually relevant. Their description of agreement and association between two observers is conditional on ratings by the others. It is more relevant to study this marginally, without conditioning on the other ratings. Hence, for R observers, modelling simultaneously the pairwise agreement and association structure requires studying the $\binom{R}{2}$ pairs of two-way marginal distributions (Becker and Agresti 1992).

Other approaches have also been used. For instance, generalizations of kappa summarize pairwise agreements or multiple agreements (Fleiss 1981, Sec. 13.2; Landis and Koch 1977). Or, it may make sense to use a mixture model that assumes latent classes of subjects for whom the observers agree and subjects for whom they disagree. Such an analysis is shown in Section 13.1.2.

10.6 BRADLEY–TERRY MODEL FOR PAIRED PREFERENCES

Sometimes, categorical outcomes result from pairwise evaluations. A common example is athletic competitions, when the outcome for a team or player consists of categories (win, lose). Another example is pairwise comparison of product brands, such as two brands of wine of some type. When a wine critic rates I brands of sauvignon blanc, it might be difficult to establish an outright ranking, especially if I is large. However, for any given pair, the critic could probably state a preference after tasting them at the same occasion. An overall ranking of the wines could then be based on the pairwise preferences. We present a model for this in this section.

10.6.1 Bradley–Terry Model

Bradley and Terry (1952) proposed a logit model for paired evaluations. Let Π_{ab} denote the probability that a is preferred to b. Suppose that $\Pi_{ab} + \Pi_{ba} = 1$ for all pairs; that is, a tie cannot occur. The Bradley–Terry model is

$$\log\frac{\Pi_{ab}}{\Pi_{ba}} = \beta_a - \beta_b. \tag{10.31}$$

Alternatively,

$$\Pi_{ab} = \exp(\beta_a)/\left[\exp(\beta_a) + \exp(\beta_b)\right].$$

Thus, $\Pi_{ab} = \frac{1}{2}$ when $\beta_a = \beta_b$ and $\Pi_{ab} > \frac{1}{2}$ when $\beta_a > \beta_b$.

Identifiability requires a constraint such as $\beta_I = 0$ or $\sum_a \exp(\hat{\beta}_a) = 1$. Since the model describes $\binom{I}{2}$ probabilities ($\{\Pi_{ab}\}$ for $a < b$) by $(I - 1)$ parameters, residual df $= \binom{I}{2} - (\mathrm{I} - 1)$.

For $a < b$, let N_{ab} denote the sample number of evaluations, with a preferred n_{ab} times and b preferred $n_{ba} = N_{ab} - n_{ab}$ times. A square contingency table with empty cells on the main diagonal summarizes results. When the N_{ab} comparisons are independent with probability Π_{ab} for each, n_{ab} has a bin(N_{ab}, Π_{ab}) distribution. If evaluations for different pairs are also independent, ordinary methods for logit models apply for fitting the model.

TABLE 10.10 Results of 1987 Season for American League Baseball Teams

Winning Team	Losing Team[a]						
	Milwaukee	Detroit	Toronto	New York	Boston	Cleveland	Baltimore
Milwaukee	—	7 (7.0)	9 (7.4)	7 (7.6)	7 (8.0)	9 (9.2)	11 (10.8)
Detroit	6 (6.0)	—	7 (7.0)	5 (7.1)	11 (7.6)	9 (8.8)	9 (10.5)
Toronto	4 (5.6)	6 (6.0)	—	7 (6.7)	7 (7.1)	8 (8.4)	12 (10.2)
New York	6 (5.4)	8 (5.9)	6 (6.3)	—	6 (7.0)	7 (8.3)	10 (10.1)
Boston	6 (5.0)	2 (5.4)	6 (5.9)	7 (6.0)	—	7 (7.9)	12 (9.8)
Cleveland	4 (3.8)	4 (4.2)	5 (4.6)	6 (4.7)	6 (5.1)	—	6 (8.6)
Baltimore	2 (2.2)	4 (2.5)	1 (2.8)	3 (2.9)	1 (3.2)	7 (4.4)	—

[a] Values in parentheses represent the fit of the Bradley–Terry model.

Source: American League Red Book, 1988 (St. Louis, MO: Sporting News Publishing Co.)

10.6.2 Home Team Advantage in Baseball

Table 10.10 shows results of the 1987 season for the seven baseball teams in the Eastern Division of the American League. For instance, of games between Boston and New York, Boston won 7 and New York won 6. Table 10.10 shows the population of regular-season games. We regard this as a sample estimate of a conceptual distribution representing the long-run performance of teams as constituted in 1987.

We fitted the Bradley–Terry model as a logit model for $\binom{7}{2} = 21$ independent binomial samples, using an appropriate model matrix and no intercept (e.g., for SAS, see Table A.19). The model fits adequately ($G^2 = 15.7$, df = 15). Table 10.10 contains the fitted values $\{\hat{\mu}_{ab}\}$. Table 10.11 displays the sample proportion of games each team won and the model estimates of $\{\hat{\beta}_a\}$ (setting $\hat{\beta}_7 = 0$) and $\{\exp(\hat{\beta}_a)\}$ [setting $\Sigma_a \exp(\hat{\beta}_a) = 1$]. When Boston played New York, the estimated probability that Boston won is

$$\hat{\Pi}_{54} = \exp(\hat{\beta}_5) \Big/ \left[\exp(\hat{\beta}_5) + \exp(\hat{\beta}_4)\right] = 0.46.$$

The standard error of each $\hat{\beta}_a$ and of each $\hat{\beta}_a - \hat{\beta}_b$ is about 0.3, so not much evidence exists of a difference among the top five teams.

TABLE 10.11 Results of Fitting Bradley–Terry Models to Baseball Data

Team	Winning Percentage	$\hat{\beta}_i$ (10.31)	$\exp(\hat{\beta}_i)$ (10.31)	$\exp(\hat{\beta}_i)$ (10.32)
Milwaukee	64.1	1.58	0.218	0.220
Detroit	60.2	1.44	0.189	0.190
Toronto	56.4	1.29	0.164	0.164
New York	55.1	1.25	0.158	0.157
Boston	51.3	1.11	0.136	0.137
Cleveland	39.7	0.68	0.089	0.088
Baltimore	23.1	0.00	0.045	0.044

TABLE 10.12 Wins / Losses by Home and Away Team, 1987

	Away Team						
Home Team	Milwaukee	Detroit	Toronto	New York	Boston	Cleveland	Baltimore
Milwaukee	—	4-3	4-2	4-3	6-1	4-2	6-0
Detroit	3-3	—	4-2	4-3	6-0	6-1	4-3
Toronto	2-5	4-3	—	2-4	4-3	4-2	6-0
New York	3-3	5-1	2-5	—	4-3	4-2	6-1
Boston	5-1	2-5	3-3	4-2	—	5-2	6-0
Cleveland	2-5	3-3	3-4	4-3	4-2	—	2-4
Baltimore	2-5	1-5	1-6	2-4	1-6	3-4	—

Source: American League Red Book, 1988 (St. Louis, MO: Sporting News Publishing Co.).

This model does not recognize which team is the home team. Most sports have a home field advantage: A team is more likely to win when it plays at its home city. Table 10.12 contains results for the 1987 season according to the (home team, away team) classification. For instance, when Boston was the home team, it beat New York 4 times and lost 2 times; when New York was the home team, it beat Boston 4 times and lost 3 times. Now for all $a \neq b$, let Π^*_{ab} denote the probability that team a beats team b, when a is the home team. Consider logit model

$$\log \frac{\Pi^*_{ab}}{1 - \Pi^*_{ab}} = \alpha + (\beta_a - \beta_b). \tag{10.32}$$

When $\alpha > 0$, a home field advantage exists. The home team of two evenly matched teams has probability $\exp(\alpha)/[1 + \exp(\alpha)]$ of winning.

For Table 10.12, model (10.32) describes 42 binomial distributions with 7 parameters. It has $G^2 = 38.6$ (df = 35). Table 10.11 displays $\{\exp(\hat{\beta}_a)\}$, which are similar to those obtained previously. The estimate of the home-field parameter is $\hat{\alpha} = 0.302$. For two evenly matched teams, the home team had estimated probability 0.575 of winning. When Boston played New York, the estimated probability of a Boston win was 0.54 at Boston and 0.39 at New York.

Model (10.32) is a useful generalization of the Bradley–Terry model whenever an *order effect* exists. For instance, in pairwise taste evaluations, the product tasted first may have a slight advantage.

10.6.3 Bradley–Terry Model and Quasi-symmetry

Fienberg and Larntz (1976) showed that the Bradley–Terry model is a logit formulation of the quasi-symmetry model (10.19). For quasi-symmetry, given that an observation is in cell (a, b) or (b, a), the logit of the conditional

probability of cell (a, b) equals

$$\log\frac{\mu_{ab}}{\mu_{ba}} = \left(\lambda + \lambda_a^X + \lambda_b^Y + \lambda_{ab}^{XY}\right) - \left(\lambda + \lambda_b^X + \lambda_a^Y + \lambda_{ba}^{XY}\right)$$
$$= \left(\lambda_a^X - \lambda_a^Y\right) - \left(\lambda_b^X - \lambda_b^Y\right) = \beta_a - \beta_b,$$

where $\beta_a = \lambda_a^X - \lambda_a^Y$. Estimates $\{\hat{\lambda}_a^X\}$ and $\{\hat{\lambda}_a^Y\}$ for quasi-symmetry yield $\{\hat{\beta}_a\}$ for the Bradley–Terry model.

10.6.4 Extensions to Ties and Ordinal Evaluations

The Bradley–Terry model extends to ordinal comparisons, such as the evaluation scale (much better, slightly better, the same, slightly worse, much worse) in comparing two products. With cumulative logits and an I-category evaluation scale, let Y_{ab} denote the response for a comparison of a with b. The model is

$$\text{logit}\left[P(Y_{ab} \le j)\right] = \alpha_j + (\beta_a - \beta_b).$$

Since $P(Y_{ab} \le j) = P(Y_{ba} > I - j) = 1 - P(Y_{ba} \le I - j)$, it follows that $\text{logit}[P(Y_{ab} \le j] = -\text{logit}[P(Y_{ba} \le I - j)]$. Thus, necessarily, $\alpha_j = -\alpha_{I-j}$.

The most common ordered preference scale is (win, tie, lose). Then, $\alpha_1 = -\alpha_2$.

10.7 MARGINAL AND QUASI-SYMMETRY MODELS FOR MATCHED SETS*

Methods for matched pairs extend to matched sets. Here we present mainly the loglinear modeling approach; in Chapters 11 and 12 we present extensions of the marginal and conditional logit modeling approaches.

10.7.1 Marginal Homogeneity, Complete Symmetry, and Quasi-symmetry

Let $(Y_1, Y_2, \ldots, Y_T)$ denote the T responses in each matched set. With I response categories, a contingency table with I^T cells summarizes the possible outcomes. Let $\mathbf{i} = (i_1, \ldots, i_T)$ denote the cell having $Y_t = i_t$, $t = 1, \ldots, T$. Let $\pi_{\mathbf{i}} = P(Y_t = i_t,\ t = 1, \ldots, T)$, and let $\mu_{\mathbf{i}} = n\pi_{\mathbf{i}}$. Then

$$P(Y_t = j) = \pi_{+\cdots+j+\cdots+},$$

where the j subscript is in position t, and $\{P(Y_t = j),\ j = 1, \ldots, I\}$ is the marginal distribution for Y_t.

This T-way table satisfies *marginal homogeneity* if

$$P(Y_1 = j) = P(Y_2 = j) = \cdots = P(Y_T = j) \qquad \text{for } j = 1, \ldots, I.$$

It satisfies *complete symmetry* if

$$\pi_{\mathbf{i}} = \pi_{\mathbf{j}}$$

for any permutation $\mathbf{j} = (j_1, \ldots, j_T)$ of $\mathbf{i} = (i_1, \ldots, i_T)$. Complete symmetry implies marginal homogeneity, but the converse does not hold except when $T = I = 2$.

Complete symmetry is a loglinear model. One representation is

$$\log \mu_{\mathbf{i}} = \lambda_{ab\ldots m},$$

where a is the minimum of $(i_1, \ldots, i_T)$, b is the next smallest, ..., and m is the maximum. In a three-way table, for instance, $\log \mu_{122} = \log \mu_{212} = \log \mu_{221} = \lambda_{122}$. The number of $\{\lambda_{ab\ldots m}\}$ parameters is the number of ways of selecting T out of I items with replacement, which is $\binom{I+T-1}{T}$. Thus, residual df $= I^T - \binom{I+T-1}{T}$ (Haberman 1978, p. 518).

An I^T table satisfies *quasi-symmetry* if

$$\log \mu_{\mathbf{i}} = \lambda_{1i_1} + \lambda_{2i_2} + \cdots + \lambda_{Ti_T} + \lambda_{ab\ldots m} \tag{10.33}$$

where $\lambda_{ab\ldots m}$ is defined as in the complete symmetry model. It has symmetric association and higher-order interaction terms, but permits each single-factor marginal distribution to have its own parameters. Identifiability requires constraints such as $\lambda_{tI} = 0$ for each t. One set of main-effect terms is redundant (Problem 10.31). This model has $(I-1)(T-1)$ more parameters than complete symmetry. It is fitted using iterative methods.

For ordinal responses, a simpler model with quantitative main effects uses ordered scores $\{u_a\}$. The *ordinal quasi-symmetry model* is

$$\log \mu_{\mathbf{i}} = \beta_1 u_{i_1} + \beta_2 u_{i_2} + \cdots + \beta_T u_{i_T} + \lambda_{ab\ldots m}$$

where one can set $\beta_T = 0$. Complete symmetry is the special case $\beta_1 = \cdots = \beta_T$.

When quasi-symmetry (10.33) or ordinal quasi-symmetry holds, marginal homogeneity is equivalent to complete symmetry. Marginal heterogeneity occurs if quasi-symmetry (QS) holds but complete symmetry (S) does not. The statistic

$$G^2(\mathrm{S} \mid \mathrm{QS}) = G^2(\mathrm{S}) - G^2(\mathrm{QS})$$

tests marginal homogeneity. Under complete symmetry, it is asymptotically chi-squared with df $= (I-1)(T-1)$. The corresponding test for the ordinal quasi-symmetry model has df $= (T-1)$.

10.7.2 Attitudes toward Legalized Abortion Example

Refer to Table 10.13. Subjects indicated whether they support legalized abortion in three situations: (1) if the family has a very low income and cannot afford any more children, (2) when the woman is not married and does not want to marry the man, and (3) when the woman wants it for any reason. The table also classifies subjects by gender, resulting in a 2^4 table.

Let μ_{ghij} denote the expected frequency for gender g (1 = female; 0 = male) with response sequence (h, i, j) for the three questions. Consider the model

$$\log \mu_{ghij} = \beta g + \lambda_{abc},$$

where the interaction term is λ_{111} when $(h, i, j) = (1, 1, 1)$, λ_{112} when $(h, i, j) = (1, 1, 2)$ or $(1, 2, 1)$ or $(2, 1, 1)$, λ_{122} when $(h, i, j) = (1, 2, 2)$ or $(2, 1, 2)$ or $(2, 2, 1)$, and λ_{222} when $(h, i, j) = (2, 2, 2)$. This model implies the same complete symmetry pattern of probabilities for each gender. Its fit has $G^2 = 39.2$ with df $= 11$.

Adding main-effect terms for the three issues implies the same quasi-symmetric pattern for each gender. It fits much better, having $G^2 = 10.2$ with df $= 9$. Thus, it seems plausible to assume a symmetric association structure. In fact, the loglinear model with only two-factor association terms has fitted log odds ratios of 3.2 for items 1 and 2, 2.6 for items 1 and 3, and 3.3 for items 2 and 3.

One can test marginal homogeneity, given gender, by the likelihood-ratio statistic $39.2 - 10.2 = 29.0$, with df $= 2$. An analysis of the main-effect terms in the quasi-symmetry model shows greater support for legalized abortion when the family has a low income and cannot afford any more children than in the other two instances.

TABLE 10.13 Support for Legalizing Abortion in Three Situations, by Gender

	Sequence of Responses on the Three Items[a]							
Gender	(1, 1, 1)	(1, 1, 2)	(2, 1, 1)	(2, 1, 2)	(1, 2, 1)	(1, 2, 2)	(2, 2, 1)	(2, 2, 2)
Male	342	26	6	21	11	32	19	356
Female	440	25	14	18	14	47	22	457

[a] Items are (1) if the family has a very low income and cannot afford anymore children, (2) when the woman is not married and does not want to marry the man, and (3) when the woman wants it for any reason. 1, yes; 2, no.

Source: Data from 1994 General Social Survey, National Opinion Research Center.

10.7.3 Types of Marginal Symmetry

A general type of symmetry for I^T tables has marginal homogeneity and complete symmetry as special cases. For an I^T table, $P(Y_{t_1} = j_1, \ldots, Y_{t_h} = j_h)$, where h is between 1 and T, is a h-dimensional marginal probability, $h = 1$ giving single-variable marginal probabilities. There is hth-*order marginal symmetry* if for all h-tuples $\mathbf{j} = (j_1, \ldots, j_h)$, this probability is the same for each permutation of $\mathbf{j}$ and for all combinations $\mathbf{t} = (t_1, \ldots, t_h)$ of h of the T responses.

For $h = 1$, first-order marginal symmetry is marginal homogeneity. Second-order marginal symmetry occurs if for all t and u, $P(Y_t = a, Y_u = b)$ is the same and the equality holds for all pairs of outcomes (a, b). In other words, the two-way marginal tables exhibit symmetry, and they are identical. Tth-order marginal symmetry in an I^T table is complete symmetry.

When hth-order symmetry holds, ith-order marginal symmetry holds for any $i < h$. For instance, complete symmetry implies second-order marginal symmetry, which itself implies marginal homogeneity. Although this hierarchy is mathematically attractive, the higher-order symmetries are usually too restrictive to fit well in practice.

10.7.4 Marginal Models: Multiway Tables

In practice, usually the form of the joint distribution is of secondary interest. Research questions pertain instead to the marginal distributions. The marginal models of Section 10.3 for matched pairs extend to matched sets. For instance, with ordinal classifications, a cumulative logit model is

$$\text{logit}\big[P(Y_t \le j)\big] = \alpha_j + \beta_t, \qquad j = 1, \ldots, I-1, \quad t = 1, \ldots, T. \quad (10.34)$$

In the next chapter we study marginal models in more general contexts, extending the analyses of this chapter to incorporate matched sets and explanatory variables.

NOTES

Section 10.1: Comparing Dependent Proportions

10.1. Miettinen (1969) generalized the McNemar test to case–control sets having several controls per case. The Table 10.2 representation is then useful. Each of n matched sets forms a stratum of a $2 \times 2 \times n$ table with one observation in column 1 (the case) and several observations in column 2 (the controls).

Altham (1971) and Ghosh et al. (2000) presented Bayesian analyses for binary matched pairs. Copas (1973), Gart (1969), Kenward and Jones (1994), and Miettinen (1969) studied generalizations of matched-pairs designs. With some approaches (Ghosh et al. 2000; Liang and Zeger 1988; Suissa and Shuster 1991), inferences about marginal homogeneity also use the main-diagonal observations.

Section 10.4: Symmetry, Quasi-symmetry, and Quasi-independence

10.2. For other discussion of quasi-symmetry, see Darroch (1981) and McCullagh (1982). The term *quasi-independence* originated in Goodman (1968). A more general definition of it is $\pi_{ab} = \alpha_a \beta_b$ for some fixed set of cells. See Caussinus (1966), Fienberg (1970b, 1972), and Goodman (1968). Caussinus used the concept to analyze tables that deleted a certain set of cells from consideration, and Goodman used it in earlier analyses of social mobility. Altham (1975) used it with triangular tables, for which observations occur only above or only below the main diagonal. Stigler (1999, Chap. 19) summarized early uses, including Karl Pearson's handling in 1913 of a triangular array. Booth and Butler (1999) and Smith et al. (1996) discussed exact tests for square-table models.

10.3. The effect β in ordinal quasi-symmetry relates to the occasion effect in a subject-specific adjacent-categories-logit model (Agresti 1993). Conditional symmetry is a special case of *diagonals-parameter symmetry*,

$$\log(\pi_{ab}/\pi_{ba}) = \tau_{b-a}, \qquad a < b.$$

See Goodman (1979b, 1985) and Hout et al. (1987).

10.4. In some applications a table is *a priori* symmetric or independent, but one can observe only the pair (i, j) rather than their order, thus leading to an upper-triangular table. See Khamis (1983) for examples and ML fitting of models for such three-way tables that are symmetric within layers.

Section 10.5: Measuring Agreement between Observers

10.5. Kappa and weighted kappa relate to the intraclass correlation, a measure of interrater reliability for interval scales (Fleiss 1981; Fleiss and Cohen 1973; Kraemer 1979). Banerjee et al. (1999) and Fleiss (1981, Chap. 13) reviewed kappa and its generalizations. See Becker and Agresti (1992), Goodman (1979b), Tanner and Young (1985), and Problem 10.41 for examples of modeling agreement with loglinear models. Darroch and McCloud (1986) showed that quasi-symmetry has an important role in agreement modeling.

Section 10.6: Bradley–Terry Model for Paired Preferences

10.6. Zermelo (1929) proposed a model that is equivalent to the Bradley–Terry model. Luce (1959) provided an axiomatic basis for it. Mosteller (1951) and Thurstone (1927) proposed an analogous model with probit link. An interesting interview of Ralph Bradley by M. Hollander (*Stat. Sci.* **16**: 75–100, 2001) discussed food-tasting applications that motivated its development. For extensions, see Bradley (1976). Fienberg and Larntz (1976) and Imrey et al. (1976) related it to quasi-independence. Dittrich et al. (1998) allowed covariates. Matthews and Morris (1995) gave an application with a factorial design, ties, and allowance for dependence among judgments. Böckenholt and Dillon (1997) modeled dependence with ordinal preferences. David (1988) and Imrey (1998) surveyed paired preference methods.

TABLE 10.14 Data for Problem 10.1

	Let Patient Die	
Suicide	Yes	No
Yes	1097	90
No	203	435

Source: 1994 General Social Survey, National Opinion Research Center.

PROBLEMS

Applications

10.1 Table 10.14 shows results when subjects were asked "Do you think a person has the right to end his or her own life if this person has an incurable disease?" and "When a person has a disease that cannot be cured, do you think doctors should be allowed to end the patient's life by some painless means if the patient and his family request it?" The table refers to these variables as "suicide" and "let patient die."

a. Compare the marginal proportions using a confidence interval.

b. Perform McNemar's test, and interpret.

c. Find the conditional ML estimate of β for model (10.8). Interpret.

10.2 Refer to Table 8.16 and Problem 8.1. Treat the data as matched pairs on opinion, stratified by gender. Testing independence for the 2×2 table using entries $(6, 160)$ in row 1 and $(11, 181)$ in row 2 tests equality of β for logit model (10.8) for each gender. Explain why.

10.3 A crossover experiment with 100 subjects compares two drugs for treating migraine headaches. The response scale is success (1) or failure (0). Half the study subjects, randomly selected, used drug A the first time they had a headache and drug B the next time. For them, 6 had outcomes $(1, 1)$ for (A, B), 25 had outcomes $(1, 0)$, 10 had outcomes $(0, 1)$, and 9 had outcomes $(0, 0)$. For the 50 subjects who took the drugs in the reverse order, 10 were $(1, 1)$ for (A, B), 20 were $(1, 0)$, 12 were $(0, 1)$, and 8 were $(0, 0)$.

a. Ignoring treatment order, compare the success probabilities for the two drugs. Interpret.

b. McNemar's test uses only the pairs of outcomes that differ. For this study, Table 10.15 shows such data from both treatment orders. Testing independence for this table tests whether success rates are identical for the treatments (Gart 1969). Explain why. Analyze these data, and interpret.

TABLE 10.15 Data for Problem 10.3

Treatment Order	Treatment That Is Better	
	First	Second
A, then B	25	10
B, then A	12	20

10.4 A case–control study has 8 pairs of subjects. The cases have colon cancer, and the controls are matched with the cases on gender and age. A possible explanatory variable is the extent of red meat in a subject's diet, measured as "1 = high" or "0 = low." The (case, control) observations on this were $(1, 1)$ for 3 pairs, $(0, 0)$ for 1 pair, $(1, 0)$ for 3 pairs, and $(0, 1)$ for 1 pair.

a. Cross-classify the 8 pairs in terms of diet (1 or 0) for the case against diet (1 or 0) for the control. Call this Table A. Display the $2 \times 2 \times 8$ table with eight partial tables relating diet (1 or 0) to response (case or control) for the 8 pairs. Call this Table B.

b. Calculate the McNemar z^2 for Table A and the CMH statistic for Table B. Compare.

c. Show that the Mantel–Haenszel estimate of a common odds ratio for Table B is identical to n_{12}/n_{21} for Table A.

d. For Table B with pairs deleted in which the case and the control had the same diet, show that the CMH statistic and the Mantel–Haenszel odds ratio estimate do not change.

e. This sample size is small for large-sample tests. Use the binomial distribution with Table A to find the exact P-value for testing marginal homogeneity against the alternative hypothesis of a higher incidence of colon cancer for the high-red-meat diet.

10.5 Each week *Variety* magazine summarizes reviews of new movies by critics in several cities. Each review is categorized as pro, con, or mixed, according to whether the overall evaluation is positive, negative, or a mixture of the two. Table 10.16 summarizes the ratings from

TABLE 10.16 Data for Problem 10.5

Siskel	Ebert		
	Con	Mixed	Pro
Con	24	8	13
Mixed	8	13	11
Pro	10	9	64

Source: A. Agresti and L. Winner, *CHANCE* **10**: 10–14 (1997), reprinted with permission, copyright 1997 by the American Statistical Association.

April 1995 through September 1996 for Chicago film critics Gene Siskel and Roger Ebert.

a. Fit the symmetry model, quasi-independence model, and quasi-symmetry model. Interpret.

b. Test marginal homogeneity using models, and interpret.

c. Analyze these data using agreement models and/or measures of agreement.

10.6 Refer to Table 10.5. Fit the ordinal quasi-symmetry model using $u_1 = 1$ and $u_4 = 4$ and picking u_2 and u_3 that are unequally spaced but represent sensible choices. Compare results and interpretations to those in Sections 10.3.2 and 10.4.7.

10.7 Refer to all four items in Table 8.19.

a. Fit the complete symmetry and quasi-symmetry models. Test marginal homogeneity. Interpret.

b. Fit the ordinal quasi-symmetry model. Test marginal homogeneity. Interpret the effects.

10.8 Table 10.17 shows subjects' purchase choice of instant decaffeinated coffee at two times.

a. Fit the symmetry model and use residuals to analyze changes.

b. Test marginal homogeneity. Show that the small P-value reflects a decrease in the proportion choosing High Point and an increase in the proportion choosing Sanka, with no evidence of change for the other coffees.

c. Show that quasi-independence has $G^2 = 13.8$ (df = 11). Interpret, and suggest other analyses that might be useful.

TABLE 10.17 Data for Problem 10.8

	Second Purchase				
First Purchase	High Point	Taster's Choice	Sanka	Nescafe	Brim
High Point	93	17	44	7	10
Taster's Choice	9	46	11	0	9
Sanka	17	11	155	9	12
Nescafe	6	4	9	15	2
Brim	10	4	12	2	27

Source: Based on data from R. Grover and V. Srinivasan, *J. Market. Res.* **24**: 139–153 (1987). Reprinted with permission from the American Marketing Association.

TABLE 10.18 Data for Problem 10.9

Father's Status	Son's Status					Total
	1	2	3	4	5	
1	50	45	8	18	8	129
2	28	174	84	154	55	495
3	11	78	110	223	96	518
4	14	150	185	714	447	1510
5	3	42	72	320	411	848
Total	106	489	459	1429	1017	3500

Source: Reprinted with permission from D. V. Glass (ed), *Social Mobility in Britain*, Glencoe, IL: Free Press (1954).

10.9 Table 10.18 relates father's and son's occupational status for a British sample. Analyze these data, using models of **(a)** symmetry, **(b)** quasi-symmetry, **(c)** ordinal quasi-symmetry, **(d)** conditional symmetry, **(e)** marginal homogeneity, **(f)** quasi-independence, and **(g)** quasi-uniform association. Interpret using their fit and lack of fit.

10.10 For Table 10.18, use kappa to describe agreement. Interpret.

10.11 Table 10.19 displays multiple sclerosis diagnoses for two neurologists who classified patients in two sites, Winnipeg and New Orleans. The diagnostic classes are (1) certain; (2) probable; (3) possible; and (4) doubtful, unlikely, or definitely not. For the New Orleans patients, study the agreement using **(a)** the independence model and residuals, **(b)** more complex models, and **(c)** kappa. Interpret each.

TABLE 10.19 Data for Problem 10.11

New Orleans Neurologist	Winnipeg Neurologist							
	Winnipeg Patients				New Orleans Patients			
	1	2	3	4	1	2	3	4
1	38	5	0	1	5	3	0	0
2	33	11	3	0	3	11	4	0
3	10	14	5	6	2	13	3	4
4	3	7	3	10	1	2	4	14

Source: J. R. Landis and G. G. Koch, *Biometrics* **33**: 159–174 (1977). Reprinted with permission from the Biometric Society.

10.12 For Problem 10.11, construct a model that describes agreement between neurologists for the two sites simultaneously.

10.13 Calculate kappa for a 4×4 table having $n_{ii} = 5$ all i, $n_{i,\,i+1} = 15$, $i = 1, 2, 3$, $n_{41} = 15$, and $n_{ij} = 0$ otherwise. Explain why strong association does not imply strong agreement.

10.14 Refer to Table 10.8. Based on the reported standardized residuals, explain why the linear-by-linear association model (9.6) might fit well. Fit it and describe the association.

10.15 In 1990, a sample of psychology graduate students at the University of Florida made blind, pairwise preference tests of three cola drinks. For 49 comparisons of Coke and Pepsi, Coke was preferred 29 times. For 47 comparisons of Classic Coke and Pepsi, Classic Coke was preferred 19 times. For 50 comparisons of Coke and Classic Coke, Coke was preferred 31 times. Comparisons resulting in ties are not reported.

a. Fit the Bradley–Terry model, analyze the quality of fit, and rank the drinks. Is there sufficient evidence to conclude a preference for one drink?

b. Estimate the probability that Coke is preferred to Pepsi, using the model, and compare to the sample proportion.

10.16 Table 10.20 refers to journal citations among four statistics journals during 1987–1989. The more often articles in a particular journal are cited, the more prestige that journal accrues. For citations involving pair A and B, view it as a victory for A if it is cited by B and a defeat for A if it cites B. Fit the Bradley–Terry model. Interpret the fit, and give a prestige ranking of the journals. For citations involving *Commun. Stat.* and *JRSS-B*, estimate the probability that the *Commun. Stat.* article cites the *JRSS-B* article.

TABLE 10.20 Data for Problem 10.16

	Cited Journal			
Citing Journal	*Biometrika*	*Commun. Stat.*	*JASA*	*JRSS-B*
Biometrika	714	33	320	284
Commun. Stat.	730	425	813	276
JASA	498	68	1072	325
JRSS-B	221	17	142	188

Source: Stigler (1994). Reprinted with permission from the Institute of Mathematical Statistics.

TABLE 10.21 Data for Problem 10.17

	Loser				
Winner	Seles	Graf	Sabatini	Navratilova	Sanchez
Seles	—	2	1	3	2
Graf	3	—	6	3	7
Sabatini	0	3	—	1	3
Navratilova	3	0	2	—	3
Sanchez	0	1	2	1	—

10.17 Table 10.21 refers to matches for several women tennis players during 1989 and 1990.

a. Fit the Bradley–Terry model. Interpret, and rank the players.

b. Estimate the probability of Seles beating Graf. Compare the model estimate to the sample proportion. Construct a 90% confidence interval for the probability.

c. Which pairs of players are significantly different according to a 80% simultaneous Bonferroni comparison?

10.18 Refer to Problem 3.3 on basketball free-throw shooting. Analyze these data.

10.19 Refer to Table 2.12 and Problem 2.19. Using models, describe the relationship between husband's and wife's sexual fun.

10.20 Refer to Table 8.19. The two-way table relating responses for the environment (as rows) and cities (as columns) has cell counts, by row, $(108, 179, 157 \,/\, 21, 55, 52 \,/\, 5, 6, 24)$. Analyze these data.

Theory and Methods

10.21 Explain the following analogy: McNemar's test is to binary data as the paired difference t test is to normally distributed data.

10.22 For a 2×2 table, derive $\text{cov}(p_{+1}, p_{1+})$, and show that $\text{var}[\sqrt{n}\,(p_{+1} - p_{1+})]$ equals (10.1).

10.23 Refer to the subject-specific model (10.8) for binary matched pairs.

a. Show that $\exp(\beta)$ is a conditional odds ratio between observation and outcome. Explain the distinction between it and the odds ratio $\exp(\beta)$ for model (10.6).

b. Using the conditional distribution (10.9), show that $\hat{\beta} = \log(n_{21}/n_{12})$.

c. For a random sample of n pairs, explain why

$$E(n_{21}/n) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+\exp(\alpha_i)} \frac{\exp(\alpha_i+\beta)}{1+\exp(\alpha_i+\beta)}.$$

Similarily, state $E(n_{12}/n)$. Using their ratio for fixed n and as $n \to \infty$, explain why $n_{21}/n_{12} \xrightarrow{p} \exp(\beta)$. (*Hint:* Apply the law of large numbers due to A. A. Markov for independent but not identically distributed random variables, or use Chebyshev's inequality.)

d. Show that the Mantel–Haenszel estimator (6.7) of a common odds ratio in the $2 \times 2 \times n$ form of the data simplifies to $\exp(\hat{\beta}) = n_{21}/n_{12}$.

e. Use the delta method to show (10.10) for the SE of $\hat{\beta}$.

f. For a table of the form shown in Table 10.2, show that the CMH statistic (6.6) is algebraically identical to the McNemar statistic $(n_{21} - n_{12})^2/(n_{21} + n_{12})$ for tables of Table 10.1 type.

10.24 Refer to Problem 10.23. Unlike the conditional ML estimator of β, the unconditional ML estimator is inconsistent (Andersen 1980, pp. 244–245; first shown by him in 1973). Show this as follows:

a. Assuming independence of responses for different subjects and different observations by the same subject, find the log likelihood. Show that the likelihood equations are $y_{+t} = \sum_i P(Y_{it} = 1)$ and $y_{i+} = \sum_t P(Y_{it} = 1)$.

b. Substituting $\exp(\alpha_i)/[1 + \exp(\alpha_i)] + \exp(\alpha_i + \beta)/[1 + \exp(\alpha_i + \beta)]$ in the second likelihood equation, show that $\hat{\alpha}_i = -\infty$ for the n_{22} subjects with $y_{i+} = 0$, $\hat{\alpha}_i = \infty$ for the n_{11} subjects with $y_{i+} = 2$, and $\hat{\alpha}_i = -\hat{\beta}/2$ for the $n_{21} + n_{12}$ subjects with $y_{i+} = 1$.

c. By breaking $\sum_i P(Y_{it} = 1)$ into components for the sets of subjects having $y_{i+} = 0$, $y_{i+} = 2$, and $y_{i+} = 1$, show that the first likelihood equation is, for $t = 1$, $y_{+1} = n_{22}(0) + n_{11}(1) + (n_{21} + n_{12})\exp(-\hat{\beta}/2)/[1 + \exp(-\hat{\beta}/2)]$. Explain why $y_{+1} = n_{11} + n_{12}$, and solve the first likelihood equation to show that $\hat{\beta} = 2\log(n_{21}/n_{12})$. Hence, as a result of Problem 10.23, $\hat{\beta} \xrightarrow{p} 2\beta$.

10.25 Consider marginal model (10.6) when Y_1 and Y_2 are independent and conditional model (10.8) when $\{\alpha_i\}$ are identical. Explain why they are equivalent.

10.26 Let $\hat{\beta}_M = \log(p_{+1}p_{2+}/p_{+2}p_{1+})$ refer to marginal model (10.6) and $\hat{\beta}_C = \log(n_{21}/n_{12})$ to conditional model (10.8). Using the delta method, show that the asymptotic variance of $\sqrt{n}(\hat{\beta}_M - \beta_M)$ is

$$(\pi_{1+}\pi_{2+})^{-1} + (\pi_{+1}\pi_{+2})^{-1} - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})/(\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2}).$$

Under the independence condition of the previous problem, $\beta_M = \beta_C$. In that case, show that the asymptotic variances satisfy

$$\begin{aligned}\text{var}\left[\sqrt{n}\left(\hat{\beta}_M\right)\right] &= (\pi_{1+}\pi_{2+})^{-1} + (\pi_{+1}\pi_{+2})^{-1} \\ &\leq (\pi_{1+}\pi_{+2})^{-1} + (\pi_{+1}\pi_{2+})^{-1} \\ &= \pi_{12}^{-1} + \pi_{21}^{-1} = \text{var}\left[\sqrt{n}\left(\hat{\beta}_C\right)\right]\end{aligned}$$

10.27 Refer to model (10.12) for a matched-pairs study. For the conditional ML approach, show that the conditional distribution satisfies (10.13) and does not depend on $\boldsymbol{\beta}$ when $S_i = 0$ or 2. Show what happens to β_j in the conditional distribution for a predictor for which $x_{ji1} = x_{ji2}$ all i.

10.28 Consider model (10.12) for a study with matched sets of T observations rather than matched pairs. Explain how (10.13) generalizes and construct the form of the conditional likelihood.

10.29 Give an example illustrating that when $I > 2$, marginal homogeneity does not imply symmetry.

10.30 Derive the likelihood equations and residual df for (**a**) symmetry, (**b**) quasi-symmetry, (**c**) quasi-independence, and (**d**) ordinal quasi-symmetry.

10.31 For the quasi-symmetry model (10.19), let $\lambda_a = \lambda_a^X - \lambda_a^Y$. Show that one can express it equivalently as $\log \mu_{ab} = \lambda + \lambda_a + \lambda_{ab}^*$, with $\lambda_{ab}^* = \lambda_{ba}^*$. Hence, one needs only one set of main-effect parameters.

10.32 Show that quasi-symmetry is equivalent (Caussinus 1966) to

$$(\pi_{ab}\pi_{bc}\pi_{ca})/(\pi_{ba}\pi_{cb}\pi_{ac}) = 1 \quad \text{all } a, b, \text{ and } c.$$

10.33 Derive the covariance matrix (10.16) for the difference vector $\mathbf{d}$.

10.34 Construct the loglinear model satisfying both marginal homogeneity and statistical independence. Show that $\hat{\pi}_{ab} = (p_{+a} + p_{a+})(p_{+b} + p_{b+})/4$ and residual df $= I(I-1)$.

10.35 Consider the conditional symmetry (CS) model (10.28).

a. Show that it has the loglinear representation

$$\log \mu_{ab} = \lambda_{\min(a,b),\,\max(a,b)} + \tau I(a < b),$$

where $I(\cdot)$ is an indicator (see also Bishop et al. 1975, pp. 285–286).

b. Show that the likelihood equations are

$$\hat{\mu}_{ab} + \hat{\mu}_{ba} = n_{ab} + n_{ba} \quad \text{for all } a \le b, \quad \sum\sum_{a<b} \hat{\mu}_{ab} = \sum\sum_{a<b} n_{ab}.$$

c. Show that $\hat{\tau} = \log[(\Sigma\Sigma_{a<b} n_{ab})/(\Sigma\Sigma_{a>b} n_{ab})]$, $\hat{\mu}_{aa} = n_{aa}$, $a = 1, \ldots, I$, $\hat{\mu}_{ab} = \exp[\hat{\tau} I(a<b)](n_{ab} + n_{ba})/[\exp(\hat{\tau}) + 1]$ for $a \neq b$.

d. Show that the estimated asymptotic variance of $\hat{\tau}$ is

$$\left(\sum\sum_{a<b} n_{ab}\right)^{-1} + \left(\sum\sum_{a>b} n_{ab}\right)^{-1}.$$

e. Show that residual df $= (I+1)(I-2)/2$.

f. Show that conditional symmetry + marginal homogeneity = symmetry. Explain why $G^2(\mathrm{S} \mid \mathrm{CS})$ tests marginal homogeneity (df = 1). When the model holds $G^2(\mathrm{S} \mid \mathrm{CS})$ is more powerful asymptotically than $G^2(\mathrm{S} \mid \mathrm{QS})$. Why?

10.36 Identify loglinear models that correspond to the logit models, for $a < b$, $\log(\pi_{ab}/\pi_{ba}) =$ **(a)** 0, **(b)** τ, **(c)** $\alpha_a - \alpha_b$, and **(d)** $\beta(b-a)$.

10.37 A nonmodel-based ordinal measure of marginal heterogeneity is

$$\hat{\Delta} = \sum\sum_{a<b} p_{a+}p_{+b} - \sum\sum_{a>b} p_{a+}p_{+b}.$$

Show that $\hat{\Delta}$ estimates $\Delta = P(Y_1 > Y_2) - P(Y_2 > Y_1)$, where Y_1 has distribution $\{\pi_{a+}\}$ and Y_2 is independent from $\{\pi_{+b}\}$. Show that marginal homogeneity implies that $\Delta = 0$. Show that the estimated

asymptotic variance of $\hat{\Delta}$ is

$$\left[\sum_a \sum_b \hat{\phi}_{ab}^2 p_{ab} - \left(\sum_a \sum_b \hat{\phi}_{ab} p_{ab}\right)^2\right] \Big/ n,$$

where $\hat{\phi}_{ab} = \hat{F}_{b1} + \hat{F}_{b-1,1} - \hat{F}_{a2} - \hat{F}_{a-1,2}$ with $\hat{F}_{a1} = (p_{1+} + \cdots + p_{a+})$ and $\hat{F}_{a2} = (p_{+1} + \cdots + p_{+a})$ (Agresti 1984, pp. 208–209).

10.38 For ordered scores $\{u_a\}$, let $\bar{y}_1 = \sum_a u_a p_{a+}$ and $\bar{y}_2 = \sum_a u_a p_{+a}$. Show that marginal homogeneity implies that $E(\bar{Y}_1) = E(\bar{Y}_2)$ and

$$\left[\sum_a \sum_b (u_a - u_b)^2 p_{ab} - (\bar{y}_1 - \bar{y}_2)^2\right] \Big/ n.$$

estimates $\text{var}(\bar{Y}_1 - \bar{Y}_2)$. Construct a test of marginal homogeneity (Bhapkar 1968).

10.39 Consider the multiplicative model for a square table,

$$\pi_{ab} = \begin{cases} \alpha_a \alpha_b (1 - \beta), & a \neq b \\ \alpha_a^2 + \beta \alpha_a (1 - \alpha_a), & a = b. \end{cases}$$

a. Show that the model satisfies (**i**) symmetry, (**ii**) marginal homogeneity, (**iii**) quasi-symmetry, (**iv**) quasi-independence.

b. Show that $\alpha_a = \pi_{a+} = \pi_{+a}$, $a = 1, \ldots, I$.

c. Show that $\beta =$ Cohen's kappa, and interpret $\kappa = 0$ and $\kappa = 1$ for this model.

10.40 A 2×2 table has a true odds ratio of 10. Find the cell probabilities for which (**a**) $\pi_{1+} = \pi_{+1} = 0.5$, (**b**) $\pi_{1+} = \pi_{+1} = 0.3$, and (**c**) $\pi_{1+} = \pi_{+1} = 0.1$. Find the value of kappa for each. (This shows that for a given association, kappa depends strongly on the marginal probabilities; see also Sprott 2000, p. 59.)

10.41 A model for agreement on an ordinal response partitions beyond-chance agreement into that due to a baseline association and a main-diagonal increment (A. Agresti, *Biometrics* **44**: 539–548, 1988). For ordered scores $\{u_a\}$, the model is

$$\log \mu_{ab} = \lambda + \lambda_a^A + \lambda_b^B + \beta u_a u_b + \delta I(a = b). \quad (10.35)$$

a. Show that this is a special case of quasi-symmetry and of quasi-association (10.29).

b. For agreement odds (10.30), show that $\log \tau_{ab} = (u_b - u_a)^2\beta + 2\delta$. For unit-spaced scores, show the local odds ratios have $\log \theta_{ab} = \beta$ when none of the four cells falls on the main diagonal.

c. Find the likelihood equations and show that $\{\hat{\mu}_{ab}\}$ and $\{n_{ab}\}$ share the same marginal distributions, correlation, and prevalence of exact agreement.

d. For Table 10.8 using $\{u_a = a\}$, show that (10.35) has $G^2 = 4.8$ (df = 7), with $\hat{\delta} = 0.842$ (SE = 0.427) and $\hat{\beta} = 1.316$ (SE = 0.420). Interpret using $\hat{\tau}_{a,a+1}$ and $\hat{\theta}_{ab}$ for $|a - b| > 1$.

10.42 Refer to the Bradley–Terry model.

a. Show that $\log(\Pi_{ac}/\Pi_{ca}) = \log(\Pi_{ab}/\Pi_{ba}) + \log(\Pi_{bc}/\Pi_{cb})$.

b. With this model, is it possible that a could be preferred to b (i.e., $\Pi_{ab} > \Pi_{ba}$) and b could be preferred to c, yet c could be preferred to a? Explain.

c. Explain why $\{\beta_a\}$ are not identifiable without a constraint such as $\beta_I = 0$. (*Hint:* Show the model holds when $\{\beta_a^* = \beta_a - c\}$ for any c.)

10.43 Refer to model (10.32).

a. Construct a more general model having home-team parameters $\{\beta_{Hi}\}$ and away-team parameters $\{\beta_{Ai}\}$, such that the probability team i beats team j when i is the home team is $\exp(\beta_{Hi})/[\exp(\beta_{Hi}) + \exp(\beta_{Aj})]$, where $\beta_{AI} = 0$ but β_{Hi} is unrestricted.

b. Interpret the case $\{\beta_{Hi} = \beta_{Ai} + c\}$, when **(i)** $c = 0$, and **(ii)** $c > 0$.

c. Fit the model to Table 10.12. Compare the fit to model (10.32). Compare $\{\hat{\beta}_{Hi}\}$ and $\{\hat{\beta}_{Ai}\}$ to describe how teams play at home and away.

10.44 Find the log likelihood for the Bradley–Terry model. From the kernel, show that (given $\{N_{ab}\}$) the minimal sufficient statistics are $\{n_{a+}\}$. Thus, explain how "victory totals" determine the estimated ranking.

10.45 Explain how to fit the complete symmetry model in T dimensions.

10.46 Prove that if kth-order marginal symmetry holds, jth-order marginal symmetry holds for any $j < k$.

10.47 Suppose that quasi-symmetry holds for an I^T table. When the table is collapsed over a variable, show that the model holds for the I^{T-1} table with the same main effects.

Categorical Data Analysis, Second Edition. Alan Agresti

ISBN: 0-471-36093-7

CHAPTER 11

Analyzing Repeated Categorical Response Data

Many studies observe the response variable for each subject repeatedly, at several times or under various conditions. Repeated categorical response data occur commonly in health-related applications, especially in longitudinal studies. For example, a physician might evaluate patients at weekly intervals regarding whether a new drug treatment is successful. In some cases explanatory variables may also vary over time. But the repeated responses need not refer to different times. A dental study might measure whether there is decay for each tooth in a subject's mouth.

Often, the responses refer to matched sets, or *clusters*, of subjects. An example is a (survival, nonsurvival) response for each fetus in a litter, for a sample of pregnant mice exposed to various dosages of a toxin. A multistage sample to study factors affecting obesity in children may regard children from the same family as a cluster. Observations within a cluster tend to be more alike than observations from different clusters. Ordinary analyses that ignore this may be badly inappropriate.

In this chapter we generalize methods of Chapter 10, which referred to matched pairs. In Section 11.1 we compare marginal distributions in T-way tables. The remaining sections extend models to include explanatory variables. For instance, many studies compare the repeated measurements for different groups or treatments. In Section 11.2 we use ML methods for fitting marginal models. In Section 11.3 we use *generalized estimating equations* (GEE), a multivariate version of quasi-likelihood that is computationally simpler than ML. Section 11.4 covers technical details about the GEE approach. In the final section we introduce a *transitional* approach that models observations in terms of previous outcomes.

11.1 COMPARING MARGINAL DISTRIBUTIONS: MULTIPLE RESPONSES

Usually, the multivariate dependence among repeated responses is of less interest than their marginal distributions. For instance, in treating a chronic condition (such as a phobia) with some treatment, the primary goal might be to study whether the probability of success increases over the T weeks of a treatment period. The T success probabilities refer to the T first-order marginal distributions. In Sections 10.2.1 and 10.3 we compared marginal distributions for matched pairs ($T = 2$) using models that apply directly to the marginal distributions. In this section we extend this approach to $T > 2$.

11.1.1 Binary Marginal Models and Marginal Homogeneity

Denote T binary responses by $(Y_1, Y_2, \ldots, Y_T)$. The marginal logit model (10.6) for matched pairs extends to

$$\text{logit}\big[P(Y_t = 1)\big] = \alpha + \beta_t, \qquad t = 1, \ldots, T, \tag{11.1}$$

with a constraint such as $\beta_T = 0$ or $\alpha = 0$. For a possible sequence of outcomes $\mathbf{i} = (i_1, i_2, \ldots, i_T)$ where each $i_t = 0$ or 1, let

$$\pi_{\mathbf{i}} = P(Y_1 = i_1, Y_2 = i_2, \ldots, Y_T = i_T).$$

Let $\boldsymbol{\pi}$ denote the vector of these probabilities for the possible $\mathbf{i}$. They refer to a 2^T table that cross-classifies the T responses and describes the joint distribution of $(Y_1, \ldots, Y_T)$. The sample cell proportions are the ML estimates of $\boldsymbol{\pi}$, and the sample proportion with $y_t = 1$ is the ML estimate of $P(Y_t = 1)$.

Model (11.1) is saturated, describing T marginal probabilities by T parameters. Marginal homogeneity, for which $P(Y_1 = 1) = \cdots = P(Y_T = 1)$, is the special case $\beta_1 = \cdots = \beta_T$. Even though this case has only one parameter, ML fitting is not simple. The multinomial likelihood refers to the 2^T joint cell probabilities $\boldsymbol{\pi}$ rather than the T marginal probabilities $\{P(Y_t = 1)\}$. Fitting methods are described in Section 11.2.5.

Let $n_{\mathbf{i}}$ denote the sample cell count in cell $\mathbf{i}$. The kernel of the log likelihood $L(\boldsymbol{\pi})$ is $\sum_{\mathbf{i}} n_{\mathbf{i}} \log \pi_{\mathbf{i}}$. Let $L(\mathbf{p})$ denote the log likelihood evaluated at the sample proportions $\{p_{\mathbf{i}} = n_{\mathbf{i}}/n\}$, the ML fit of model (11.1). Let $L(\hat{\boldsymbol{\pi}}^{MH})$ denote the maximized log likelihood assuming marginal homogeneity. The likelihood-ratio test of marginal homogeneity (Lipsitz et al. 1990; Madansky 1963) uses

$$-2\big[L(\hat{\boldsymbol{\pi}}^{MH}) - L(\mathbf{p})\big] = 2\sum_{\mathbf{i}} n_{\mathbf{i}} \log\big(p_{\mathbf{i}}/\hat{\pi}_{\mathbf{i}}^{MH}\big). \tag{11.2}$$

TABLE 11.1 Responses to Three Drugs in a Crossover Study

	Drug A Favorable		Drug A Unfavorable	
	B Favorable	B Unfavorable	B Favorable	B Unfavorable
C Favorable	6	2	2	6
C Unfavorable	16	4	4	6

Source: Reprinted with permission from the Biometric Society (Grizzle et al. 1969).

The asymptotic null chi-squared distribution has df $= T - 1$, since the general model (11.1) has $T - 1$ more parameters than marginal homogeneity.

11.1.2 Crossover Drug Comparison Example

Table 11.1 comes from a crossover study in which each subject used each of three drugs for treatment of a chronic condition at three times. The response measured the reaction as favorable or unfavorable. The 2^3 table gives the (favorable, unfavorable) classification for reaction to drug A in the first dimension, drug B in the second, and drug C in the third. We assume that the drugs have no carryover effects and that the severity of the condition remained stable for each subject throughout the experiment. These assumptions are reasonable for many chronic conditions, such as migraine headache.

The sample proportion favorable was $(0.61, 0.61, 0.35)$ for drugs (A, B, C). The likelihood-ratio statistic for testing marginal homogeneity is 5.95 (df $= 2$), for a P-value of 0.05. For simultaneous confidence intervals comparing pairs of treatments with overall error probability no greater than 0.05, the Bonferroni method uses confidence coefficient $(1 - 0.05/3) = 0.9833$ for each. For instance, from formula (10.1), the estimate $0.261 = 0.609 - 0.348$ of the difference between drugs A and C has an estimated standard error of 0.108. The confidence interval for the true difference is $0.261 \pm 2.39(0.108)$, or $(0.002, 0.520)$. The same interval holds for comparison of drugs B and C. There is some evidence that the proportion of favorable responses is lower for drug C.

The sample size is not large, however, so we view these results with caution. For each pair of drugs, a 2×2 table relates the two responses. An exact binomial test (Section 10.4.1) uses its off-diagonal counts. These yield P-values of 1.0 for comparing drugs A and B and 0.036 for comparing A with C and for comparing B with C.

11.1.3 Modeling Margins of a Multicategory Response

The binary marginal model (11.1) extends to multinomial responses. With baseline-category logits for I outcome categories, the saturated model is

$$\log\big[P(Y_t = j)/P(Y_t = I)\big] = \beta_{tj}, \qquad t = 1, \ldots, T, \quad j = 1, \ldots, I - 1. \tag{11.3}$$

Marginal homogeneity, whereby $P(Y_1 = j) = \cdots = P(Y_T = j)$ for $j = 1, \ldots, I-1$, is the special case in which

$$\beta_{1j} = \beta_{2j} = \cdots = \beta_{Tj}, \qquad j = 1, \ldots, I-1.$$

The likelihood-ratio test of marginal homogeneity comparing the two models has form (11.2) and df $= (T-1)(I-1)$.

For an ordinal response, an unsaturated model that is more complex than marginal homogeneity focuses on shifts up and down in the T margins. One such model is

$$\text{logit}\big[P(Y_t \le j)\big] = \alpha_j + \beta_t, \qquad t = 1, \ldots, T, \quad j = 1, \ldots, I-1, \quad (11.4)$$

with constraint such as $\beta_T = 0$. Marginal homogeneity is the special case $\beta_1 = \cdots = \beta_T$. Its test has df $= T-1$. The $\{\alpha_j\}$ satisfy $\alpha_1 < \ldots < \alpha_{I-1}$ because of the ordering of the cumulative probabilities. These models can be fitted using ML methodology presented in Section 11.2.5.

11.1.4 Wald and Generalized CMH Score Tests of Marginal Homogeneity

In this chapter we focus on modeling the marginal distributions rather than merely testing marginal homogeneity. However, a variety of tests are available besides the likelihood ratio, so we briefly summarize a couple of them.

Let $p_j(t)$ denote the sample proportion in category j for response Y_t, let

$$\bar{p}_j = \sum_t p_j(t)/T, \qquad d_j(t) = p_j(t) - \bar{p}_j,$$

and let $\mathbf{d}$ denote the vector of $\{d_j(t),\ t = 1, \ldots, T-1,\ j = 1, \ldots, I-1\}$. Let $\hat{\mathbf{V}}$ denote the estimated covariance matrix of $\sqrt{n}\,\mathbf{d}$. Bhapkar (1973) proposed the Wald statistic

$$W = n\mathbf{d}'\hat{\mathbf{V}}^{-1}\mathbf{d}. \quad (11.5)$$

for the general alternative. This generalizes (10.16) and has a large-sample chi-squared distribution with df $= (I-1)(T-1)$.

Other statistics are special cases of the generalized Cochran–Mantel–Haenszel (CMH) statistic (Section 7.5.3). Recall that for the binary case ($I = 2$) with matched pairs ($T = 2$), the CMH statistic applies to a three-way table (see, e.g., Table 10.2) in which each stratum shows the two outcomes for a given subject. A generalization of Table 10.2 provides n strata of $T \times I$ tables. The kth stratum gives the T outcomes for subject k. Row t in a stratum has a 1 in the column that is the outcome for observation t, and 0 in all other columns (or 0 in every column if that observation is missing). Probability distributions for the subject-stratified setup naturally relate to

subject-specific models such as logit model (10.8), rather than to marginal models. However, conditional independence in this three-way table (given subject) corresponds to an exchangeability among variables in the I^T table that implies marginal homogeneity. A generalized CMH test of conditional independence in the $T \times I \times n$ table also tests marginal homogeneity using a sampling distribution generated under the stronger exchangeability condition (Darroch 1981). For an ordinal response with fixed scores, the generalized CMH statistic for detecting variability among T means is appropriate.

When $I = 2$ and $T = 2$, this CMH approach is equivalent to McNemar's statistic. When $I = 2$ but $T > 2$, the generalized CMH statistic treating the T responses as unordered is identical to a statistic Cochran (1950) proposed. His statistic, called *Cochran's Q*, has df $= T - 1$ (Problem 11.22).

11.2 MARGINAL MODELING: MAXIMUM LIKELIHOOD APPROACH

Analyses above compared marginal distributions, but without accounting for explanatory variables. We now include such predictors. In this section we use ML, but we defer model fitting details to the end of the section.

11.2.1 Longitudinal Mental Depression Example

We use Table 11.2 to illustrate a variety of analyses in this and the next chapter. It refers to a longitudinal study comparing a new drug with a standard drug for treatment of subjects suffering mental depression (Koch et al. 1977). Subjects were classified into two initial diagnosis groups according to whether severity of depression was mild or severe. In each group, subjects were randomly assigned to one of the two drugs. Following 1 week, 2 weeks, and 4 weeks of treatment, each subject's suffering from mental depression was classified as normal or abnormal.

TABLE 11.2 Cross-Classification of Responses on Depression at Three Times by Diagnosis and Treatment

		Response at Three Times[a]							
Diagnosis	Treatment	NNN	NNA	NAN	NAA	ANN	ANA	AAN	AAA
Mild	Standard	16	13	9	3	14	4	15	6
	New drug	31	0	6	0	22	2	9	0
Severe	Standard	2	2	8	9	9	15	27	28
	New drug	7	2	5	2	31	5	32	6

[a] N, normal; A, abnormal.

Source: Reprinted with permission from the Biometric Society (Koch et al. 1977).

Table 11.2 shows four groups, the combinations of categories of the two explanatory variables: treatment type and severity of initial diagnosis. Since the study observed the binary response (depression assessment) at $T = 3$ occasions, Table 11.2 shows a 2^3 table for each group. The three depression assessments form a multivariate response variable with three components, with $Y_t = 1$ for normal and 0 for abnormal. The 12 marginal distributions result from three repeated observations for each of the four groups.

Let s denote the severity of the initial diagnosis, with $s = 1$ for severe and $s = 0$ for mild. Let d denote the drug, with $d = 1$ for new and $d = 0$ for standard. Let t denote the time of measurement. Koch et al. (1977) noted that if the time metric reflects cumulative drug dosage, a logit scale often has a linear effect for the logarithm of time. They used scores (0, 1, 2), the logs to base 2 of the week numbers (1, 2, and 4), for time.

Table 11.3 shows sample proportions of normal responses (i.e., $y_t = 1$) for the 12 marginal distributions. For instance, from Table 11.2, the sample proportion of normal responses after week 1 for subjects with mild initial diagnosis using the standard drug was $(16 + 13 + 9 + 3)/(16 + 13 + 9 + 3 + 14 + 4 + 15 + 6) = 0.51$. The sample proportion of normal responses (1) increased over time for each group; (2) increased at a faster rate for the new drug than the standard, for each fixed initial diagnosis; and (3) was higher for the mild than the severe initial diagnosis, for each treatment at each occasion. In such a study the company that developed the new drug would hope to show that patients have a significantly higher rate of improvement with it.

The marginal logit model

$$\text{logit}\big[P(Y_t = 1)\big] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t$$

has the main effects of the explanatory variables (severity of initial diagnosis and drug) and of the variable (time) that specifies the different components of the multivariate response. Its linear time effect β_3 is the same for each group.

The natural sampling assumption is multinomial for the eight cells in the 2^3 cross-classification of the three responses, independently for the four

TABLE 11.3 Sample Marginal Proportions of Normal Response for Depression Data of Table 11.2

		Sample Proportion		
Diagnosis	Treatment	Week 1	Week 2	Week 4
Mild	Standard	0.51	0.59	0.68
	New drug	0.53	0.79	0.97
Severe	Standard	0.21	0.28	0.46
	New drug	0.18	0.50	0.83

groups. However, the model refers to 12 marginal probabilities (for 2 drug treatments $\times$ 2 initial severity diagnoses $\times$ 3 time points) rather than the $4 \times 2^3 = 32$ cell probabilities in the product multinomial likelihood function. The three marginal binomial variates for each group are dependent. ML estimation requires an iterative routine for maximizing the product multinomial likelihood, subject to the constraint that the marginal probabilities satisfy the model. An algorithm for this is given in Section 11.2.5.

A check of model fit compares the 32 cell counts in Table 11.2 to their ML fitted values. Since the model describes 12 marginal logits using four parameters, residual df = 8. The deviance $G^2 = 34.6$. The poor fit is not surprising. The model assumes a common rate of improvement β_3, but the sample shows a higher rate for the new drug.

A more realistic model permits the time effect to differ by drug,

$$\text{logit}\big[P(Y_t = 1)\big] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 dt.$$

Its time effect estimate is $\hat{\beta}_3 = 0.48$ (SE = 0.12) for the standard drug ($d = 0$) and $\hat{\beta}_3 + \hat{\beta}_4 = 1.49$ (SE = 0.14) for the new one ($d = 1$). For the new drug, the slope is $\hat{\beta}_4 = 1.01$ (SE = 0.18) higher than for the standard, giving strong evidence of faster improvement. This model fits much better, with $G^2 = 4.2$ (df = 7). The G^2 decrease of $34.6 - 4.2 = 30.4$ compared to the simpler model is the likelihood-ratio test of H_0: $\beta_4 = 0$, a common time effect for each drug.

The severity of initial diagnosis estimate is $\hat{\beta}_1 = -1.29$ (SE = 0.14); for each drug–time combination, the estimated odds of a normal response when the initial diagnosis was severe equal $\exp(-1.29) = 0.27$ times the estimated odds when the initial diagnosis was mild. The estimate $\hat{\beta}_2 = -0.06$ (SE = 0.22) indicates an insignificant difference between the drugs after 1 week (for which $t = 0$). At time t, the estimated odds of normal response with the new drug are $\exp(-0.06 + 1.01\,t)$ times the estimated odds for the standard drug, for each initial diagnosis level. In summary, severity of initial diagnosis, drug treatment, and time all have substantial effects on the probability of a normal response.

11.2.2 Modeling a Repeated Multinomial Response

Models for marginal distributions of a repeated binary response generalize to multicategory responses. At observation t, the marginal response distribution has $I - 1$ logits. With nominal responses, baseline-category logit models describe the odds of each outcome relative to a baseline. For ordinal responses, one might use cumulative logit models.

For a particular marginal logit, a model has the form

$$\text{logit}_j(t) = \alpha_j + \boldsymbol{\beta}_j' \mathbf{x}_t, \quad j = 1, \ldots, I-1, \quad t = 1, \ldots .$$

For an ordinal response, perhaps $\text{logit}_j(t) = \text{logit}[P(Y_t \le j)]$. Then, $\boldsymbol{\beta}_j$ may simplify to $\boldsymbol{\beta}$, in which case the model takes the proportional odds form with the same effects for each logit. Some parameters in $\boldsymbol{\beta}$ may refer to the variable subscripted by t (e.g., time) that indexes the repeated measurements. One can then compare marginal distributions at particular settings of $\mathbf{x}$ or evaluate effects of $\mathbf{x}$ on the response. In either case, checking for interaction is crucial. For instance, are the effects of $\mathbf{x}$ the same at each t?

11.2.3 Insomnia Example

Table 11.4 shows results of a randomized, double-blind clinical trial comparing an active hypnotic drug with a placebo in patients who have insomnia problems. The response is the patient's reported time (in minutes) to fall asleep after going to bed. Patients responded before and following a two-week treatment period. The two treatments, active and placebo, form a binary explanatory variable. The subjects receiving the two treatments were independent samples.

Table 11.5 displays sample marginal distributions for the four treatment–occasion combinations. From the initial to follow-up occasion, time to falling asleep seems to shift downward for both treatments. The degree of shift seems greater for the active treatment, indicating possible interaction. The response variable is a discrete version of a continuous variable, so by the derivation in Section 7.2.3 a cumulative link model is natural. The proportional odds model

$$\text{logit}\big[P(Y_t \le j)\big] = \alpha_j + \beta_1 t + \beta_2 x + \beta_3 tx \qquad (11.6)$$

permits interaction between t = occasion (0 = initial, 1 = follow-up) and

TABLE 11.4 Time to Falling Asleep, by Treatment and Occasion

	Time to Falling Asleep				
		Follow-up			
Treatment	Initial	< 20	20–30	30–60	> 60
Active	< 20	7	4	1	0
	20–30	11	5	2	2
	30–60	13	23	3	1
	> 60	9	17	13	8
Placebo	< 20	7	4	2	1
	20–30	14	5	1	0
	30–60	6	9	18	2
	> 60	4	11	14	22

Source: From S. F. Francom, C.Chuang-Stein, and J. R. Landis, *Statist. Med.* **8**: 571–582 (1989). Reprinted with permission from John Wiley & Sons Ltd.

TABLE 11.5 Sample Marginal Distributions of Table 11.4

		Response			
Treatment	Occasion	< 20	20–30	30–60	> 60
Active	Initial	0.101	0.168	0.336	0.395
	Follow-up	0.336	0.412	0.160	0.092
Placebo	Initial	0.117	0.167	0.292	0.425
	Follow-up	0.258	0.242	0.292	0.208

x = treatment (0 = placebo, 1 = active), but assumes the same effects for each response cutpoint.

For ML model fitting, $G^2 = 8.0$ (df = 6) for comparing observed to fitted cell counts in modeling the 12 marginal logits using these six parameters. The ML estimates are $\hat{\beta}_1 = 1.074$ (SE = 0.162), $\hat{\beta}_2 = 0.046$ (SE = 0.236), and $\hat{\beta}_3 = 0.662$ (SE = 0.244). This shows evidence of interaction. At the initial observation, the estimated odds that time to falling asleep for the active treatment is below any fixed level equal $\exp(0.046) = 1.04$ times the estimated odds for the placebo treatment; at the follow-up observation, the effect is $\exp(0.046 + 0.662) = 2.03$. In other words, initially the two groups had similar distributions, but at the follow-up those with the active treatment tended to fall asleep more quickly.

For simpler interpretation, it can be helpful to report sample marginal means and their differences. With response scores $\{10, 25, 45, 75\}$ for time to fall asleep, the initial means were 50.0 for the active group and 50.3 for the placebo. The difference in means between the initial and follow-up responses was 22.2 for the active group and 13.0 for the placebo. The difference between these differences of means equals 9.2, with SE = 3.0, indicating that the change was significantly greater for the active group.

11.2.4 Comparisons That Control for Initial Response

For data such as Table 11.4, suppose that the marginal distributions for initial response are identical for the treatment groups. This is true, apart from sampling error, with random assignment of subjects to the groups. Suppose also that conditional on the initial response, the follow-up response distribution is identical for the treatment groups. Then, the follow-up marginal distributions are also identical.

If the initial marginal distributions are not identical, however, the difference between follow-up and initial marginal distributions may differ between treatment groups, even though their conditional distributions for follow-up response are identical. In such cases, although marginal models can be useful, they may not tell the entire story. It may be more informative to construct models that compare the follow-up responses while controlling for the initial response.

Let Y_2 denote the follow-up response, for treatment x with initial response y_1. In the model

$$\text{logit}\big[P(Y_2 \le j)\big] = \alpha_j + \beta_1 x + \beta_2 y_1, \tag{11.7}$$

β_1 compares the follow-up distributions for the treatments, controlling for initial observation. This is an analog of an analysis-of-covariance model, with ordinal rather than continuous response. This cumulative logit model refers to a univariate response (Y_2) rather than marginal distributions of a multivariate response (Y_1, Y_2). It is an example of a *transitional model*, discussed in the final section of this chapter.

11.2.5 ML Fitting of Marginal Logit Models*

ML fitting of marginal logit models is awkward. For T observations on an I-category response, at each setting of predictors the likelihood refers to I^T multinomial joint probabilities, but the model applies to T sets of marginal multinomial parameters $\{P(Y_t = k),\ k = 1, \ldots, I\}$. The marginal multinomial variates are not independent.

Let $\boldsymbol{\pi}$ denote the complete set of multinomial joint probabilities for all settings of predictors. Marginal logit models have the generalized loglinear model form

$$\mathbf{C}\log(\mathbf{A}\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta} \tag{11.8}$$

introduced in Section 8.5.4. In the binary case, the matrix $\mathbf{A}$ applied to $\boldsymbol{\pi}$ forms the T marginal probabilities $\{P(Y_t = 1)\}$ and their complements at each setting of predictors. The matrix $\mathbf{C}$ applied to the log marginal probabilities forms the T marginal logits for each setting; each row of $\mathbf{C}$ has 1 in the position multiplied by the log numerator probability for a given marginal logit, -1 in the position multiplied by the log denominator probability, and 0 elsewhere.

For instance, for the model of marginal homogeneity in a 2^T table with no covariates, $\boldsymbol{\beta}$ is a single parameter, denoted by α in (11.1). For $T = 2$, $\boldsymbol{\pi}$ has four elements, and this model is

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \log \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{21} \\ \pi_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \alpha,$$

which sets both $\text{logit}(\pi_{11} + \pi_{12}) = \text{logit}[P(Y_1 = 1)]$ and $\text{logit}(\pi_{11} + \pi_{21}) = \text{logit}[P(Y_2 = 1)]$ equal to α.

The likelihood function $\ell(\boldsymbol{\pi})$ for a marginal logit model is the product of the multinomial mass functions from the various predictor settings. One

approach for ML fitting views the model as a set of constraints and uses methods for maximizing a function subject to constraints. In model (11.8), let $\mathbf{U}$ denote a full column rank matrix such that the space spanned by the columns of $\mathbf{U}$ is the orthogonal complement of the space spanned by the columns of $\mathbf{X}$. Then, $\mathbf{U}'\mathbf{X} = \mathbf{0}$, and the model has the equivalent constraint form

$$\mathbf{U}'\mathbf{C}\log(\mathbf{A}\boldsymbol{\pi}) = \mathbf{0}.$$

For instance, for marginal homogeneity in a 2×2 table with (11.8) as expressed above, $\mathbf{U}' = (1, -1)$. Then $\mathbf{U}'$ applied to $\mathbf{C}\log(\mathbf{A}\boldsymbol{\pi})$ sets the difference between the row and column marginal logits equal to 0.

This method of maximizing the likelihood incorporates these model constraints as well as identifiability constraints, which constrain the response probabilities at each predictor setting to sum to 1. We express this collection of model constraints $\mathbf{U}'\mathbf{C}\log(\mathbf{A}\boldsymbol{\pi}) = \mathbf{0}$ and identifiability constraints as $\mathbf{f}(\boldsymbol{\pi}) = \mathbf{0}$. The method introduces Lagrange multipliers corresponding to these constraints and solves the Lagrangian likelihood equations using a Newton–Raphson algorithm (Aitchison and Silvey 1958; Haber 1985). Let $\boldsymbol{\theta}$ be a vector having elements $\boldsymbol{\pi}$ and the Lagrange multipliers $\boldsymbol{\lambda}$. The Lagrangian likelihood equations have form $\mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$, where

$$\mathbf{h}(\boldsymbol{\theta}) = \mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\lambda}) = \left(\mathbf{f}(\boldsymbol{\pi}), \partial \log[\, l(\boldsymbol{\pi})]/\partial\boldsymbol{\pi} + [\,\partial \mathbf{f}(\boldsymbol{\pi})/\partial\boldsymbol{\pi}]'\boldsymbol{\lambda}\right)'$$

is a vector with terms involving the contrasts in marginal logits that the model specifies as constraints as well as log-likelihood derivatives.

The Newton–Raphson method then is

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \left[\frac{\partial h(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}}\right]^{-1} h(\boldsymbol{\theta}^{(t)}), \quad t = 1, \ldots .$$

This can be computationally intensive because the derivative matrix inverted has dimensions larger than the number of elements in $\boldsymbol{\pi}$. A refinement (Lang 1996a; Lang and Agresti 1994) uses an asymptotic approximation to a reparameterized derivative matrix that has a much simpler form, requiring inverting only a diagonal matrix and a symmetric positive definite matrix.

This ML marginal fitting method is available in specialized software (Appendix A mentions an S-Plus function). It makes no assumption about the model that describes the joint distribution $\boldsymbol{\pi}$. Thus, when the marginal model holds, the ML estimate of $\boldsymbol{\beta}$ in (11.8) is consistent regardless of the dependence structure for that distribution. Several alternative fitting approaches have been considered. Lang and Agresti (1994) simultaneously fitted a marginal model and an unsaturated loglinear model for $\boldsymbol{\pi}$. The complete model can be specified as a special case of (11.8) and fitted using the constraint approach with Lagrange multipliers just described. In standard cases, the marginal and joint model parameters are orthogonal. If the

marginal model holds, the ML estimator of the marginal model parameters is consistent even if the model for the joint distribution is incorrect.

Fitzmaurice and Laird (1993) gave a related ML approach. A one-to-one correspondence holds between $\boldsymbol{\pi}$ and parameters of the saturated loglinear model. They used a further one-to-one correspondence between the main effect and the higher-order parameters of that loglinear model with the marginal probabilities and those same higher-order loglinear parameters. Models were then specified separately for the marginal probabilities and the higher-order (conditional) loglinear parameters. The likelihood is then maximized in terms of the two sets of model parameters. Again, the two sets of parameters are orthogonal, so the ML estimator of marginal model parameters is consistent when the marginal model holds. This *mixed parameter* approach is also available in specialized software (Kastner et al. 1997; see also Appendix A).

Yet another ML approach uses a one-to-one correspondence between $\boldsymbol{\pi}$ and parameters that describe the marginal distributions, the bivariate distributions, the trivariate distributions, and so on (e.g., Glonek and McCullagh 1995; Molenberghs and Lesaffre 1994). Multivariate logistic models then apply to the component distributions, although some higher-order effects may be assumed to vanish, for simplicity. Glonek (1996) proposed a hybrid of this and the Fitzmaurice and Laird (1993) approach.

11.3 MARGINAL MODELING: GENERALIZED ESTIMATING EQUATIONS (GEE) APPROACH

At each combination of predictor values, ML fitting assumes a multinomial distribution for the I^T cell probabilities for the T observations on an I-category response. As the number of predictors increases, the number of multinomial probabilities increases dramatically. Currently, all the ML approaches described above are not practical when T is large or there are many predictors, especially when some are continuous. Compared to the continuous-response case using the multivariate normal, marginal modeling of multivariate categorical responses is also hindered by the lack of a simple multivariate distribution for describing correlations among the T responses. For instance, with T means and a common variance and correlation, the multivariate normal has only $T + 2$ parameters, compared to the $I^T - 1$ parameters for the multinomial.

An alternative to ML fitting uses a multivariate generalization of quasi-likelihood (Section 4.7). Rather than assuming a particular distribution for Y, the quasi-likelihood method specifies only the first two moments; it links the mean to a linear predictor and also specifies how the variance depends on the mean. The estimates are solutions of estimating equations that are likelihood equations under the further assumption of a distribution in the exponential family with that mean and variance (Wedderburn 1974).

11.3.1 Generalized Estimating Equation Methodology: Basic Ideas

Repeated measurement provides a multivariate response $(Y_1, Y_2, \ldots, Y_T)$, where T sometimes varies by subject. As in the univariate case, the quasi-likelihood method specifies a model for $\mu = E(Y)$ and specifies a variance function $v(\mu)$ describing how var(Y) depends on μ. Now, though, that model applies to the marginal distribution for each Y_t. The method also requires a working guess for the correlation structure among $\{Y_t\}$. The estimates are solutions of quasi-likelihood equations called *generalized estimating equations*. The method is often referred to as the GEE method. Liang and Zeger (1986) proposed it for marginal modeling with GLMs. Their work built on related material in the econometrics literature (e.g., Gourieroux et al. 1984; Hansen 1982; White 1982). We outline concepts here and give more details in Section 11.4.

The GEE approach utilizes an assumed covariance structure for $(Y_1, Y_2, \ldots, Y_T)$, specifying a variance function and a pairwise correlation pattern, without assuming a particular multivariate distribution. The GEE estimates of model parameters are valid even if one misspecifies the covariance structure. Consistency (i.e., estimates converging in probability to the true parameters) depends on the first moment but not the second. Specifically, suppose that the model is correct in the sense that the chosen link function and linear predictor truly describe how $E(Y_t)$ depend on the predictors, $t = 1, \ldots, T$. Then the GEE model parameter estimators are consistent.

In practice, a chosen model is never exactly correct. This result is useful, however, for suggesting that the correlation structure need not adversely affect the quality of estimates for whatever model one uses. Often, no a priori information is available about this structure, and the correlation is regarded as a nuisance. A simple implementation of the GEE method naively treats $\{Y_t\}$ as pairwise independent. Although parameter estimates are usually fine under this naive assumption, standard errors are not. More appropriate standard errors result from an adjustment the GEE method makes using the empirical dependence the data exhibit. The naive standard errors based on the independence assumption are updated using the information the data provide about the actual dependence structure to yield more appropriate (*robust*) standard errors.

As an alternative to estimates that treat $\{Y_t\}$ as pairwise independent, the GEE method can use a working guess about the correlation structure but again empirically adjust the standard error. The *exchangeable* working correlation structure treats corr(Y_t, Y_s) as identical for all s and t. This is more flexible and realistic than the naive independence assumption. Even more realistic is an unstructured working correlation that permits a separate correlation for each pair. When T is large, however, this approach suffers some efficiency loss because of the many additional parameters.

In theory, choosing the working correlation wisely can pay benefits of improved efficiency of estimation. However, Liang and Zeger (1986) noted

that estimators based on independence working correlation can have surprisingly good efficiency when the actual correlation is weak to moderate. One can check the sensitivity to the selection by comparing results for different working correlation assumptions. In our experience, when the correlations are modest, all working correlation structures yield similar GEE estimates and standard errors, as the empirical dependence has a large impact on adjusting the naive standard errors. (If they differed substantially, a more careful study of the correlation structure would be necessary.) Unless one expects dramatic differences among the correlations, we recommend the exchangeable working correlation structure. This recognizes the dependence at the cost of only one extra parameter.

The GEE approach is appealing for categorical data because of its computational simplicity compared to ML. Advantages include not requiring a multivariate distribution and the consistency of estimation even with misspecified correlation structure. However, it has limitations. Since the GEE approach does not completely specify the joint distribution, it does not have a likelihood function. Likelihood-based methods are not available for testing fit, comparing models, and conducting inference about parameters. Instead, inference uses Wald statistics constructed with the asymptotic normality of the estimators together with their estimated covariance matrix. However, unless the sample size is quite large, the empirically based standard errors tend to underestimate the true ones (e.g., Firth 1993b). As estimators, those standard errors can also show more variability than parametric estimators (Kauermann and Carroll 2001). Boos (1992) and Rotnitzky and Jewell (1990) proposed analogs of score tests for effects of predictors, using quasi-log-likelihood, that may be more trustworthy than Wald tests. Some statisticians (e.g., Lindsey 1999) are critical of the GEE approach because of the lack of likelihood. Others do not find this problematic, as they regard GEE as an estimation method rather than a model.

11.3.2 Longitudinal Mental Depression Example

For Table 11.2 comparing two treatments for mental depression, ML fitting of a logit model with drug $\times$ time interaction was used in Section 11.2.1. The GEE analysis provides similar results, regardless of the choice of working correlation structure. With the exchangeable structure, the GEE estimated slope (on the logit scale) for the standard drug is $\hat{\beta}_3 = 0.48$ (SE = 0.12). For the new drug the slope increases by $\hat{\beta}_4 = 1.02$ (SE = 0.19). Table 11.6 shows results using the independence working correlations. Estimates are the same to two decimal places. The initial estimates and standard errors there are those that apply if the repeated responses are truly independent. They equal those obtained by using ordinary logistic regression with $3 \times 340 = 1020$ independent observations rather than treating the data as three dependent observations for each of 340 subjects. The empirical standard errors incorporate the sample dependence to adjust the independence-based standard errors.

TABLE 11.6 Output from Using GEE to Fit Logit Model to Table 11.2

```
    Initial Parameter Estimates          GEE Parameter Estimates
                                      Empirical Std Error Estimates
Parameter     Estimate    Std Error    Parameter     Estimate    Std Error
 Intercept    -0.0280      0.1639      Intercept    -0.0280      0.1742
 diagnose     -1.3139      0.1464      diagnose     -1.3139      0.1460
 drug         -0.0596      0.2222      drug         -0.0596      0.2285
 time          0.4824      0.1148      time          0.4824      0.1199
 drug*time     1.0174      0.1888      drug*time     1.0174      0.1877

                    Working Correlation Matrix
                        Col1        Col2        Col3
                Row1    1.0000      0.0000      0.0000
                Row2    0.0000      1.0000      0.0000
                Row3    0.0000      0.0000      1.0000
```

With exchangeable correlation structure, the estimated common correlation between pairs of the three responses is -0.003. The successive observations apparently have pairwise appearance like independent observations. This is quite unusual for repeated measurement data. For this reason, similar results occur from fitting the model assuming the three observations for a subject actually come from three separate subjects (i.e., assuming 1020 independent observations).

11.3.3 GEE Approach for Multinomial Responses: Insomnia Example

Liang and Zeger (1986) originally specified the GEE methodology for modeling univariate marginal distributions, such as the binomial and Poisson. It extends to marginal modeling of multinomial responses. Lipsitz et al. (1994) outlined a GEE approach for cumulative logit models with repeated ordinal responses. With this approach, for each pair of outcome categories one selects a working correlation matrix for the pairs of repeated observations. Each multinomial response at a fixed observation uses the $(I-1)\times(I-1)$ multinomial covariance matrix. Section 11.4.4 has details.

We illustrate for the insomnia data of Table 11.4. In Section 11.2.3 we used ML to fit the marginal model

$$\text{logit}\big[P(Y_t \le j)\big] = \alpha_j + \beta_1 t + \beta_2 x + \beta_3 tx$$

for Y_t = time to fall asleep with treatment x at occasion t. With independence working correlation structure, the GEE estimates are $\hat{\beta}_1 = 1.038$ (SE = 0.168), $\hat{\beta}_2 = 0.034$ (SE = 0.238), and $\hat{\beta}_3 = 0.708$ (SE = 0.244). The estimates are similar to the ML estimates, and the substantive conclusions are the same. Considerable evidence exists that the distribution of time to fall asleep decreased more for the treatment group than for the placebo group.

11.4 QUASI-LIKELIHOOD AND ITS GEE MULTIVARIATE EXTENSION: DETAILS*

A GLM assumes a certain distribution for the response variable. Sometimes it is unclear how to select it. However, often there is a plausible relationship between the mean and variance, such as $v(\mu_i) = \phi\mu_i$ for count data. Then, an alternative to ML estimation is quasi-likelihood estimation (Section 4.7). We next present some details about this method and its GEE extension for marginal modeling of multivariate responses.

We begin with models for a single response and later discuss marginal models for a multivariate response. For subject $i, i = 1, \ldots, n$, let y_i be the outcome on Y with $\mu_i = E(Y_i)$ and variance function $v(\mu_i)$, and let x_{ij} be the value of explanatory variable j. For link function g, the linear predictor is $\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij} = \mathbf{x}_i'\boldsymbol{\beta}$. The quasi-likelihood (QL) parameter estimates $\hat{\boldsymbol{\beta}}$ are the solutions of quasi-score equations

$$\mathbf{u}(\boldsymbol{\beta}) = \sum_i \left(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}\right)' v(\mu_i)^{-1}(y_i - \mu_i) = \mathbf{0}, \tag{11.9}$$

where $\mu_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$. These *estimating equations* are the same as the likelihood equations (4.22) for GLMs when we substitute

$$\frac{\partial\mu_i}{\partial\beta_j} = \frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j} = \frac{\partial\mu_i}{\partial\eta_i}x_{ij}.$$

They are not likelihood equations, however, without the extra assumption that $\{y_i\}$ has distribution in the natural exponential family. Under that assumption, $v(\mu_i)$ characterizes the distribution within the natural exponential family (Jørgensen 1987). Another motivation for equations (11.9) is that with $v(\mu_i)$ replaced by known variance v_i, they result from the weighted least squares problem of minimizing $\sum_i (y_i - \mu_i)^2 v_i^{-1}$.

The likelihood equations (4.22) for a GLM depend only on the mean and variance of $\{y_i\}$ and the link function g, which determines $\partial\mu_i/\partial\eta_i$. Thus, Wedderburn (1974) suggested using them as estimating equations for *any* link and variance function, even if they do not correspond to a particular member of the natural exponential family.

11.4.1 Properties of Quasi-likelihood Estimators

In the quasi-likelihood (QL) method, the *quasi-score function* $u_j(\boldsymbol{\beta})$ in (11.9) is called an *unbiased estimating function*; this term refers to any function $h(\mathbf{y}; \boldsymbol{\beta})$ of $\mathbf{y}$ and $\boldsymbol{\beta}$ such that $E[h(\mathbf{Y}; \boldsymbol{\beta})] = 0$ for all $\boldsymbol{\beta}$. The equations (11.9) that determine $\hat{\boldsymbol{\beta}}$ are called *estimating equations*.

The quasi-likelihood method treats the quasi-score function as the derivative of a function called the *quasi-log likelihood*. This function may not be a

proper log likelihood function. Nonetheless, McCullagh (1983) showed that QL estimators have properties similar to those of ML estimators. For instance, the QL estimators $\hat{\boldsymbol{\beta}}$ are asymptotically normal with covariance matrix approximated by

$$\mathbf{V} = \left[\sum_i \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)' \left[v(\mu_i) \right]^{-1} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right) \right]^{-1}. \tag{11.10}$$

This is equivalent to the formula for the large-sample covariance matrix of the ML estimator in a GLM [which is estimated by (4.28)].

A key result is that the QL estimator $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ (i.e., $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$) even if the variance function is misspecified, as long as the specification is correct for the link function and linear predictor. That is, assuming that the model form $g(\mu_i) = \sum_j \beta_j x_{ij}$ is correct, the consistency of $\hat{\boldsymbol{\beta}}$ holds even if the true variance function is not $v(\mu_i)$. We now give a heuristic explanation for this.

When truly $\mu_i = g^{-1}(\sum_j \beta_j x_{ij})$, then from (11.9), $E[u_j(\boldsymbol{\beta})] = 0$ for all j. From (11.9), $\mathbf{u}(\boldsymbol{\beta})/n$ is a vector of sample means. By a law of large numbers, it converges in probability to its expected value of $\mathbf{0}$. The solution $\hat{\boldsymbol{\beta}}$ of the quasi-score equations is a continuous function of these sample means, so it converges to $\boldsymbol{\beta}$, since $\hat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ for which the sum is exactly equal to $\mathbf{0}$. The consistency also follows from general results for unbiased estimating functions (Liang and Zeger 1995).

11.4.2 Sandwich Covariance Adjustment for Variance Misspecification

If one assumes that $\text{var}(Y_i) = v(\mu_i)$ but the true $\text{var}(Y_i) \neq v(\mu_i)$, then the actual asymptotic covariance matrix of the QL estimator $\hat{\boldsymbol{\beta}}$ is not $\mathbf{V}$ as given in (11.10). Instead, it is (Diggle et al. 2001; White 1982)

$$\mathbf{V}\left[\sum_i \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)' \left[v(\mu_i) \right]^{-1} \text{var}(Y_i) \left[v(\mu_i) \right]^{-1} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right) \right] \mathbf{V}. \tag{11.11}$$

Even though the variances are scalar, we express the matrices in this form to motivate the GEE multivariate extension discussed below. Matrix (11.11) simplifies to $\mathbf{V}$ if $\text{var}(Y_i) = v(\mu_i)$. In practice, the true variance function is unknown. A consistent estimator of (11.11) is a sample analog, replacing μ_i by $\hat{\mu}_i$ and $\text{var}(Y_i)$ by $(y_i - \hat{\mu}_i)^2$ (Liang and Zeger 1986). The estimated covariance matrix is valid regardless of whether the variance specification $v(\mu_i)$ is correct. This estimated covariance matrix is called a *sandwich estimator*, because the empirical evidence is sandwiched between the model-driven covariance matrices.

In summary, even with incorrect specification of the variance function, one can still consistently estimate $\boldsymbol{\beta}$ and one can estimate the asymptotic variance

of $\hat{\boldsymbol{\beta}}$ by estimating the sandwich adjustment (11.11). However, some efficiency loss occurs when the variance chosen, $v(\mu_i)$, is wildly inaccurate. Also, the number of clusters n may need to be large for the sample version of (11.11) to work well; otherwise, it can be biased downward. Of course, a modeling process never gets anything exactly correct. Just as the variance function chosen only approximates the true one (hopefully, closely), so is the specification for the mean only approximate.

11.4.3 GEE Methodology: Technical Details

Now we consider the generalized estimating equations (GEE) multivariate generalization of QL. For subject i, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT_i})'$ and $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iT_i})'$, where $\mu_{it} = E(Y_{it})$. The number T_i of responses may vary by cluster. Let $\mathbf{x}_{it}$ denote a $p \times 1$ vector of explanatory variable values for y_{it}. The notation allows for cases where explanatory variables also vary for the repeated measurements. The linear predictor of the model is $\eta_{it} = g(\mu_{it}) = \mathbf{x}_{it}'\boldsymbol{\beta}$ for link function g. The model refers to the marginal distribution at each t rather than the joint distribution. Let $\mathbf{X}_i$ be the $T_i \times p$ matrix of predictor values for cluster (or subject) i, for which row t is $\mathbf{x}_{it}'$.

We assume that y_{it} has probability mass function of form

$$f(y_{it}; \theta_{it}, \phi) = \exp\Big\{\big[y_{it}\theta_{it} - b(\theta_{it})\big]\big/\phi + c(y_{it}, \phi)\Big\}.$$

When ϕ is known, this is the natural exponential family with natural parameter θ_{it}. From Section 4.4.1,

$$\mu_{it} = E(Y_{it}) = b'(\theta_{it}), \qquad v(\mu_{it}) = \text{var}(Y_{it}) = b''(\theta_{it})\phi.$$

The GEE method also assumes a working correlation matrix $\mathbf{R}(\boldsymbol{\alpha})$ for $\mathbf{Y}_i$, depending on parameters $\boldsymbol{\alpha}$. The exchangeable working correlation has $\text{corr}(Y_{it}, Y_{is}) = \alpha$ for each pair in $\mathbf{Y}_i$. Let $\mathbf{b}_i(\boldsymbol{\theta}) = (b(\theta_{i1}), \ldots, b(\theta_{iT_i}))$, and let $\mathbf{B}_i$ denote a diagonal matrix with main diagonal elements $\mathbf{b}_i''(\boldsymbol{\theta})$. Then the working covariance matrix for $\mathbf{Y}_i$ is

$$\mathbf{V}_i = \mathbf{B}_i^{1/2}\mathbf{R}(\boldsymbol{\alpha})\mathbf{B}_i^{1/2}\phi. \tag{11.12}$$

Note that $\mathbf{V}_i = \text{cov}(\mathbf{Y}_i)$ if $\mathbf{R}$ is the true correlation matrix for $\mathbf{Y}_i$.

Now let $\boldsymbol{\Delta}_i$ be the diagonal matrix with elements $\partial\theta_{it}/\partial\eta_{it}$ on the main diagonal for $t = 1, \ldots, T_i$. (For the canonical link, this is the identity matrix.) Let $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta} = \mathbf{B}_i\boldsymbol{\Delta}_i\mathbf{X}_i$ be a $T_i \times p$ matrix with typical element expressing $\partial\mu_{it}/\partial\beta_j$ in the form $(\partial\mu_{it}/\partial\theta_{it})(\partial\theta_{it}/\partial\eta_{it})(\partial\eta_{it}/\partial\beta_j)$. From (11.9), for univariate GLMs the quasi-likelihood estimating equations have the form

$$\sum_i (\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta})' v(\boldsymbol{\mu}_i)^{-1}\big[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\big] = \mathbf{0},$$

where $\mu_i = \mu_i(\boldsymbol{\beta}) = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$. The analog of this in the multivariate case is the set of *generalized estimating equations*

$$\sum_{i=1}^{n} \mathbf{D}_i' \mathbf{V}_i^{-1} \left[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}) \right] = \mathbf{0}.$$

The GEE estimator $\hat{\boldsymbol{\beta}}$ is the solution of these equations.

The naive approach, which sets $\mathbf{R}(\boldsymbol{\alpha}) = \mathbf{I}$, treats pairs of responses as independent. In that case, (11.12) simplifies to $\mathbf{V}_i = \mathbf{B}_i \phi$, and the generalized estimating equations simplify to

$$\begin{aligned} \sum_i \mathbf{D}_i' \mathbf{V}_i^{-1} \left[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}) \right] &= \sum_i \mathbf{X}_i' \boldsymbol{\Delta}_i \mathbf{B}_i \mathbf{V}_i^{-1} \left[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}) \right] \\ &= (1/\phi) \sum_i \mathbf{X}_i' \boldsymbol{\Delta}_i \left[\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}) \right] = \mathbf{0}, \end{aligned}$$

or $\sum_i \mathbf{X}_i' \boldsymbol{\Delta}_i [\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] = \mathbf{0}$. The solution $\hat{\boldsymbol{\beta}}$ is then the same as the ordinary estimator for a GLM with the chosen link function and variance function, treating $(y_{i1}, \ldots, y_{iT_i})$ as independent observations.

Normally, one selects a working correlation matrix permitting dependence, such as the exchangeable structure. For time-series data, also popular is the autoregressive structure, $\text{corr}(Y_{it}, Y_{is}) = \alpha^{|t-s|}$, which treats observations farther apart in time as more weakly correlated. Liang and Zeger (1986) suggested computing the GEE estimates by iterating between a modified Fisher scoring algorithm for solving the generalized estimating equations for $\boldsymbol{\beta}$ (given current estimates of $\boldsymbol{\alpha}$ and ϕ) and using residuals for moment estimation of $\boldsymbol{\alpha}$ and ϕ (based on the current estimates of $\boldsymbol{\beta}$). They suggested estimates of $\mathbf{R}(\boldsymbol{\alpha})$ for a variety of correlation structures. Alternative algorithms simultaneously solve estimating equations for $\boldsymbol{\beta}$ and for association parameters (e.g., Liang et al. 1992; see also Note 11.8). GEE algorithms need not converge, but often one iteration gives adequate results (Lipsitz et al. 1991).

Liang and Zeger (1986) showed asymptotic normality and consistency as the number of clusters n increases. Under certain regularity conditions,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_G) .$$

Here, generalizing (11.11), $\mathbf{V}_G = \lim_{n \to \infty} \mathbf{V}_{G,n}$ with

$$\mathbf{V}_{G,n} = n \left[\sum_i \mathbf{D}_i' \mathbf{V}_i^{-1} \ \mathbf{D}_i \right]^{-1} \left[\sum_i \mathbf{D}_i' \mathbf{V}_i^{-1} \text{cov}(\mathbf{Y}_i) \mathbf{V}_i^{-1} \mathbf{D}_i \right] \left[\sum_i \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i \right]^{-1} .$$

The estimated covariance matrix $\hat{\mathbf{V}}_{G,n}/n$ of $\hat{\boldsymbol{\beta}}$ replaces $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$, ϕ with $\hat{\phi}$, $\boldsymbol{\alpha}$ with $\hat{\boldsymbol{\alpha}}$, and $\text{cov}(\mathbf{Y}_i)$ by $[\mathbf{y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}})][\mathbf{y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}})]'$. The purpose of the

sandwich estimator is to use the data's empirical evidence about covariation to adjust the standard errors in case the true covariance differs substantially from the working guess.

When the working correlation structure is the true one and $\text{cov}(\mathbf{Y}_i) = \mathbf{V}_i$, the asymptotic covariance matrix $\mathbf{V}_{G,n}/n$ simplifies to $(\sum_i \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i)^{-1}$. This is the relevant covariance if we put complete faith in our guess about the correlation structure.

With binary data, the correlation may not be the best way to express the within-cluster association. The marginal probabilities constrain the possible correlation values, since the range of possible values for $E(Y_{it}Y_{is}) = P(Y_{it} = 1, Y_{is} = 1)$ depends on $P(Y_{it} = 1)$ and $P(Y_{is} = 1)$. An alternative approach uses the odds ratio, for instance by modeling the log odds ratios for pairs in a cluster as exchangeable. This has the advantage that the association parameters are distinct from the means. See Fitzmaurice et al. (1993) and Lipsitz et al. (1991). Carey et al. (1993) suggested an iterative *alternating logistic regressions* algorithm. It alternates between a GEE step for the regression parameters in the model for the mean and a step for an association model for the log odds ratio. This is useful when the structure of the association is itself a major focus rather than a nuisance.

11.4.4 GEE Approach: Multinomial Responses

We now briefly describe the Lipsitz et al. (1994) GEE approach for marginal modeling with a multinomial response. This is appropriate, for instance, with cumulative logit models. Let $y_{it}(j) = 1$ if observation t in cluster i has outcome j ($j = 1, \ldots, I - 1$). Let $\mathbf{y}_i$ be the $T_i(I - 1)$ binary indicators for cluster i. Then, one selects a $[T_i(I - 1)] \times [T_i(I - 1)]$ working covariance matrix $\mathbf{V}_i$ for $\mathbf{y}_i$, specifying a pattern for $\text{corr}(Y_{it}(j), Y_{is}(k))$ for each pair of outcome categories (j, k) and each pair (t, s). The $(I - 1) \times (I - 1)$ block of $\mathbf{V}_{it}$ for $(y_{it}(1), \ldots, y_{it}(I - 1))$ is a multinomial covariance matrix with $v_{it}(j) = P(Y_{it}(j) = 1)[1 - P(Y_{it}(j) = 1)]$ on the main diagonal and $-P(Y_{it}(j) = 1)P(Y_{it}(k) = 1)$ off it. The remaining elements of $\mathbf{V}_i$ contain elements $\text{cov}(Y_{it}(j), Y_{is}(k))$. For instance, one possibility is the exchangeable structure, $\text{corr}(Y_{it}(j), Y_{is}(k)) = \rho_{jk}$ for all t and s.

In this approach the generalized estimating equations for $\boldsymbol{\beta}$ again have the form

$$\mathbf{u}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0},$$

where $\boldsymbol{\mu}_i$ is the vector of probabilities associated with $\mathbf{y}_i$, $\mathbf{D}_i' = \partial \boldsymbol{\mu}_i'/\partial \boldsymbol{\beta}$, and the parameters are evaluated at their current estimates. Lipsitz et al. suggested a Fisher scoring algorithm for solving these equations and a method of moments update for estimating $\{\rho_{jk}\}$ at each step of the iteration. An

empirically adjusted sandwich covariance matrix of $\hat{\boldsymbol{\beta}}$ is again

$$\left[\sum_{i=1}^{n} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right]^{-1} \left[\sum_{i=1}^{n} \mathbf{D}_i' \mathbf{V}_i^{-1} \text{cov}(\mathbf{Y}_i) \mathbf{V}_i^{-1} \mathbf{D}_i\right] \left[\sum_{i=1}^{n} \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i\right]^{-1}.$$

This is estimated by substituting $\hat{\boldsymbol{\mu}}_i$ from the model fit and replacing $\text{cov}(\mathbf{Y}_i)$ by the empirical covariance matrix of $\mathbf{y}_i$.

11.4.5 Dealing with Missing Data

Unfortunately, studies with repeated measurement often have cases for which at least one response in a cluster is missing. In a longitudinal study, for instance, some subjects may drop out before its conclusion. When data are missing, analyzing the observed data alone as if no data are missing can result in biased estimates.

An advantage of the GEE method is that different clusters can have different numbers of observations. The data input file has a separate line for each observation, and for longitudinal studies, computations use those times for which a subject has an observation. However, bias can arise in GEE estimates unless one can make certain assumptions about why the data are missing.

Let $\mathbf{Y}^{(o)}$ denote the observed responses, $\mathbf{Y}^{(m)}$ the missing responses, and $\mathbf{Y}$ their union. Let M denote a missing data indicator that equals 1 when an observation is missing and 0 otherwise. Little and Rubin (1987) called the data *missing completely at random* if M is statistically independent of $\mathbf{Y}$; that is, the probability that an observation is missing is independent of that observation's value, although it may depend on the explanatory variables. Less restrictively, they called the data *missing at random* if the distribution of $(M \mid \mathbf{Y})$ equals that of $(M \mid \mathbf{Y}^{(o)})$; that is, missingness depends only on $\mathbf{Y}^{(o)}$ and not on the missing values.

When either of these is plausible, with a likelihood-based analysis it is not necessary to model the missingness mechanism. An analysis using only $\mathbf{Y}^{(o)}$ is not systematically biased. The same is true with GEE methods when estimating equations can be weighted by response probabilities (Robins et al. 1995). Otherwise, however, with non-likelihood-based methods such as GEE, the missingness process can be ignored only when data are missing completely at random. Kenward et al. (1994) illustrated the breakdown in GEE estimates when the data are not missing completely at random.

Often, missingness depends on the missing values. For instance, in a longitudinal study measuring pain, perhaps a subject dropped out when the pain got above some threshhold. Then, more complex analyses are needed that model the joint distribution of $\mathbf{Y}$ and M (Little 1998). Let $f(\cdot)$ denote a generic probability mass function, which also depends on explanatory variables $\mathbf{x}$ and parameters. *Selection models* factor the joint distribution of $\mathbf{Y}$

and M as

$$f(\mathbf{y}, M; \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\psi}) = f(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}) f(M \mid \mathbf{y}; \mathbf{x}, \boldsymbol{\psi}),$$

where $f(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta})$ is the model in the absence of missing values and $f(M \mid \mathbf{y}; \mathbf{x}, \boldsymbol{\psi})$ is the model for the missing-data mechanism. *Pattern mixture models* use the alternative factorization,

$$f(\mathbf{y}, M; \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) = f(\mathbf{y} \mid M, \mathbf{x}, \boldsymbol{\phi}) f(M; \mathbf{x}, \boldsymbol{\theta}),$$

which conditions the distribution of $\mathbf{Y}$ on the missing data pattern. The two specifications are equivalent when M is independent of $\mathbf{Y}$, with $\boldsymbol{\beta} = \boldsymbol{\phi}$ and $\boldsymbol{\psi} = \boldsymbol{\theta}$. For discussion of advantages of each modeling approach and details on ways of modeling missingness, see Little (1998) and references in Note 11.9. See Stokes et al. (2000, p. 524) for an example of building the missingness pattern into a model to check whether it is associated with the response or interacts with effects of explanatory variables.

Analyses in the presence of much missingness should be made with caution. Typically, little is known about the missing data mechanism, and assumptions about it cannot be checked. Since inferences may not be robust, a sensitivity study is necessary to check how results depend on specification of that mechanism. In the absence of a model for the missingness, one should at least compare results of the analysis using all available cases for all clusters to the analysis using only clusters having no missing observations. If results differ substantially, conclusions should be very tentative until the reasons for missingness can be studied.

11.5 MARKOV CHAINS: TRANSITIONAL MODELING

When Y_t denotes the response at time t, $t = 0, 1, 2, \ldots,$ the indexed family of random variables $(Y_0, Y_1, Y_2, \ldots)$ is a *stochastic process*. The *state space* of the process is the set of possible values for Y_t. The value Y_0 is the *initial state*. When the state space is categorical and observations occur at a discrete set of times, $\{Y_t\}$ has *discrete state space* and *discrete time*.

11.5.1 Transitional Models

The main focus is usually on the dependence of Y_t on the responses $\{y_0, y_1, \ldots, y_{t-1}\}$ observed previously as well as any explanatory variables. Models of this type are called *transitional models*. Let $f(y_0, \ldots, y_T)$ denote the joint probability mass function of $(Y_0, \ldots, Y_T)$ (ignoring, for now, ex-

planatory variables). Transitional models use the factorization

$$f(y_0, \ldots, y_T) = f(y_0) f(y_1 | y_0) f(y_2 | y_0, y_1) \cdots f(y_T | y_0, y_1, \ldots, y_{T-1}).$$

Unlike the marginal models in the other sections of this chapter, this modeling is conditional on previous responses.

In this section we introduce discrete-time *Markov chains*, a simple stochastic process having discrete state space. Many transitional models have Markov chain structure for at least part of the model.

11.5.2 First-Order Markov Chains

A *Markov chain* is a stochastic process for which, for all t, the conditional distribution of Y_{t+1}, given $Y_0, \ldots, Y_t$, is identical to the conditional distribution of Y_{t+1} given Y_t alone. That is, given Y_t, Y_{t+1} is conditionally independent of $Y_0, \ldots, Y_{t-1}$. Knowing the present state of a Markov chain, information about past states does not help us predict the future. For Markov chains,

$$f(y_0, \ldots, y_T) = f(y_0) f(y_1 | y_0) f(y_2 | y_1) \ldots f(y_T | y_{T-1}). \quad (11.13)$$

A stochastic process is a kth-*order Markov chain* if, for all t, the conditional distribution of Y_{t+1}, given $Y_0, \ldots, Y_t$, is identical to the conditional distribution of Y_{t+1}, given $(Y_t, \ldots, Y_{t-k+1})$. Given the states at the previous k times, the future behavior of the chain is independent of past behavior before those k times. Our discussion here focuses mainly on ordinary Markov chains as in (11.13), which are first order ($k = 1$).

Denote the conditional probability $P(Y_t = j \,|\, Y_{t-1} = i)$ by $\pi_{j|i}(t)$. The $\{\pi_{j|i}(t)\}$, which satisfy $\sum_j \pi_{j|i}(t) = 1$, are called *transition probabilities*. The $I \times I$ matrix $\{\pi_{j|i}(t),\ i = 1, \ldots, I,\ j = 1, \ldots, I\}$ is a *transition probability matrix*. It is called *one-step*, to distinguish it from the matrix of probabilities for k-step transitions from time $t - k$ to time t.

From (11.13), the joint distribution for a Markov chain depends only on one-step transition probabilities and the marginal distribution for the initial state. It also follows that the joint distribution satisfies loglinear model

$$(Y_0 Y_1, Y_1 Y_2, \ldots, Y_{T-1} Y_T).$$

For a sample of realizations of a stochastic process, a contingency table displays counts of the possible sequences. A test of fit of this loglinear model checks whether the process plausibly satisfies the Markov property.

Statistical inference for Markov chains uses standard methods of categorical data analysis. For example, consider ML estimation of transition probabilities. Let $n_{ij}(t)$ denote the number of transitions from state i at time $t - 1$ to state j at time t. For fixed t, $\{n_{ij}(t)\}$ form the two-way marginal table for dimensions $t - 1$ and t of an I^{T+1} contingency table. For the $n_{i+}(t)$ subjects

in category i at time $t-1$, suppose that $\{n_{ij}(t), j = 1, \ldots, I\}$ have a multinomial distribution with parameters $\{\pi_{j|i}(t)\}$. Let $\{n_{i0}\}$ denote the initial counts. Suppose that they also have a multinomial distribution, with parameters $\{\pi_{i0}\}$. If subjects behave independently, from (11.13) the likelihood function is proportional to

$$\left(\prod_{i=1}^{I} \pi_{i0}^{n_{i0}}\right)\left\{\prod_{t=1}^{T}\prod_{i=1}^{I}\left[\prod_{j=1}^{I} \pi_{j|i}(t)^{n_{ij}(t)}\right]\right\}. \tag{11.14}$$

The transition probabilities are parameters of IT independent multinomial distributions. From Anderson and Goodman (1957), the ML estimates are

$$\hat{\pi}_{j|i}(t) = n_{ij}(t)/n_{i+}(t) .$$

11.5.3 Respiratory Illness Example

Table 11.7 refers to a longitudinal study at Harvard of effects of air pollution on respiratory illness in children. The children were examined annually at ages 9 through 12 and classified according to the presence or absence of wheeze.

Denote the binary response (wheeze, no wheeze) by Y_t at age t, $t = 9, 10, 11, 12$. The loglinear model $(Y_9Y_{10}, Y_{10}Y_{11}, Y_{11}Y_{12})$ represents a first-order Markov chain. It fits poorly, with $G^2 = 122.9$ (df = 8). Given the state at time t, classification at time $t+1$ depends on states at times previous to time t. The model $(Y_9Y_{10}Y_{11}, Y_{10}Y_{11}Y_{12})$ represents a second-order Markov chain, satisfying conditional independence at ages 9 and 12, given states at ages 10 and 11. This model also fits poorly, with $G^2 = 23.9$ (df = 4). The poor fits may partly reflect subject heterogeneity, since these analyses ignore possibly relevant covariates such as parental smoking behavior.

The loglinear model $(Y_9Y_{10}, Y_9Y_{11}, Y_9Y_{12}, Y_{10}Y_{11}, Y_{10}Y_{12}, Y_{11}Y_{12})$ that permits association at each pair of ages fits well, with $G^2 = 1.5$ (df = 5). Table

TABLE 11.7 Results of Breath Test at Four Ages[a]

Y_9	Y_{10}	Y_{11}	Y_{12}	Count	Y_9	Y_{10}	Y_{11}	Y_{12}	Count
1	1	1	1	94	2	1	1	1	19
1	1	1	2	30	2	1	1	2	15
1	1	2	1	15	2	1	2	1	10
1	1	2	2	28	2	1	2	2	44
1	2	1	1	14	2	2	1	1	17
1	2	1	2	9	2	2	1	2	42
1	2	2	1	12	2	2	2	1	35
1	2	2	2	63	2	2	2	2	572

[a] 1, wheeze; 2, no wheeze.

Source: Ware et al. (1988).

TABLE 11.8 Estimated Conditional Log Odds Ratios for Table 11.7

Association	Estimate	Simpler Structure
Y_9Y_{10}	1.81	1.75
$Y_{10}Y_{11}$	1.65	1.75
$Y_{11}Y_{12}$	1.85	1.75
Y_9Y_{11}	0.95	1.04
Y_9Y_{12}	1.05	1.04
$Y_{10}Y_{12}$	1.07	1.04

11.8 shows its ML estimates of pairwise conditional log odds ratios. The association seems similar for pairs of ages 1 year apart, and somewhat weaker for pairs of ages more than 1 year apart. The simpler model in which

$$\lambda_{ij}^{Y_9Y_{10}} = \lambda_{ij}^{Y_{10}Y_{11}} = \lambda_{ij}^{Y_{11}Y_{12}} \quad \text{and} \quad \lambda_{ij}^{Y_9Y_{11}} = \lambda_{ij}^{Y_9Y_{12}} = \lambda_{ij}^{Y_{10}Y_{12}}$$

fits well, with $G^2 = 2.3$ (df = 9). The estimated log odds ratios are 1.75 in the first case, and 1.04 in the second.

11.5.4 Transitional Models with Explanatory Variables

Transitional models usually also include explanatory variables $\mathbf{x}$. The joint mass function of T sequential responses is then

$$\begin{aligned} &f(y_1, \ldots, y_T; \mathbf{x}) \\ &\quad = f(y_1; \mathbf{x}) f(y_2 | y_1; \mathbf{x}) f(y_3 | y_1, y_2; \mathbf{x}) \cdots f(y_T | y_1, y_2, \ldots, y_{T-1}; \mathbf{x}) . \end{aligned}$$

With binary y, for instance, one might specify a logistic regression model for each term in this factorization,

$$\begin{aligned} &f(y_t | y_1, \ldots, y_{t-1}; \mathbf{x}_t) \\ &\quad = \frac{\exp[y_t(\alpha + \beta_1 y_1 + \cdots + \beta_{t-1} y_{t-1} + \boldsymbol{\beta}' \mathbf{x}_t)]}{1 + \exp(\alpha + \beta_1 y_1 + \cdots + \beta_{t-1} y_{t-1} + \boldsymbol{\beta}' \mathbf{x}_t)}, \quad y_t = 0,1. \end{aligned}$$

Here, the predictor $\mathbf{x}$ may take different value for each component. The model treats previous responses as explanatory variables. It is called a *regressive logistic model* (Bonney 1987).

The interpretation and magnitude of $\hat{\boldsymbol{\beta}}$ depends on how many previous observations are in the model. Within-cluster effects may diminish markedly

by conditioning on previous responses. This is an important difference from marginal models, for which the interpretation does not depend on the specification of the dependence structure. In the special case of first-order Markov structure, the coefficients of $\{y_1, \ldots, y_{t-2}\}$ equal 0 in the model for y_t (e.g., Azzalini 1994; Bonney 1987). It may help to allow interaction between $\mathbf{x}_t$ and y_{t-1} in their effects on y_t.

For a given subject, the product of the conditional mass functions determines that subject's contribution to the likelihood function. (One usually ignores the contribution of the marginal distribution for the first term.) That is, given the predictor, the model treats repeated transitions by a subject as independent. Thus, one can fit the model with ordinary GLM software, treating each transition as a separate observation (Bonney 1986).

11.5.5 Child's Respiratory Illness and Maternal Smoking

Table 11.9 is also from the Harvard study of air pollution and health. At ages 7 through 10, children were evaluated annually on the presence of respiratory illness. A predictor is maternal smoking at the start of the study, where $s = 1$ for smoking regularly and $s = 0$ otherwise. Let y_t denote the response at age t ($t = 7, 8, 9, 10$). We consider the regressive logistic model

$$\text{logit}\big[P(Y_t = 1)\big] = \alpha + \beta_1 s + \beta_2 t + \beta_3 y_{t-1}, \quad t = 8, 9, 10.$$

Each subject contributes three observations to the model fitting. The data set consists of 12 binomials, for the $2 \times 3 \times 2$ combinations of (s, t, y_{t-1}). For instance, for the combination $(0, 8, 0)$, $y_8 = 0$ for $237 + 10 + 15 + 4 =$

TABLE 11.9 Child's Respiratory Illness by Age and Maternal Smoking

Child's Respiratory Illness			No Maternal Smoking		Maternal Smoking	
			Age 10		Age 10	
Age 7	Age 8	Age 9	No	Yes	No	Yes
No	No	No	237	10	118	6
		Yes	15	4	8	2
	Yes	No	16	2	11	1
		Yes	7	3	6	4
Yes	No	No	24	3	7	3
		Yes	3	2	3	1
	Yes	No	6	2	4	2
		Yes	5	11	4	7

Source: Data courtesy of James Ware.

266 subjects and $y_8 = 1$ for $16 + 2 + 7 + 3 = 28$ subjects. The ML fit is

$$\text{logit}\left[\hat{P}(Y_t = 1)\right] = -0.293 + 0.296s - 0.243t + 2.211y_{t-1},$$

with SE values $(0.846, 0.156, 0.095, 0.158)$. Not surprisingly, the previous observation has a strong effect. Given that and the child's age, there is slight evidence of a positive effect of maternal smoking: The likelihood-ratio statistic for H_0: $\beta_1 = 0$ is 3.55 (df = 1, $P = 0.06$). The model itself does not show any evidence of lack of fit ($G^2 = 3.1$, df = 8).

NOTES

Section 11.1: Comparing Marginal Distributions: Multiple Responses

11.1. Darroch (1981) surveyed thoroughly the relationships among statistics for testing marginal homogeneity and their connections with generalized CMH analyses. See also Mantel and Byar (1978) and White et al. (1982). Croon et al. (2000) studied a variety of hypotheses for longitudinal data in the context of the generalized loglinear model.

Section 11.2: Marginal Modeling: Maximum Likelihood Approach

11.2. For other work on ML fitting of marginal models, see Bergsma and Rudas (2002), Ekholm et al. (2000), Fitzmaurice et al. (1993), and Lang et al. (1999).

Section 11.3: Marginal Modeling: Generalized Estimating Equations Approach

11.3. Liang et al. (1992) discussed GEE methods for categorical (primarily binary) responses. For multinomial responses, see Heagerty and Zeger (1996), Lipsitz et al. (1994), Miller et al. (1993), and references in Agresti and Natarajan (2001). More general models with ordinal responses allow for dispersion parameters that also depend on covariates (Toledano and Gatsonis 1996).

11.4. LaVange et al. (2001) used GEE methods to adjust for clustered sampling in surveys and clinical trials. Boos (1992) discussed generalized score tests that incorporate empirical variance estimates, illustrating with tests for trend and lack of fit in binary regression.

11.5. Koch et al. (1977) used weighted least squares (WLS) to fit marginal models to Table 11.2. WLS for categorical modeling is described in Section 15.1. It has severe limitations (e.g., covariates must be categorical and marginal tables cannot be sparse) but led naturally to the GEE approach.

Section 11.4: Quasi-likelihood and Its GEE Multivariate Extension: Details

11.6. Firth (1993b) provided a useful overview of quasi-likelihood methods. McCullagh (1983) showed that under correct specification of the mean and the variance function, quasi-likelihood estimators are asymptotically efficient among estimators that are locally linear in $\{y_i\}$. His result generalizes the Gauss–Markov theorem, although in an asymptotic rather than exact manner. See also Heyde (1997) and Liang and Zeger (1995) for discussions of unbiased estimating functions and their connections with

asymptotic consistency and efficiency. Godambe showed in 1960 that ML estimators are optimal solutions with an unbiased estimating function. When quasi-likelihood estimators are not ML, Cox (1983) and Firth (1987) suggested that they still retain good efficiency when the departure from the natural exponential family is at most moderate, such as modest overdispersion relative to such a family.

11.7. The generalized estimating equations are likelihood equations, and hence the GEE estimates are also ML, in certain cases. Examples are multivariate normal data or binary data when the working covariance is correct (Fitzmaurice et al. 1993). Results about effects of model misspecification arise in a variety of model-building contexts. For general theory, see Gourieroux et al. (1984), Hansen (1982), Liang and Zeger (1995), and White (1982).

11.8. A GEE2 analysis adds estimating equations for the correlation structure (Prentice and Zhao 1991). This has the potential to increase efficiency. A disadvantage is that, unlike with ordinary GEE, $\hat{\boldsymbol{\beta}}$ is no longer consistent if this part of the model is misspecified. Qu et al. (2000) showed how to increase efficiency by representing the working correlation matrix by a linear combination of basis matrices.

11.9. For surveys of ways to handle missing data, see Little (1998), Little and Rubin (1987, Chap. 9), Schafer (1997), and Verbeke and Molenberghs (2000). See also Baker and Laird (1988), Fay (1986), Fitzmaurice et al. (1994), Forster and Smith (1998), Fuchs (1982), Molenberghs and Goetghebeur (1997), Molenberghs et al. (1997), Park and Brown (1994), and Stokes et al. (2000).

Section 11.5: Markov Chains: Transitional Modeling

11.10. For statistical inference with Markov chains, see Andersen (1980, Sec. 7.7), Anderson and Goodman (1957), Billingsley (1961), Bishop et al. (1975, Chap. 7), and Kalbfleisch and Lawless (1985). See Conaway (1989), Stiratelli et al. (1984), and Ware et al. (1988) for other analyses focusing on the conditional dependence structure.

PROBLEMS

Applications

11.1 Refer to Table 8.3. Viewing the table as matched triplets, construct the marginal distribution for each substance. Find the sample proportions of students who used marijuana, alcohol, and cigarettes. Test the hypothesis of marginal homogeneity. Interpret results.

11.2 Refer to Table 9.1. Fit a marginal model to describe main effects of race, gender, and substance type (marijuana, alcohol, cigarettes) on whether a subject had used that substance. Summarize effects.

11.3 Refer to Problem 11.2. Further study shows evidence of an interaction between gender and substance type. Using GEE with exchangeable working correlation, the model fit for the probability π of using

a particular substance is

$$\text{logit}(\hat{\pi}) = -0.57 + 1.93S_1 + 0.86S_2 + 0.38R - 0.20G + 0.37G \times S_1 + 0.22G \times S_2,$$

where R, G, S_1, S_2 are dummy variables for race (1 = white), gender (1 = female), and substance type ($S_1 = 1$, $S_2 = 0$ for alcohol; $S_1 = 0, S_2 = 1$ for cigarettes; $S_1 = S_2 = 0$ for marijuana). Show that:

a. The estimated odds a nonwhite male has used marijuana are $\exp(-0.57) = 0.57$.

b. Given gender, the estimated odds a white subject used a given substance are 1.46 times the estimated odds for a black subject.

c. Given race, the estimated odds a female has used alcohol are 1.19 times the estimated odds for males; for cigarettes and for marijuana, the estimated odds ratios are 1.02 and 0.82.

d. Given race, the estimated odds a female has used alcohol (cigarettes) are 9.97 (2.94) times the estimated odds she has used marijuana.

e. Given race, the estimated odds a male has used alcohol (cigarettes) are 6.89 (2.36) times the estimated odds he has used marijuana. Interpret the interaction.

11.4 Refer to Table 11.2. Analyze the data using the scores $(1, 2, 4)$ for the week number, using ML or GEE. Interpret estimates and compare substantive results to those in the text with scores $(0, 1, 2)$.

11.5 Analyze Table 11.9 using a marginal logit model with age and maternal smoking as predictors. Compare interpretations to the Markov model of Section 11.5.5.

11.6 Table 11.10 refers to a three-period crossover trial to compare placebo (treatment A) with a low-dose analgesic (treatment B) and high-dose analgesic (treatment C) for relief of primary dysmenorrhea. Subjects in the study were divided randomly into six groups, the possible sequences for administering the treatments. At the end of each period, each subject rated the treatment as giving no relief (0) or some relief (1). Let $y_{i(k)t} = 1$ denote relief for subject i using treatment t ($t = A, B, C$), where subject i is nested in treatment sequence k ($k = 1, \ldots, 6$). Assuming common treatment effects for each sequence, and setting $\beta_A = 0$, obtain and interpret $\{\hat{\beta}_t\}$ (using ML or GEE) for the model

$$\text{logit}\left[P\left(Y_{i(k)t} = 1\right)\right] = \alpha_k + \beta_t.$$

How would you order the drugs, taking significance into account?

TABLE 11.10 Data for Problem 11.6

Treatment Sequence	Response Pattern for Treatments (A, B, C)							
	000	001	010	011	100	101	110	111
A B C	0	2	2	9	0	0	1	1
A C B	2	0	0	9	1	0	0	4
B A C	0	1	1	8	1	3	0	1
B C A	0	1	1	8	1	0	0	1
C A B	3	0	0	7	0	1	2	1
C B A	1	5	0	4	0	3	1	0

Source: Jones and Kenward (1987).

11.7 Table 11.11 is from a Kansas State University survey of 262 pig farmers. For the question "What are your primary sources of veterinary information?," the categories were (A) professional consultant, (B) veterinarian, (C) state or local extension service, (D) magazines, and (E) feed companies and reps. Farmers sampled were asked to select all relevant categories. The $2^5 \times 2 \times 4$ table shows the (yes, no) counts for each of these five sources cross-classified with the farmers' education (whether they had at least some college education) and size of farm (number of pigs marketed annually, in thousands).

TABLE 11.11 Data for Problem 11.7

			Response on D															
			A = yes								A = no							
			B = yes				B = no				B = yes				B = no			
			C = yes		C = no		C = yes		C = no		C = yes		C = no		C = yes		C = no	
Educ	Pigs	E	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
No	< 1	Y	1	0	0	0	0	0	0	0	2	1	1	2	1	1	5	3
		N	0	0	0	0	0	0	0	1	1	0	0	5	4	7	7	0
	1–2	Y	2	0	0	0	0	0	0	0	4	0	0	4	1	0	0	4
		N	0	0	0	0	0	0	0	0	0	0	0	5	0	3	4	0
	2–5	Y	3	0	0	0	0	0	0	0	3	0	0	1	2	0	1	1
		N	1	0	0	0	0	0	0	3	0	0	0	2	0	1	4	0
	> 5	Y	2	0	0	0	0	0	0	0	1	0	1	0	0	1	0	2
		N	1	0	0	2	1	0	1	6	0	1	1	1	0	0	6	0
Some	< 1	Y	3	0	0	0	0	0	0	0	4	0	1	1	0	0	2	11
		N	0	0	0	0	0	0	0	0	4	0	1	2	4	6	14	0
	1–2	Y	0	0	0	0	0	0	0	0	2	0	0	1	0	0	1	6
		N	0	0	0	0	1	0	0	1	2	1	0	4	2	7	14	0
	2–5	Y	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	3
		N	1	0	0	0	0	0	0	0	0	0	0	5	0	4	4	0
	> 5	Y	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	2
		N	1	1	0	0	0	1	0	10	0	0	0	4	1	2	4	0

Source: Data courtesy of Tom Loughin, Kansas State University.

a. Explain why it is not proper to analyze the data by fitting a multinomial model to the counts in the $2 \times 4 \times 5$ contingency table cross-classifying education by size of farm by the source of veterinary information, treating source as the response variable. (This table contains 453 positive responses of sources from the 262 farmers.)

b. For a farmer with education i and size of farm s, let $\pi_j(is)$ denote the probability of responding "yes" on the jth source. Table 11.12 shows output for using GEE with exchangeable working correlation to estimate parameters in the model lacking an education effect,

$$\text{logit}\big[\pi_j(is)\big] = \alpha_j + \beta_j s, \qquad s = 1, 2, 3, 4.$$

Explain how to interpret the working correlation matrix. Explain why the results suggest a strong positive size of farm effect for source A and perhaps a weak negative size effect of similar magnitude for C, D, and E.

c. Constraining $\beta_3 = \beta_4 = \beta_5$, the ML estimate of the common slope is -0.184 (SE = 0.063). Explain why it is advantageous to fit the marginal model simultaneously for all sources rather than separately to each. [Agresti and Liu (1999) and Loughin and Scherer (1998) discussed analyses for data of this form.]

TABLE 11.12 Output for Problem 11.7

Working Correlation Matrix

	Col1	Col2	Col3	Col4	Col5
Row1	1.0000	0.0997	0.0997	0.0997	0.0997
Row2	0.0997	1.0000	0.0997	0.0997	0.0997
Row3	0.0997	0.0997	1.0000	0.0997	0.0997
Row4	0.0997	0.0997	0.0997	1.0000	0.0997
Row5	0.0997	0.0997	0.0997	0.0997	1.0000

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter		Estimate	Std Error	Z	Pr>|Z|
source	1	−4.4994	0.6457	−6.97	<.0001
source	2	−0.8279	0.2809	−2.95	0.0032
source	3	−0.1526	0.2744	−0.56	0.5780
source	4	0.4875	0.2698	1.81	0.0708
source	5	−0.0808	0.2738	−0.30	0.7680
size*source	1	1.0812	0.1979	5.46	<.0001
size*source	2	0.0792	0.1105	0.72	0.4738
size*source	3	−0.1894	0.1121	−1.69	0.0912
size*source	4	−0.2206	0.1081	−2.04	0.0412
size*source	5	−0.2387	0.1126	−2.12	0.0341

TABLE 11.13 Output for Problem 11.8

Working Correlation Matrix

	Col1	Col2	Col3
Row1	1.0000	0.8173	0.8173
Row2	0.8173	1.0000	0.8173
Row3	0.8173	0.8173	1.0000

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Std Error	Z	Pr>|Z|
Intercept	−0.1253	0.0676	−1.85	0.0637
question 1	0.1493	0.0297	5.02	<.0001
question 2	0.0520	0.0270	1.92	0.0544
question 3	0.0000	0.0000	.	.
female	0.0034	0.0878	0.04	0.9688

11.8 Refer to Table 11.13 on attitudes toward legalized abortion. For the response Y_t (1 = support legalization, 0 = oppose) for question t ($t = 1, 2, 3$) and for gender g (1 = female, 0 = male), consider the model $\text{logit}[P(Y_t = 1)] = \alpha + \gamma g + \beta_t$ with $\beta_3 = 0$.

a. A GEE analysis using unstructured working correlation gives correlation estimates 0.826 for questions 1 and 2, 0.797 for 1 and 3, and 0.832 for 2 and 3. What does this suggest about a reasonable working correlation structure?

b. Table 11.13 shows a GEE analysis with exchangeable working correlation. Interpret effects.

c. Treating the three responses for each subject as independent observations and performing ordinary logistic regression, $\hat{\beta}_1 = 0.149$ (SE = 0.066), $\hat{\beta}_2 = 0.052$ (SE = 0.066), and $\hat{\gamma} = 0.004$ (SE = 0.054). Give a heuristic explanation of why within-subject standard errors are much larger than with GEE, yet the between-subject standard error is smaller.

11.9 Refer to the air pollution data in Table 11.7. Using ML or GEE, fit marginal logit models that assume **(a)** marginal homogeneity, **(b)** a linear effect of time, and **(c)** no pattern. Interpret and compare.

11.10 Refer to the clinical trials data in Table 12.5, analyzed with random effects models in Section 12.3.4. Use GEE methods to analyze them, treating each center as a correlated cluster.

11.11 Refer to Table 10.5. Using GEE methods with cumulative logits, compare the two marginal distributions. Compare results to those using ML in Section 10.3.2.

11.12 Refer to the 3^4 table on government spending in Table 8.19. Analyze these data with a marginal cumulative logit model. Interpret effects.

11.13 Refer to Table 11.4.

a. To compare effects while controlling for initial response, fit model (11.7), using scores {10, 25, 45, 75} for time to falling asleep. Also fit the interaction model, and describe the lack of fit. (Note that for the first two baseline levels, the active and placebo treatments have similar sample response distributions at the follow-up; at higher baseline levels, the active treatment seems more successful.)

b. Fit the interaction model

$$\text{logit}\big[P(Y_2 \le j)\big] = \alpha_j + \beta_1 x + \beta_2 y_1 + \beta_3 x y_1$$

that constrains effects $\{\beta_1 x + \beta_2 y_1 + \beta_3 x y_1\}$ to follow the pattern $(\tau, \tau, \lambda + \sigma, \lambda)$ for the active group and $(\tau, \tau, \sigma, 0)$ for the placebo group. Interpret $\hat{\lambda}$.

11.14 Find a marginal model with another type of logit that fits the insomnia data of Table 11.4 well. Interpret parameter estimates, and compare conclusions to those using cumulative logits.

11.15 Refer to Table 11.9. Combine the data for the two levels of maternal smoking. Does a first-order Markov chain model these data adequately? Find a loglinear model that does fit adequately.

11.16 Analyze Table 11.9 using a transitional model with two previous responses. Does it fit better than the first-order model of Section 11.5.5? Interpret.

11.17 Analyze Table 11.2 using a first-order transitional model. Compare interpretations to those in this chapter using marginal models.

11.18 Table 11.14 is from a longitudinal study of coronary risk factors in schoolchildren (Woolson and Clarke 1984). A sample of children aged 11–13 in 1977 were classified by gender and by relative weight (obese, not obese) in 1977, 1979, and 1981. Analyze these data.

TABLE 11.14 Data for Problem 11.18

	Responses[a]							
Gender	NNN	NNO	NON	NOO	ONN	ONO	OON	OOO
Male	119	7	8	3	13	4	11	16
Female	129	8	7	9	6	2	7	14

[a] NNN indicates not obese in 1977, 1979, and 1981; NNO indicates not obese in 1977 and 1979 but obese in 1981; and so on.

Source: Reproduced with permission from the Royal Statistical Society, London (Woolson and Clarke 1984).

11.19 Refer to the pig farmer survey of Problem 11.7 (Table 11.11). Analyze these data using marginal models with all the variables.

11.20 Refer to the cereal diet and cholesterol study of Problem 7.18 (Table 7.23). Analyze these data with marginal models.

Theory and Methods

11.21 Refer to Problem 11.1. Suppose that we expressed the data with a 3×2 partial table of drug-by-response for each subject, to use a generalized CMH procedure to test marginal homogeneity. Explain why the 911 + 279 subjects who make the same response for every drug have no effect on the test.

11.22 Let $y_{it} = 1$ or 0 for observation t on subject i, $i = 1, \ldots, n$, $t = 1, \ldots, T$. Let $y_{.t} = \Sigma_i y_{it}/n$, $y_{i.} = \Sigma_t y_{it}/T$, and $y_{..} = \Sigma_i \Sigma_t y_{it}/nT$.

a. Regard $\{y_{i+}\}$ as fixed. Suppose that each way to allocate the y_{i+} "successes" to y_{i+} of the observations is equally likely. Show that $E(Y_{it}) = y_{i.}$, $\text{var}(Y_{it}) = y_{i.}(1 - y_{i.})$, and $\text{cov}(Y_{it}, Y_{ik}) = -y_{i.}(1 - y_{i.})/(T - 1)$ for $t \neq k$. [*Hint:* The covariance is the same for any pair of cells in the same row, and $\text{var}(\Sigma_t Y_{it}) = 0$ since y_{i+} is fixed.]

b. Refer to part (a). For large n with independent subjects, explain why $(Y_{.1}, \ldots, Y_{.T})$ is approximately multivariate normal with pairwise correlation $\rho = -1/(T - 1)$. Conclude that Cochran's Q statistic (Cochran 1950)

$$Q = \frac{n^2(T-1)\Sigma_{t=1}^{T}(y_{.t} - y_{..})^2}{T\Sigma_{i=1}^{n} y_{i.}(1 - y_{i.})}$$

is approximately chi-squared with df $= (T - 1)$. [One way notes that if $(X_1, \ldots, X_T)$ is multivariate normal with common mean and common variance σ^2 and common correlation ρ for pairs (X_t, X_k), then $\Sigma(X_t - \bar{X})^2/\sigma^2(1 - \rho)$ is chi-squared with df $= (T - 1)$. See Bhapkar and Somes (1977) for slightly weaker conditions for a chi-squared limiting distribution for Q than those in part (a).]

c. Show that Q is unaffected by deleting cases in which $y_{i1} = \cdots = y_{iT}$.

11.23 Consider the model $\mu_i = \beta$, $i = 1, \ldots, n$, assuming that $v(\mu_i) = \mu_i$. Suppose that actually $\text{var}(Y_i) = \mu_i^2$. Using the univariate version of GEE described in Section 11.4, show that $u(\beta) = \Sigma_i(y_i - \beta)/\beta$ and $\hat{\beta} = \bar{y}$. Show that V in (11.10) equals β/n, the actual asymptotic variance (11.11) simplifies to β^2/n, and its consistent estimate is $\Sigma_i(y_i - \bar{y})^2/n^2$.

11.24 Repeat Problem 11.23 assuming that $v(\mu_i) = \sigma^2$ when actually $\text{var}(Y_i) = \mu_i$.

11.25 Consider the model $\mu_i = \beta$, $i = 1, \ldots, n$, for independent Poisson observations. For $\hat{\beta} = \bar{y}$, show that the model-based asymptotic variance estimate is $\bar{y}/n$, whereas the robust estimate of the asymptotic variance is $\sum_i (y_i - \bar{y})^2/n^2$. Which would you expect to be better (**a**) if the Poisson model holds, and (**b**) if there is severe overdispersion?

11.26 Show that (11.10) is equivalent to the formula for the large-sample covariance of the ML estimator in a GLM, estimated by (4.28).

11.27 **a.** For a univariate response, how is quasi-likelihood (QL) inference different from ML inference? When are they equivalent?

b. Explain the sense in which GEE methodology is a multivariate version of QL.

c. Summarize the advantages and disadvantages of the QL approach.

d. Describe conditions under which GEE parameter estimators are consistent and conditions under which they are not. For conditions in which they are consistent, explain why.

11.28 Formulate a model using adjacent-categories logits or continuation-ratio logits that is analogous to (11.4). Interpret parameters.

11.29 Refer to the analysis of mean time to falling asleep at the end of Section 11.2.3. Explain how to calculate SE for the difference between the difference of means reported there. (Note that one difference uses paired samples and the other uses independent samples.)

11.30 What is wrong with this statement?: "For a first-order Markov chain, Y_t is independent of Y_{t-2}."

11.31 Suppose that loglinear model $(Y_0, Y_1, \ldots, Y_T)$ holds. Is this a Markov chain?

11.32 Gamblers A and B have a total of I dollars. They play games of pool repeatedly. Each game they each bet \$1, and the winner takes the other's dollar. The outcomes of the games are statistically independent, and A has probability π and B has probability $1 - \pi$ of winning any game. Play stops when one player has all the money. Let Y_t denote A's monetary total after t games.

a. Show that $\{Y_t\}$ is a first-order Markov chain.

b. State the transition probability matrix. (For this *gambler's ruin* problem, 0 and I are *absorbing* states. Eventually, the chain enters one of these and stays. The other states are *transient*.)

11.33 A first-order Markov chain has *stationary* (or *time-homogeneous*) transition probabilities if the one-step transition probability matrices are identical, that is, if for all i and j,

$$\pi_{j|i}(1) = \pi_{j|i}(2) = \cdots = \pi_{j|i}(T) = \pi_{j|i}.$$

Let X, Y, and Z denote the classifications for the $I \times I \times T$ table consisting of $\{n_{ij}(t),\ i = 1, \ldots, I,\ j = 1, \ldots, I,\ t = 1, \ldots, T\}$.

a. Explain why all transition probabilities are stationary if expected frequencies for this table satisfy loglinear model (XY, XZ). [Thus, the likelihood-ratio statistic for testing stationary transition probabilities equals G^2 for testing fit of model (XY, XZ).]

b. Let $n_{ij} = \sum_t n_{ij}(t)$. Under the assumption of stationary transition probabilities, show how the likelihood in (11.14) simplifies, and show that the ML estimators are

$$\hat{\pi}_{j|i} = n_{ij}/n_{i+}.$$

c. For a Markov chain with stationary transition probabilities, let y_{ijk} denote the number of transitions from i to j to k over two successive steps. For $\{y_{ijk}\}$, argue that the goodness of fit of loglinear model (Y_1Y_2, Y_2Y_3) tests that the chain is first order against the alternative that it is second order (Anderson and Goodman 1957).

Categorical Data Analysis, Second Edition. Alan Agresti

ISBN: 0-471-36093-7

CHAPTER 12

Random Effects: Generalized Linear Mixed Models for Categorical Responses

In Chapter 11 we noted that observations often occur in clusters. For instance, cluster i might consist of repeated measurements on subject i or observations for all subjects in family i. Observations within a cluster tend to be more alike than observations from different clusters. Thus, they are usually positively correlated. Ordinary analyses that ignore the correlation and treat within-cluster observations the same as between-cluster observations produce invalid standard errors.

In Chapter 11 we focused on modeling the *marginal* distributions of clustered responses, treating the joint dependence structure as a nuisance. In this chapter we present an alternative approach using cluster-level terms in the model. These terms take the same value for each observation in a cluster but different values for different clusters. They are unobserved and, when treated as varying randomly among clusters, are called *random effects*. In Section 10.2.4 we introduced this approach in a model for matched pairs. The models have *conditional* interpretations, referred to as *subject-specific* when each cluster is a subject. This contrasts with marginal models, which have *population-averaged* interpretations.

Random effects models for normal responses are well established. By contrast, only recently have random effects been used much in models for categorical data. In this chapter we extend generalized linear models to include random effects. In Section 12.1 we introduce this extension, the *generalized linear mixed model*. In Section 12.2 we discuss an important special case for binary data, the *logistic-normal model*. Several examples are shown in Section 12.3. Section 12.4 covers extensions for multinomial responses, and Section 12.5 covers models with multivariate random effects. In Section 12.6 we discuss model fitting, assuming normality for the random effects. Parts of this chapter are from Agresti et al. (2000).

12.1 RANDOM EFFECTS MODELING OF CLUSTERED CATEGORICAL DATA

Parameters that describe a factor's effects in ordinary linear models are called *fixed effects*. They apply to *all* categories of interest, such as genders, age groupings, or treatments. By contrast, random effects usually apply to a *sample*. For a study using a sample of clinics, for example, the model treats observations from a given clinic as a cluster, and it has a random effect for each clinic.

GLMs extend ordinary regression by allowing nonnormal responses and a link function of the mean. The *generalized linear mixed model* (GLMM) is a further extension that permits random effects as well as fixed effects in the linear predictor.

12.1.1 Generalized Linear Mixed Model

Let y_{it} denote observation t in cluster i, $t = 1, \ldots, T_i$. As in the GEE analyses in Chapter 11, the number of observations may vary by cluster. In a longitudinal study, even if clusters have equal size, many of them may have missing observations. Let $\mathbf{x}_{it}$ denote a column vector of values of explanatory variables, for fixed effect model parameters $\boldsymbol{\beta}$. Let $\mathbf{u}_i$ denote the vector of random effect values for cluster i. This is common to all observations in the cluster. Let $\mathbf{z}_{it}$ denote a column vector of their explanatory variables. Often, the random effect is univariate.

Conditional on $\mathbf{u}_i$, a GLMM resembles an ordinary GLM. Let $\mu_{it} = E(Y_{it} | \mathbf{u}_i)$. The linear predictor for a GLMM has the form

$$g(\mu_{it}) = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{u}_i \tag{12.1}$$

for link function $g(\cdot)$. The random effect vector $\mathbf{u}_i$ is assumed to have a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$. The covariance matrix $\boldsymbol{\Sigma}$ depends on unknown *variance components* and possibly also correlation parameters.

Denote $\text{var}(Y_{it} | \mathbf{u}_i) = \phi_{it} v(\mu_{it})$, where the variance function $v(\cdot)$ describes how the (conditional) variance depends on the mean. As in Section 4.4, often $\phi_{it} = 1$ or $\phi_{it} = \phi/\omega_{it}$, where ω_{it} is a known weight (e.g., number of trials for a binomial count) and ϕ is an unknown dispersion parameter. Conditional on $\mathbf{u}_i$, the model treats $\{y_{it}\}$ as independent over i and t. As discussed in Section 10.2.2, the variability among $\mathbf{u}_i$ induces a nonnegative association among the responses, for the marginal distribution averaged over the subjects. This is caused by the shared random effect $\mathbf{u}_i$ for each observation in a cluster.

In (12.1), the random effect enters the model on the same scale as the predictor terms. This is convenient but also natural for many applications. For instance, random effects sometimes represent heterogeneity caused by

omitting certain explanatory variables. Consider the special case with univariate random effect and $z_{it} = 1$. With u_i replaced by $u_i^*\sigma$ where $\{u_i^*\}$ are $N(0, 1)$, the GLMM has the form

$$g(\mu_{it}) = \mathbf{x}_{it}'\boldsymbol{\beta} + u_i^*\sigma.$$

This has the form of an ordinary GLM with unobserved values $\{u_i^*\}$ of a particular covariate. Thus, random effects models relate to methods of dealing with unmeasured predictors and other forms of missing data. The random effects part of the linear predictor reflects terms that would be in the fixed effects part if those explanatory variables had been included. Random effects also sometimes represent random measurement error in the explanatory variables. If we replace a particular predictor x_{it} by $x_{it}^* + \epsilon_i$, with x_{it}^* the true value and ϵ_i the measurement error, then ϵ_i times the regression parameter can be absorbed in the random effects term. Related to these motivations, random effects also provide a mechanism for explaining overdispersion in basic models not having those effects (Breslow and Clayton 1993).

12.1.2 Logit GLMM for Binary Matched Pairs

We illustrate the GLMM expression (12.1) using a simple case, that of binary matched pairs. The data form two dependent binomial samples (Section 10.1). Cluster i consists of the responses (y_{i1}, y_{i2}) for matched pair i. Observation t in cluster i has $y_{it} = 1$ (a success) or 0 (a failure), $t = 1, 2$.

In Section 10.2.2 we introduced the model (Cox 1958b, Rasch 1961)

$$\text{logit}\big[P(Y_{it} = 1)\big] = \alpha_i + \beta x_t \tag{12.2}$$

where $x_1 = 0$ and $x_2 = 1$. For it, β is a cluster-specific log odds ratio. That section treated α_i as a fixed effect and eliminated it using conditional ML. An equivalent representation of (12.2) is

$$\text{logit}\big[P(Y_{i1} = 1 \,|\, u_i)\big] = \alpha + u_i, \qquad \text{logit}\big[P(Y_{i2} = 1 \,|\, u_i)\big] = \alpha + \beta + u_i, \tag{12.3}$$

where $u_i = \alpha_i - \alpha$ for some constant α. Now, we treat u_i as a random effect for cluster i, with $\{u_i\}$ independent from a $N(0, \sigma^2)$ distribution with σ unknown. Conditionally on u_i, we assume that y_{i1} and y_{i2} are independent.

Model (12.3) is the special case of (12.1) in which $\mu_{it} = P(Y_{it} = 1 \,|\, u_i)$, $g(\cdot)$ is the logit link, $\boldsymbol{\beta}' = (\alpha, \beta)$, $\mathbf{x}_{i1}' = (1, 0)$ and $\mathbf{x}_{i2}' = (1, 1)$ for all i, and $z_{it} = 1$ for all i and t. The univariate random effect adjusts the intercept but does not modify the fixed effect. A GLMM with random effect of this form is called a *random intercept* model. Instead of the usual fixed intercept α, it has a random intercept $\alpha + u_i$.

Let $Y_1 = \sum_i y_{i1}$ and $Y_2 = \sum_i y_{i2}$. Marginally, Y_1 is binomial with n trials and parameter $E\{\exp(\alpha + U)/[1 + \exp(\alpha + U)]\}$, and Y_2 is binomial with parameter $E\{\exp(\alpha + \beta + U)/[1 + \exp(\alpha + \beta + U)]\}$. The expectations refer to U, a $N(0, \sigma^2)$ random variable. The model implies a nonnegative correlation between Y_1 and Y_2, with greater association resulting from greater heterogeneity (i.e., larger σ). Clusters with a large positive u_i have a relatively large $P(Y_{it} = 1 \mid u_i)$ for each t, whereas clusters with a large negative u_i have a relatively small $P(Y_{it} = 1 \mid u_i)$ for each. For this model, Y_1 and Y_2 are independent only if $\sigma = 0$.

A 2×2 population-averaged table with (success, failure) for both the row and column categories summarizes the number of observations for which $(y_{i1}, y_{i2}) = (1, 1)$, $(1, 0)$, $(0, 1)$, or $(0, 0)$. Let $\{n_{ab}\}$ denote these counts. Table 12.1, analyzed first in Section 10.1, is an example. Let $\{\hat{\mu}_{ab}\}$ denote marginal fitted values for model (12.3). We defer discussion of model fitting until Section 12.6. However, model (12.3) is a rare instance in which the fixed effect in a random effects model has a closed-form ML estimate,

$$\hat{\beta} = \log(\hat{\mu}_{21}/\hat{\mu}_{12}).$$

When the sample log odds ratio $\log(n_{11}n_{22}/n_{12}n_{21}) \geq 0$, then $\{\hat{\mu}_{ab} = n_{ab}\}$ and $\hat{\beta} = \log(n_{21}/n_{12})$. This is the same as the conditional ML estimate (Section 10.2.3). Neuhaus et al. (1994) showed that this is true for *any* parametric choice of random effects distribution for which the model (12.3) can generate $\{n_{ab}\}$ as fitted values. Lindsay et al. (1991) showed that this estimate also results with a nonparametric approach discussed in Section 13.2.4. The model implies that the true log odds ratio for this 2×2 table is at least 0. When $\log(n_{11}n_{22}/n_{12}n_{21}) < 0$, however, then $\hat{\sigma} = 0$ and the fitted values $\{\hat{\mu}_{ab} = n_{a+}n_{+b}/n\}$ satisfy independence. Then, $\hat{\beta}$ is identical to the estimate for the marginal model (10.6) by which β is the difference between logits for the two marginal distributions, namely $\hat{\beta} = \log[(n_{2+}n_{+1})/(n_{1+}n_{+2})]$.

12.1.3 Ratings of Prime Minister Revisited

For Table 12.1, the ML fit of model (12.3), treating $\{u_i\}$ as normal, yields $\hat{\beta} = \log(86/150) = -0.556$ (SE = 0.135), with $\hat{\sigma} = 5.16$. This is identical to the conditional ML estimate (10.10), with standard error $[(1/86) + (1/150)]^{1/2}$. For a given subject, the estimated odds of approval at the second

TABLE 12.1 Rating of Performance of Prime Minister

First Survey	Second Survey: Approve	Second Survey: Disapprove	Total
Approve	794	150	944
Disapprove	86	570	656
Total	880	720	1600

survey equal $\exp(-0.556) = 0.57$ times those at the first survey. The large $\hat{\sigma}$ reflects the very strong association between the two responses, with sample odds ratio 35.1.

12.1.4 Extension: Rasch Model and Item Response Models

An extension of the logit matched-pairs model (12.3) allows $T > 2$ observations in each cluster. The random intercept model then has form

$$\text{logit}\big[P(Y_{it} = 1 \,|\, u_i)\big] = u_i + \beta_t, \tag{12.4}$$

where $\{u_i\}$ are independent $N(0, \sigma^2)$. Equivalently, the model can add an intercept α or let $E(u_i) = \alpha$, but then identifiability requires a constraint such as $\beta_T = 0$.

Early applications of this GLMM were in psychometrics. The model describes responses to a battery of T questions on an exam. The probability $P(Y_{it} = 1 \,|\, u_i)$ that subject i makes the correct response on question t depends on the overall ability of subject i, characterized by u_i, and the easiness of question t, characterized by β_t. Such models are called *item-response models*. The logit form (12.4) is called the *Rasch model* (Rasch 1961). In estimating $\{\beta_t\}$, Rasch treated $\{u_i\}$ as fixed effects and used conditional ML, as outlined in Section 10.2.3 for matched pairs. Later authors used the normal random effects approach for this model and the model with probit link (e.g., Bock and Aitkin 1981).

The $\{\beta_t\}$ in the Rasch model differ from parameters in corresponding marginal models such as (11.1), since the effects are subject specific. The Rasch model refers to a $T \times 2 \times n$ table of observation by outcome by subject, whereas the marginal model refers to the $T \times 2$ observation-by-outcome table of the T marginal distributions, collapsed over subjects. For observations s and t for a given subject i with model (12.4),

$$\beta_s - \beta_t = \text{logit}\big[P(Y_{is} = 1|u_i)\big] - \text{logit}\big[P(Y_{it} = 1|u_i)\big],$$

which is a log odds ratio conditional on the subject. By contrast, the corresponding population-averaged effect in marginal model (11.1) is

$$\beta_s - \beta_t = \text{logit}\big[P(Y_{hs} = 1)\big] - \text{logit}\big[P(Y_{it} = 1)\big],$$

with subject h randomly selected for observation s and subject i randomly selected for observation t (i.e., h and i are *independent* observations).

12.1.5 Random Effects versus Conditional ML Approaches

Suppose that one treated $\{u_i\}$ in model (12.4) as fixed effects instead of random effects. Then, consider ordinary ML estimation of $\{\beta_t\}$ and $\{u_i\}$. As n increases, so does the number of parameters, since each subject has a u_i.

Even though the number of $\{\beta_t\}$ does not increase as n does, the ordinary ML estimators $\{\hat{\beta}_t\}$ are not consistent. This happens in many models when the number of parameters has an order similar to that of the number of subjects. Asymptotic optimality properties of ML estimators, such as consistency, require the number of parameters to be fixed as n increases. For model (12.4), ML estimators of $\{\beta_t\}$ have bias of order $T/(T-1)$ (Andersen 1980, pp. 244–245). For the matched-pairs model (12.2), for instance, $\hat{\beta} \to 2\beta$ in probability (Problem 10.24).

For this reason, the preferable approach for the fixed effects model is *conditional ML*. One eliminates $\{u_i\}$ by conditioning on their sufficient statistics $\{S_i = \sum_t y_{it},\ i = 1, \ldots, n\}$. In the item response context, these are the numbers of correct responses for each subject. Conditional on $\{S_i\}$, the distribution of $\{y_{it}\}$ is independent of $\{u_i\}$. Maximizing the resulting likelihood then yields consistent estimators of $\{\beta_t\}$. The analysis generalizes the one in Section 10.2.3 for the subject-specific logistic model (10.8) for matched pairs. See Andersen (1980) for details.

Compared with the random effects approach, the conditional ML approach has certain advantages. One does not need to assume a parametric distribution for $\{u_i\}$. It is difficult to check this assumption in the random effects approach. Conditional ML is also appropriate with retrospective sampling. In that case, bias can occur with a random effects approach because the clusters are not randomly sampled (Neuhaus and Jewell 1990b).

However, the conditional ML approach has severe disadvantages. It is restricted to the canonical link (the logit), for which reduced sufficient statistics exist for $\{u_i\}$. More important, as discussed in Section 10.2.7, it is restricted to inference about within-cluster fixed effects. The conditioning removes the source of variability needed for estimating between-cluster effects in models with explanatory variables such as those considered next. Also, this approach does not provide information about $\{u_i\}$, such as predictions of their values and estimates of their variability or of the probabilities they determine. Finally, in more general models with covariates, conditional ML can be less efficient than the random effects approach for estimating the fixed effects (see Note 12.2).

12.2 BINARY RESPONSES: LOGISTIC-NORMAL MODEL

The item response model (12.4) with random intercept is a special case of an important class of random effects models for binary data called *logistic-normal models*. With univariate random effect, the model form is

$$\text{logit}\big[P(Y_{it} = 1 \,|\, u_i)\big] = \mathbf{x}_{it}'\boldsymbol{\beta} + u_i \tag{12.5}$$

where $\{u_i\}$ are independent $N(0, \sigma^2)$ variates. This is the special case of the GLMM (12.1) in which $g(\cdot)$ is the logit link and the random effects structure

simplifies to a random intercept. The logistic-normal model has a long history, dating at least to Cox (1970, Prob. 20 in that text) for the matched-pairs model (12.3) and Pierce and Sands (1975).

More generally, the link function in model (12.5) can be an arbitrary inverse cdf. For such models, Y_{is} and Y_{it} are treated conditionally (given u_i) as independent but are marginally nonnegatively correlated. Let Φ denote the cdf that is the inverse link function. Then, for $s \neq t$,

$$\begin{aligned}\operatorname{cov}(Y_{is}, Y_{it}) &= E\big[\operatorname{cov}(Y_{is}, Y_{it} \mid u_i)\big] + \operatorname{cov}\big[E(Y_{is} \mid u_i), E(Y_{it} \mid u_i)\big] \\ &= 0 + \operatorname{cov}\big[\Phi(\mathbf{x}'_{is}\boldsymbol{\beta} + u_i), \Phi(\mathbf{x}'_{it}\boldsymbol{\beta} + u_i)\big]. \qquad (12.6)\end{aligned}$$

The functions in the last covariance term are both monotone increasing in u_i, and hence are nonnegatively correlated. For common predictor value $\mathbf{x}$ at each t, the joint distribution for the model is exchangeable. This is often plausible for clustered data. In longitudinal studies, however, observations closer together in time may tend to be more highly correlated.

Usually, the main focus in using a GLMM is inference about the fixed effects. The random effects part of the model is a mechanism for representing how the positive correlation occurs between observations within a cluster. Parameters pertaining to the random effects may themselves be of interest, however. For instance, the estimate $\hat{\sigma}$ of the standard deviation of a random intercept may be a useful summary of the degree of heterogeneity of a population.

12.2.1 Interpreting Heterogeneity in Logistic-Normal Models

When $\sigma = 0$, the logistic-normal model (12.5) simplifies to the ordinary logistic regression model treating all observations as independent. When $\sigma > 0$, how can we interpret the variability in effects this model implies?

Consider observation y_{it} at setting $\mathbf{x}_{it}$ of predictors and observation y_{hs} at setting $\mathbf{x}_{hs}$. Their log odds ratio is

$$\operatorname{logit}\big[P(Y_{it} = 1 \mid u_i)\big] - \operatorname{logit}\big[P(Y_{hs} = 1 \mid u_h)\big] = (\mathbf{x}_{it} - \mathbf{x}_{hs})'\boldsymbol{\beta} + (u_i - u_h).$$

We cannot observe $(u_i - u_h)$, which has a $N(0, 2\sigma^2)$distribution. However, $100(1 - \alpha)\%$ of those log odds ratios fall within

$$(\mathbf{x}_{it} - \mathbf{x}_{hs})'\boldsymbol{\beta} \pm z_{\alpha/2}\sqrt{2}\,\sigma. \qquad (12.7)$$

When $\sigma = 0$, $(\mathbf{x}_{it} - \mathbf{x}_{hs})'\boldsymbol{\beta}$ is the usual form of log odds ratio for a model without random effects. When $\sigma > 0$, $(\mathbf{x}_{it} - \mathbf{x}_{hs})'\boldsymbol{\beta}$ is the log odds ratio for two observations in the same cluster $(h = i)$ or with the same random effect value. Suppose that $\mathbf{x}_{it} = \mathbf{x}_{hs}$ for observations from different clusters. Then, using $z_{0.25} = 0.674$, the middle 50% of the log odds ratios fall within

$\pm 0.674\sqrt{2}\,\sigma = \pm 0.95\sigma$. Hence, the median odds ratio between the observation with higher random effect and the observation with lower random effect equals $\exp(0.95\sigma)$. With a single predictor and $x_{it} - x_{hs} = 1$, the median such odds ratio equals $\exp(\beta + 0.95\sigma)$. Larsen et al. (2000) presented related interpretations.

12.2.2 Connections between Conditional Models and Marginal Models

The fixed effects parameters $\boldsymbol{\beta}$ in GLMMs have conditional intepretations, given the random effect. Those fixed effects are of two types. First, consider an explanatory variable that varies in value among observations in a cluster. For instance, in a crossover study comparing T drugs, for each subject the drug taken varies from observation to observation in that subject's cluster of T observations. For such an explanatory variable, its coefficient in the model refers to the effect on the response of a within-cluster (e.g., subject-specific) 1-unit increase of that predictor. The random effect as well as other explanatory variables in the model are constant while that predictor increases by 1. The effect of that explanatory variable is a "within-cluster" or "within-subject" one.

Second, consider an explanatory variable with constant value among observations in a cluster. An example is gender when each subject forms a cluster. For such an explanatory variable, its coefficient refers to the effect on the response of a "between-cluster" 1-unit increase of that predictor. An example is a comparison of females and males using a dummy variable and its coefficient. However, this fixed effect in the GLMM applies only when the random effect (as well as other explanatory variables in the model) takes the same value in both groups: for instance, a male and a female with the same value for their random effects.

It is in this sense that random effects models are conditional models, as both within- and between-cluster effects apply conditional on the random effect value. By contrast, effects in marginal models are averaged over all clusters (i.e., population averaged), so those effects do not refer to a comparison at a fixed value of a random effect. In fact, a fundamental difference between the two model types is that when the link function is nonlinear, such as the logit, the population-averaged effects of marginal models often are smaller than the cluster-specific effects of GLMMs.

Specifically, the GLMM (12.1) refers to the conditional mean, $\mu_{it} = E(Y_{it} \mid \mathbf{u}_i)$. By inverting the link function,

$$E(Y_{it} \mid \mathbf{u}_i) = g^{-1}(\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{u}_i).$$

Marginally, averaging over the random effects, the mean is

$$E(Y_{it}) = E\big[E(Y_{it} \mid \mathbf{u}_i)\big] = \int g^{-1}(\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{u}_i) f(\mathbf{u}_i; \boldsymbol{\Sigma})\, d\mathbf{u}_i,$$

where $f(\mathbf{u}; \boldsymbol{\Sigma})$ is the $N(\mathbf{0}, \boldsymbol{\Sigma})$ density function for the random effects. For the identity link,

$$E(Y_{it}) = \int (\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{u}_i) f(\mathbf{u}_i; \boldsymbol{\Sigma})\, d\mathbf{u}_i = \mathbf{x}'_{it}\boldsymbol{\beta}\ .$$

The marginal model has the same model form and effects $\boldsymbol{\beta}$. This is not true for other links. For instance, for the logistic-normal model (12.5),

$$E(Y_{it}) = E\left[\frac{\exp(\mathbf{x}'_{it}\boldsymbol{\beta} + u_i)}{1 + \exp(\mathbf{x}'_{it}\boldsymbol{\beta} + u_i)}\right].$$

This expectation does not have form $\exp(\mathbf{x}'_{it}\boldsymbol{\beta})/[1 + \exp(\mathbf{x}'_{it}\boldsymbol{\beta})]$ except when u_i has a degenerate distribution ($\sigma = 0$).

Approximate relationships exist between estimates from the two model types. In the logistic-normal case with effect $\boldsymbol{\beta}$ and small σ, Zeger et al. (1988) showed that

$$E(Y_{it}) \approx \exp(c\mathbf{x}'_{it}\boldsymbol{\beta})/\left[1 + \exp(c\mathbf{x}'_{it}\boldsymbol{\beta})\right], \tag{12.8}$$

where $c = [1 + 0.6\sigma^2]^{-1/2}$. Since the effect in the marginal model multiplies that of the conditional model by about c, it is typically smaller in absolute value. The discrepancy increases as σ increases. For $\boldsymbol{\beta}$ near 0, Neuhaus et al. (1991) showed that the marginal model effect is approximately $\boldsymbol{\beta}(1 - \rho)$, where $\rho = \text{corr}(Y_{it}, Y_{is})$ at $\boldsymbol{\beta} = \mathbf{0}$. Again, the discrepancy increases as σ increases, since ρ increases with σ.

For Table 12.1 on ratings of the prime minister, the ML estimate for model (12.3) is $\hat{\beta} = -0.556$, with $\hat{\sigma} = 5.16$ for variability of $\{u_i\}$. Approximation (12.8) suggests that $\hat{\beta} = -0.556$ with $\hat{\sigma} = 5.16$ corresponds to a marginal estimate of about $[1 + 0.6(5.16)^2]^{-1/2}(-0.556) = -0.135$. The actual marginal estimate is the log odds ratio for the sample marginal distributions, equaling

$$\log\left[(880/720)/(944/656)\right] = -0.163.$$

In fact, the marginal effect is much smaller than the conditional effect, but this approximation connecting the two estimates works better for smaller $\hat{\sigma}$. At $\beta = 0$, the fit of the model is that of the symmetry model, for which $\hat{\mu}_{12} = \hat{\mu}_{21} = (n_{12} + n_{21})/2$. The correlation for that 2×2 table equals 0.699, from which the conditional estimate of -0.556 suggests a marginal estimate of $-0.556(1 - 0.699) = -0.167$, very close to the actual value of -0.163.

Figure 12.1 illustrates why the marginal effect is smaller than the conditional effect. For a single explanatory variable x, the figure shows subject-specific curves for $P(Y_{it} = 1|u_i)$ for several subjects when considerable heterogeneity exists. This corresponds to a relatively large σ for random effects.

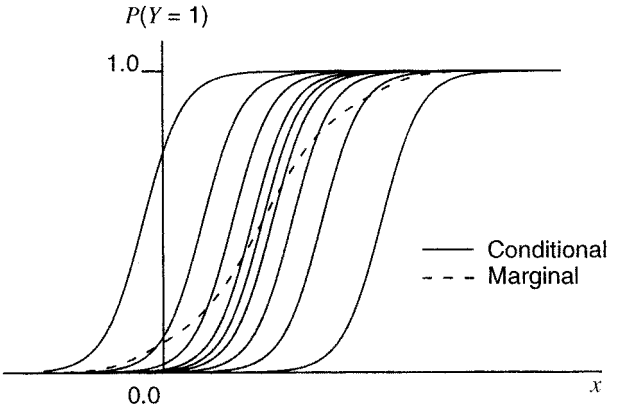


FIGURE 12.1 Logistic random-intercept model, showing the conditional (subject-specific) curves and the marginal (population-averaged) curve averaging over these.

At any fixed value of x, variability occurs in the conditional means, $E(Y_{it} \mid u_i) = P(Y_{it} = 1 \mid u_i)$. The average of these is the marginal mean, $E(Y_{it})$. These averages for various x values yield the superimposed curve. It has a shallower slope. In fact, it does not exactly follow the logistic formula. Similar remarks apply to other GLMMs. For the probit link with binary data, however, the conditional probit model with normal random effect does imply a marginal model of probit form (Problem 12.29). With univariate random intercept, the marginal effect equals the conditional effect multiplied by $[1 + \sigma^2]^{-1/2}$ (Zeger et al. 1988). In Section 13.5.1 we explore the conditional–marginal connection for loglinear GLMMs.

12.2.3 Comments about Conditional versus Marginal Models

Random effects models describe conditional (subject-specific) effects, whereas marginal models describe population-averaged effects. Some statisticians prefer one of these types, but most feel that both are useful, depending on the application.

The conditional modeling approach is preferable if one wants to specify a mechanism that could generate positive association among clustered observations, estimate cluster-specific effects, estimate their variability, or model the joint distribution. Latent variable constructions used to motivate model forms (e.g., the tolerance motivation for binary models of Section 6.6.1 and the related threshold motivation in Problem 6.28 and utility motivation in Problem 6.29) usually apply more naturally at the cluster level than at the marginal level. Given a conditional model, one can recover information about marginal distributions. That is, a conditional model implies a marginal model,

but a marginal model does not itself imply a conditional model (although see Note 12.10 for an implicit connection).

In many surveys or epidemiological studies, a goal is to compare the relative frequency of occurrence of some outcome for different groups in a population. Then, quantities of primary interest include between-group odds ratios among marginal probabilities for the different groups. That is, effects of interest are between-cluster rather than within-cluster. When marginal effects are the main focus, it is usually simpler and may be preferable to model the margins directly. One can then parameterize the model so that regression parameters have a direct marginal interpretation. Developing a more detailed model of the joint distribution that generates those margins, as a random effects model does, provides greater opportunity for misspecification. For instance, with longitudinal data the assumption that observations are independent, given the random effect, need not be realistic. With the marginal model approach, we showed in Chapter 11 that ML is sometimes possible but that the GEE approach is computationally simpler and more versatile. A drawback of the GEE approach is that it does not explicitly model random effects and therefore does not allow these effects to be estimated. In addition, likelihood-based inferences are not possible because the joint distribution of the responses is not specified.

In Section 12.2.2 it was noted that conditional effects are usually larger than marginal effects, and increase as variance components increase. Usually, though, the significance of an effect (e.g., as measured by the ratio of estimate to standard error) is similar in the two model types. If one effect seems more important than another in a conditional model, the same is usually true with a marginal model. So the choice of the model is usually not crucial to inferential conclusions.

This statement requires a caveat, however, since sizes of effects in marginal models depend on the degree of heterogeneity in conditional models. In comparing effects for two groups or two variables that have quite different variance components, relative sizes of effects will differ for marginal and conditional models. From (12.8), with binary data the attenuation from the conditional to the marginal effect will tend to be greater for the group having the larger variance component. For instance, suppose that two groups, one young in age and the other elderly, both show the same conditional effect in a crossover study comparing two drugs. If the elderly group has more heterogeneity on the response, their marginal effect may be smaller than that for the younger group. The marginal effects differ even though the conditional effects are the same, because of the greater variance component for the elderly. In such cases, the conditional effect (appropriately modeled) may have more relevance.

Finally, with either marginal or conditional models, missing data are a common problem with multivariate responses. Unless data are missing at random, potential bias occurs in ML inference. GEE methods usually require

the stronger condition that data are missing completely at random (Section 11.4.5). Thus, modeling missingness or conducting a sensitivity study to discern its potential effects can be an important component of an analysis.

Regardless of the choice of paradigm, it is a challenge for statisticians even to explain to practitioners why marginal and conditional effects differ with a nonlinear link function. Graphics such as Figure 12.1 can help. Neuhaus (1992) and Pendergast et al. (1996) surveyed ways of analyzing clustered binary data, including conditional and marginal models. Agresti and Natarajan (2001) surveyed conditional and marginal modeling of clustered ordinal data.

12.3 EXAMPLES OF RANDOM EFFECTS MODELS FOR BINARY DATA

In the next three sections we present a variety of examples of random effects models. In this section we consider binary responses.

12.3.1 Small-Area Estimation of Binomial Proportions

Small-area estimation refers to estimation of parameters for a large number of geographical areas when each has relatively few observations. For instance, one might want county-specific estimates of characteristics such as the unemployment rate or the proportion of families having health insurance coverage. With a national or statewide survey, some counties may have few observations. Then, sample proportions in the counties may poorly estimate the true countywide proportions. Random effects models that treat each county as a cluster can provide improved estimates. In assuming that the true proportions vary according to some distribution, the fitting process "borrows from the whole"—it uses data from all the counties to estimate the proportion in any given one.

Let π_i denote the true proportion in area i, $i = 1, \ldots, n$. These areas may be all the ones of interest, or only a sample. Let $\{y_i\}$ denote independent bin(T_i, π_i) variates; that is, $y_i = \sum_{t=1}^{T_i} y_{it}$, where $\{y_{it},\ t = 1, \ldots, T_i\}$ are independent with $P(Y_{it} = 1) = \pi_i$ and $P(Y_{it} = 0) = 1 - \pi_i$. The sample proportions $\{p_i = y_i/T_i\}$ are ML estimates of $\{\pi_i\}$ for the fixed-effects model

$$\text{logit}(\pi_i) = \alpha + \beta_i, \qquad i = 1, \ldots, n.$$

This model is saturated, having n nonredundant parameters (with a constraint such as $\sum_i \beta_i = 0$) for the n binomial observations.

For small $\{T_i\}$, $\{p_i\}$ have large standard errors. Thus, $\{p_i\}$ may display much more variability than $\{\pi_i\}$, especially when $\{\pi_i\}$ are similar. Then, it is helpful

to shrink $\{p_i\}$ toward their overall mean. One can accomplish this with the random effects model

$$\text{logit}\big[P(Y_{it} = 1 \,|\, u_i)\big] = \alpha + u_i, \tag{12.9}$$

where $\{u_i\}$ are independent $N(0, \sigma^2)$ variates. This model is a logit analog of one-way random effects ANOVA. When $\sigma = 0$, all π_i are identical.

For this model,

$$\hat{\pi}_i = \exp(\hat{\alpha} + \hat{u}_i)/\big[1 + \exp(\hat{\alpha} + \hat{u}_i)\big].$$

This estimate differs from the sample proportion p_i. If $\hat{\sigma} = 0$, then all $\hat{u}_i = 0$. Then, the random effects estimate of each π_i is $(\sum_{i=1}^{n}\sum_{t=1}^{T_i} y_{it})/(\sum_i T_i)$, the overall sample proportion after pooling all n samples. When truly all π_i are equal, this is a much better estimator of that common value than the sample proportion from a single sample.

Generally, the random effects model estimators shrink the separate sample proportions toward the overall sample proportion. The amount of shrinkage decreases as $\hat{\sigma}$ increases. The shrinkage also decreases as the $\{T_i\}$ grow; as each sample has more data, we put more trust in the separate sample proportions. The predicted random effect $\hat{u}_i$ is the estimated mean of the distribution of u_i, given the data (see Section 12.6.7). This prediction depends on all the data, not just data from area i. A benefit is potential reduction in the mean-squared error of the estimates around the true values.

We illustrate model (12.9) with a simulated sample of size 2000 to mimic a poll taken before the 1996 U.S. presidential election. For T_i observations in state i ($i = 1, \ldots, 51$, where $i = 51$ is DC = District of Columbia), y_i is bin(T_i, π_i), where π_i is the actual proportion of votes in state i for Bill Clinton in the 1996 election, conditional on voting for Clinton or the Republican candidate, Bob Dole. Here, T_i is proportional to the state's population size, subject to $\sum_i T_i = 2000$. Table 12.2 shows $\{T_i\}$, $\{\pi_i\}$, and $\{p_i = y_i/T_i\}$.

For the ML fit of model (12.9), $\hat{\alpha} = 0.163$ and $\hat{\sigma} = 0.29$. The predicted random effect values (obtained using PROC NLMIXED in SAS) yield the proportion estimates $\{\hat{\pi}_i\}$, also shown in Table 12.2. Since $\{T_i\}$ are mostly small and since $\hat{\sigma}$ is relatively small, considerable shrinkage of these estimates occurs from the sample proportions toward the overall proportion supporting Clinton, which was 0.548. The $\{\hat{\pi}_i\}$ vary only between 0.468 (for TX = Texas) and 0.696 (for NY = New York), whereas the sample proportions vary between 0.111 (for Idaho) and 1.0 (for DC). Sample proportions based on fewer observations, such as DC, tended to shrink more. Although the estimates incorporating random effects are relatively homogeneous, they tend to be closer than the sample proportions to the true values.

TABLE 12.2 Estimates of Proportion of Vote for Clinton, Conditional on Voting for Clinton or Dole in 1996 U.S. Presidential Election[a]

State	T_i	π_i	p_i	$\hat{\pi}_i$	State	T_i	π_i	p_i	$\hat{\pi}_i$
AK	5	0.394	0.200	0.508	MT	7	0.483	0.429	0.526
AL	32	0.463	0.500	0.524	NC	55	0.475	0.455	0.494
AR	19	0.594	0.526	0.537	ND	5	0.461	0.600	0.546
AZ	34	0.512	0.618	0.573	NE	13	0.395	0.462	0.524
CA	240	0.572	0.538	0.538	NH	9	0.567	0.556	0.543
CO	29	0.492	0.586	0.558	NJ	60	0.600	0.667	0.611
CT	25	0.604	0.720	0.602	NM	13	0.540	0.462	0.524
DC	4	0.903	1.000	0.576	NV	12	0.506	0.500	0.533
DE	5	0.586	0.400	0.527	NY	137	0.660	0.752	0.696
FL	108	0.532	0.602	0.583	OH	84	0.536	0.488	0.507
GA	56	0.494	0.554	0.548	OK	23	0.456	0.478	0.520
HI	9	0.643	0.556	0.543	OR	24	0.547	0.625	0.569
IA	22	0.557	0.500	0.528	PA	90	0.552	0.567	0.558
ID	9	0.391	0.111	0.472	RI	7	0.689	0.571	0.545
IL	89	0.596	0.539	0.540	SC	28	0.469	0.571	0.552
IN	44	0.468	0.432	0.488	SD	6	0.479	0.667	0.555
KS	19	0.400	0.316	0.477	TN	40	0.513	0.500	0.522
KY	29	0.506	0.448	0.506	TX	144	0.473	0.444	0.468
LA	33	0.566	0.667	0.592	UT	15	0.380	0.333	0.490
MA	46	0.686	0.739	0.637	VA	51	0.489	0.412	0.473
MD	38	0.586	0.474	0.511	VT	4	0.633	0.500	0.538
ME	9	0.627	0.778	0.578	WA	42	0.572	0.619	0.578
MI	73	0.573	0.589	0.570	WI	39	0.559	0.487	0.517
MN	35	0.594	0.571	0.554	WV	14	0.584	0.571	0.548
MO	41	0.535	0.561	0.550	WY	4	0.426	0.250	0.518
MS	21	0.472	0.333	0.477					

[a] π_i, True; p_i, sample; $\hat{\pi}_i$, estimate using random effects model.

12.3.2 Modeling Repeated Binary Responses

In Section 12.1.4 we introduced a random effects version of the Rasch model for repeated binary measurement. This model extends to incorporate covariates.

We illustrate using Table 10.13, first analyzed in Section 10.7.2. The subjects indicated whether they supported legalizing abortion in each of three situations. Table 10.13 also classified the subjects by gender. Let y_{it} denote the response for subject i on item t, with $y_{it} = 1$ representing support. Consider the model

$$\text{logit}\big[P(Y_{it} = 1 \,|\, u_i)\big] = u_i + \beta_t + \gamma x_i, \tag{12.10}$$

where $x_i = 1$ for females and 0 for males, and where $\{u_i\}$ are independent $N(0, \sigma^2)$. (Equivalently, one could place a constraint on $\{\beta_t\}$ and allow an

intercept α.) Here, the gender effect γ is assumed the same for each item, and the $\{\beta_t\}$ refer to the items.

Since model (12.10) implies nonnegative association among responses on the items, one should use items and scales for which this should occur. For opinions about legalized abortion with scale (yes, no), it would not be appropriate for one question to ask "Do you agree that abortion should be legal when a woman is not married?" and another to ask "Do you agree that abortion should be illegal during the last three months of pregnancy?"

Table 12.3 summarizes ML fitting results. The contrasts of $\{\hat{\beta}_t\}$ indicate greater support for legalized abortion with item 1 (when the family has a low income and cannot afford any more children) than with the other two. There is slight evidence of greater support with item 2 (when the woman is not married and does not want to marry the man) than with item 3 (when the woman wants the abortion for any reason). The fixed effects estimates have log odds ratio interpretations. For a given subject of either gender, for instance, the estimated odds of supporting legalized abortion for item 1 equal $\exp(0.83) = 2.3$ times the estimated odds for item 3. Since $\hat{\gamma} = 0.01$, for each item the estimated probability of supporting legalized abortion is similar for females and males with similar random effect values.

For these data, subjects are highly heterogeneous ($\hat{\sigma} = 8.6$). Thus, strong associations exist among responses on the three items. This is reflected by 1595 of the 1850 subjects making the same response on all three items: that is, response patterns (0, 0, 0) and (1, 1, 1). It implies tremendous variability in between-subject odds ratios. From (12.7), for different subjects of a given gender, the middle 50% of odds ratios comparing items 1 and 3 are estimated to vary between about $\exp(0.83 - 0.95 \times 8.6)$ and $\exp(0.83 + 0.95 \times 8.6)$.

For contingency tables, one can obtain cell fitted values. To do this, one must integrate over the estimated random effects distribution to obtain estimated marginal probabilities of any particular sequence of responses. For the ML parameter estimates, the probability of a particular sequence of responses $(y_{i1}, \ldots, y_{iT})$ for a given u_i is the appropriate product of conditional probabilities, $\prod_t P(Y_{it} = y_{it}|u_i)$, since the responses are independent given u_i. Integrating this product probability with respect to u_i for the

TABLE 12.3 Summary of ML Estimates for Random Effects Model (12.10) and ML and GEE Estimates for Corresponding Marginal Model

		GLMM ML		Marginal Model ML		Marginal Model GEE	
Effect	Parameter	Estimate	SE	Estimate	SE	Estimate	SE
Abortion	$\beta_1 - \beta_3$	0.83	0.16	0.148	0.030	0.149	0.030
	$\beta_1 - \beta_2$	0.54	0.16	0.098	0.027	0.097	0.028
	$\beta_2 - \beta_3$	0.29	0.16	0.049	0.027	0.052	0.027
Gender	γ	0.01	0.48	0.005	0.088	0.003	0.088
$\sqrt{\text{var}(u_i)}$	σ	8.6	0.54				

$N(0, \hat{\sigma}^2)$ distribution estimates the marginal probability for a given cell (averaged over subjects). This requires numerical integration methods described in Section 12.6. Multiplying this marginal probability of a given sequence by the sample size for that multinomial gives a fitted value.

Not surprisingly, for these data, the response patterns $(0, 0, 0)$ and $(1, 1, 1)$ also have the largest fitted values for the multinomial for each gender. For instance, for females 440 indicated support under all three circumstances (457 under none of the three), and the fitted value was 436.5 (459.3). Overall chi-squared statistics comparing the 16 observed and fitted counts are $G^2 = 23.2$ and $X^2 = 27.8$ (df = 9). These are not that large considering the very large sample size and the few parameters ($\beta_1, \beta_2, \beta_3, \gamma, \sigma$) used to describe the 14 multinomial cell probabilities ($8 - 1 = 7$ for each gender) in Table 10.13. Here, df = 9 since we are modeling 14 multinomial parameters using five GLMM parameters.

An extended model allows interaction between gender and item. It has different $\{\beta_t\}$ for men and women. However, it does not fit better. The likelihood-ratio statistic = 1.0 (df = 2) for testing that the extra parameters equal 0.

An alternative analysis of these data focuses on the marginal distributions, treating the dependence as a nuisance. A marginal model analog of (12.10) is

$$\text{logit}\big[P(Y_t = 1)\big] = \beta_t + \gamma x.$$

For it, Table 12.3 also shows GEE estimates for the exchangeable working correlation structure and ML estimates. The marginal model fits well, with $G^2 = 1.1$; here, df = 2 since the model describes six marginal probabilities (three for each gender) using four parameters. These population-averaged $\{\hat{\beta}_t\}$ are much smaller than the subject-specific $\{\hat{\beta}_t\}$ from the GLMM. This reflects the very large GLMM heterogeneity ($\hat{\sigma} = 8.6$) and the corresponding strong correlations among the three responses. For instance, the GEE analysis estimates a common correlation of 0.82 between pairs of responses. Although the GLMM $\{\hat{\beta}_t\}$ are about five to six times the marginal model $\{\hat{\beta}_t\}$, so are the standard errors. The two approaches provide similar substantive interpretations and conclusions.

12.3.3 Longitudinal Mental Depression Study Revisited

We now revisit Table 11.2 from a longitudinal study to compare a new drug with a standard for treating subjects suffering mental depression. In Section 11.2.1 we analyzed the data using marginal models. The response y_t for measurement t on mental depression equals 1 for normal and 0 for abnormal. For severity of initial diagnosis s (1 = severe, 0 = mild), drug treatment d (1 = new, 0 = standard), and time of measurement t, we used the model

$$\text{logit}\big[P(Y_t = 1)\big] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4\, dt$$

to evaluate the marginal distributions.

TABLE 12.4 Model Parameter Estimates for Marginal and Conditional Logit Models Fitted to Table 11.2

Parameter	ML Marginal Estimate	Std. Error	GEE Marginal Estimate	Std. Error	Random Effects ML Estimate	Std. Error
Diagnosis	−1.29	0.14	−1.31	0.15	−1.32	0.15
Drug	−0.06	0.22	−0.06	0.23	−0.06	0.22
Time	0.48	0.12	0.48	0.12	0.48	0.12
Drug × Time	1.01	0.18	1.02	0.19	1.02	0.19

Now let y_{it} denote observation t for subject i. The model

$$\text{logit}\big[P(Y_{it} = 1 \,|\, u_i)\big] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 dt + u_i$$

has subject-specific rather than population-averaged effects. Table 12.4 shows the ML estimates. The time trend estimates are $\hat{\beta}_3 = 0.48$ for the standard drug and $\hat{\beta}_3 + \hat{\beta}_4 = 1.50$ for the new one. These are nearly identical to the ML and GEE estimates for the corresponding marginal model, also shown in the table (these are discussed in Sections 11.2.1 and 11.3.2). The reason is that the repeated observations do not exhibit much correlation, as the GEE analysis observed. Here, this is reflected by $\hat{\sigma} = 0.07$, showing little heterogeneity among subjects.

Based on the model fit, integrating over the $N(0, 0.07^2)$ random effects distribution yields marginal fitted values of the possible response sequences. Comparing these to the sample counts in Table 11.2 indicates a relatively good fit. The model describes the 28 multinomial cell probabilities (seven for the trivariate response at each of the four severity–drug combinations) using six parameters. The usual fit statistics comparing the observed cell counts to their fitted values are $G^2 = 22.0$ and $X^2 = 20.8$ (df = $28 - 6 = 22$).

The deviance increases by only 0.001 when one assumes that $\sigma = 0$. From results to be discussed in Section 12.6.6, the P-value for comparing models is half what one gets by treating the deviance as chi-squared with df = 1, or $P = 0.49$. This simpler model, which gives nearly identical effect estimates and SE values, is adequate. This is also suggested by AIC values (e.g., PROC NLMIXED in SAS reports 1173.9 for the random effects model and 1171.9 for the simpler model with $\sigma = 0$).

12.3.4 Modeling Heterogeneity among Multicenter Clinical Trials

Many applications compare two groups on a response for data stratified on a third variable. With binary outcomes, the data form several 2×2 contingency tables. The main focus relates to studying the association in the 2×2 tables and whether and how it varies among the strata.

The strata are sometimes themselves a sample, such as schools or medical clinics. A random effects approach is then natural. With a random sampling of strata, it enables inferences to extend to the population of strata. The fit of the random effects model provides a simple summary such as an estimated mean and standard deviation of log odds ratios for the population of strata. In each stratum it also provides a predicted log odds ratio that shrinks the sample value toward the mean. This is especially useful when the sample size in a stratum is small and the ordinary sample odds ratio has large standard error. Even when the strata are not a random sample or not even a sample and a random effects approach is not as natural, the model is beneficial for these purposes.

We illustrate using Table 12.5, previously analyzed in Section 6.3, showing the results of a clinical trial at eight centers. The purpose was to compare an active drug and a control, for curing an infection. For a subject in center i using treatment t (1 = active drug; 2 = control), let $y_{it} = 1$ denote success. One possible model is the logistic-normal,

$$\begin{aligned} \text{logit}\big[P(Y_{i1} = 1 \,|\, u_i)\big] &= \alpha + \beta/2 + u_i \\ \text{logit}\big[P(Y_{i2} = 1 \,|\, u_i)\big] &= \alpha - \beta/2 + u_i, \end{aligned} \tag{12.11}$$

TABLE 12.5 Clinical Trial Relating Treatment to Response for Eight Centers

Center	Treatment	Response		Sample Odds Ratio	Fitted Odds Ratio
		Success	Failure		
1	Drug	11	25	1.19	2.02
	Control	10	27		
2	Drug	16	4	1.82	2.09
	Control	22	10		
3	Drug	14	5	4.80	2.19
	Control	7	12		
4	Drug	2	14	2.29	2.11
	Control	1	16		
5	Drug	6	11	∞	2.18
	Control	0	12		
6	Drug	1	10	∞	2.12
	Control	0	10		
7	Drug	1	4	2.0	2.11
	Control	1	8		
8	Drug	4	2	0.33	2.06
	Control	6	1		

Source: Beitler and Landis (1985).

where $\{u_i\}$ are independent $N(0, \sigma^2)$ variates. This model assumes that the log odds ratio β between treatment and response is constant over centers. The parameter σ summarizes center heterogeneity in the success probabilities.

A logistic-normal model permitting treatment-by-center interaction is

$$\begin{aligned} \text{logit}\big[P(Y_{i1} = 1 \,|\, u_i, b_i)\big] &= \alpha + (\beta + b_i)/2 + u_i, \\ \text{logit}\big[P(Y_{i2} = 1 \,|\, u_i, b_i)\big] &= \alpha - (\beta + b_i)/2 + u_i, \end{aligned} \tag{12.12}$$

where $\{u_i\}$ are independent $N(0, \sigma_a^2)$, $\{b_i\}$ are independent $N(0, \sigma_b^2)$, and $\{u_i\}$ are independent of $\{b_i\}$. The log odds ratio equals $\beta + b_i$ in center i. These vary among centers according to a $N(\beta, \sigma_b^2)$ distribution. That is, β is the expected center-specific log odds ratio between treatment and response, and σ_b describes variability in those log odds ratios. The model parameters are $(\alpha, \beta, \sigma_a, \sigma_b)$.

In Table 12.5 the sample success rates vary markedly among centers both for the control and drug treatments, but in all except the last center that rate is higher for the drug treatment. In using models with random center and possibly random treatment effects, it is preferable to have more than eight centers. It is difficult to get reliable variance component estimates with so few centers. Keeping this in mind, we use these data to illustrate the models. With a large number of centers it would also be sensible to allow correlation between b_i and u_i, but we shall not attempt that here. The treatment estimates are $\hat{\beta} = 0.739$ (SE = 0.300) for the model (12.11) of no interaction and $\hat{\beta} = 0.746$ (SE = 0.325) for the model (12.12) permitting interaction. Considerable evidence of a drug effect occurs. With such a small sample, however, it is unclear whether that effect is weak or moderate.

The evidence about association is weaker for the model permitting interaction. The Wald statistics are $(0.739/0.300)^2 = 6.0$ for the no-interaction model and $(0.746/0.325)^2 = 5.3$ for the interaction model. The corresponding likelihood-ratio statistics are 6.3 and 4.6 (df = 1). The extra variance component in the interaction model pertains to variability in the log odds ratios. As its estimate $\hat{\sigma}_b$ increases, so does the standard error of the estimated treatment effect $\hat{\beta}$ tend to increase. In this example, $\hat{\sigma}_b = 0.15$ is relatively small and the standard errors of $\hat{\beta}$ are not very different in the two models. When $\hat{\sigma}_b = 0$, the standard errors and the model fits are the same.

To show the effect of larger $\hat{\sigma}_b$ on the standard error of the mean treatment effect estimate $\hat{\beta}$, we alter Table 12.5 slightly. We change three failures to successes for drug in center 3 and three successes to failures for drug in center 8. With these changes, the estimated variability of the treatment effects increases from $\hat{\sigma}_b = 0.15$ to $\hat{\sigma}_b = 1.4$. The ML estimates of the mean treatment effects are then $\hat{\beta} = 0.722$ (SE = 0.299) for the no interaction model (12.11) and $\hat{\beta} = 0.767$ (SE = 0.623) for the interaction model. The Wald statistics are 5.8 and 1.5. The evidence of a treatment

effect is then dramatically weaker for the interaction model (12.12). Not surprisingly, when the treatment effect varies substantially among centers, it is more difficult to estimate the mean of that effect.

For the actual data in Table 12.5, because $\hat{\sigma}_b = 0.15$ for model (12.12) is relatively small, the model shrinks the sample odds ratios considerably. Table 12.5 shows the sample values and the model predicted values. These are based on predicting the random effects (to be explained in Section 12.6), and substituting them and the ML estimates of fixed effects into the model formula to estimate the two response probabilities for each treatment in each center. The sample odds ratios vary from 0.33 to ∞; their random effects model counterparts (computed with PROC NLMIXED in SAS) vary only between 2.0 and 2.2. The smoothed estimates are much less variable and do not have the same ordering as the sample values. For instance, the smoothed estimate of 2.2 for center 3 is greater than the estimate of 2.1 for center 6, even though the sample value is infinite for the latter. This partly reflects the greater shrinkage that occurs when sample sizes are smaller. When $\hat{\sigma}_b = 0$, model (12.12) provides the same fit as model (12.11), and estimated odds ratios are identical in each center.

For related analyses permitting heterogeneity in odds ratios with several 2×2 tables, see Liu and Pierce (1993) and Skene and Wakefield (1990).

12.3.5 Alternative Formulations of Random Effects Models

There are other ways to express the models. For instance, an equivalent expression for interaction model (12.12) is

$$\text{logit}\big[P(Y_{it} = 1|u_i, b_{it})\big] = \alpha + \beta x_t + b_{it} + u_i,$$

where x_t is a treatment dummy variable ($x_1 = 1, x_2 = 0$), $\{u_i\}$ are independent $N(0, \sigma_a^2)$, and $\{b_{i1}\}$ and $\{b_{i2}\}$ are independent $N(0, \sigma^2)$. Here, $b_{i1} - b_{i2}$ corresponds to b_i in parameterization (12.12), and $2\sigma^2$ corresponds to σ_b^2.

Formulating a random effects model requires care about implications of the model expression and the random effects correlation structure. Suppose that one expressed the interaction model (12.12) as

$$\text{logit}\big[P(Y_{it} = 1\,|\,u_i, b_i)\big] = \alpha + (\beta + b_i)x_t + u_i, \qquad (12.13)$$

with $\{b_i\}$ from $N(0, \sigma_b^2)$. This is inappropriate, since the model then imposes greater variability for the logit with the first treatment than the second, since $x_2 = 0$ and $\{u_i\}$ and $\{b_i\}$ are uncorrelated. Also, the model should not depend on the definition of the dummy variable x_t. Note, however, that if $z_t = x_t + c$ for some constant c, then model (12.13) is equivalently

$$\begin{aligned}&\text{logit}\big[P(Y_{it} = 1\,|\,u_i, b_i)\big] \\ &\quad = \alpha + (\beta + b_i)(z_t - c) + u_i = \alpha' + (\beta + b_i)z_t + v_i,\end{aligned}$$

where $\alpha' = \alpha - c\beta$ and $v_i = u_i - cb_i$. Thus, (v_i, b_i) are correlated even if (u_i, b_i) are not. In fact, expression (12.13) is sensible only with correlated random effects. It is then equivalent to (12.12) with correlated random effects. See Agresti and Hartzel (2000) for further discussion.

12.3.6 Capture–Recapture Modeling to Predict Population Size

Capture–recapture experiments are a method of using a series of samples to estimate the size of a population. Such methods have traditionally been used to estimate animal abundance in some habitat. At each sampling occasion, animals are captured and marked in some manner. The animals captured for any given sample are freed and all animals are candidates for recapture in a later sample. With T sampling occasions, a 2^T contingency table displays the data, with scale (captured, not captured) at each occasion. The count $n_{22\cdots2}$ is missing for the cell corresponding to noncapture at each occasion. If we knew this cell count, adding it to the others would yield the population size. Models specified for this 2^T table use the $2^T - 1$ observed counts to fit the model. The fit refers to those $2^T - 1$ cells, but extrapolating it yields an estimated count in the unobserved cell. Adding that to the total of the $2^T - 1$ observed counts yields an estimate of population size.

To illustrate, suppose that $T = 2$. We observe n_{11} animals at both occasions, n_{12} at the first but not the second occasion, and n_{21} at the second but not the first. We do not know the number n_{22} not captured either time. If we assumed independence in the 2×2 table, the prediction $\hat{n}_{22}$ would be the value giving an odds ratio of 1.0; but $(n_{11}\hat{n}_{22})/(n_{12}n_{21}) = 1$ implies that $\hat{n}_{22} = n_{12}n_{21}/n_{11}$. This yields a population size prediction (Sekar and Deming 1949) of

$$\begin{aligned}\hat{N} &= n_{11} + n_{12} + n_{21} + n_{12}n_{21}/n_{11} \\ &= n_{1+}n_{+1}/n_{11} \quad \text{with} \quad \widehat{\text{var}}(\hat{N}) = \frac{n_{1+}n_{+1}n_{12}n_{21}}{n_{11}^3}.\end{aligned}$$

The assumption of independence is usually unrealistic, however. With additional sampling occasions, one can try more complex models.

Table 12.6, analyzed by Cormack (1989) and others, refers to a study having $T = 6$ consecutive trapping days for a population of snowshoe hares. The study observed 68 hares. For instance, Table 12.6 indicates that 3 hares were observed on the first day but on none of the other days. For simplicity, models for studies over a brief time period assume that no deaths, births, or immigration into the population occurred during the study period. This is called a *closed population.*

Most methods for capture–recapture treat the probability of capture at a given occasion as identical for each subject (e.g., animal). This is usually

TABLE 12.6 Results of Capture–Recapture of Snowshoe Hares

Capture 6	Capture 5	Capture 4	Capture 3, Capture 2, Capture 1[a] 000	001	010	011	100	101	110	111
0	0	0	—	3	6	0	5	1	0	0
			(24.0)	(2.3)	(5.4)	(0.9)	(3.2)	(0.5)	(1.2)	(0.3)
0	0	1	3	2	3	0	0	1	0	0
			(4.8)	(0.8)	(1.8)	(0.5)	(1.1)	(0.3)	(0.6)	(0.3)
0	1	0	4	2	3	1	0	1	0	0
			(3.9)	(0.6)	(1.5)	(0.4)	(0.9)	(0.2)	(0.5)	(0.2)
0	1	1	1	0	0	0	0	0	0	0
			(1.3)	(0.3)	(0.8)	(0.3)	(0.5)	(0.2)	(0.4)	(0.3)
1	0	0	4	1	1	1	2	0	2	0
			(6.8)	(1.1)	(2.6)	(0.6)	(1.5)	(0.4)	(0.9)	(0.4)
1	0	1	4	0	3	0	1	0	2	0
			(2.3)	(0.6)	(1.3)	(0.5)	(0.8)	(0.3)	(0.7)	(0.4)
1	1	0	2	0	1	0	1	0	1	0
			(1.9)	(0.5)	(1.1)	(0.4)	(0.7)	(0.3)	(0.6)	(0.4)
1	1	1	1	1	1	0	0	0	1	2
			(1.0)	(0.4)	(0.9)	(0.5)	(0.5)	(0.3)	(0.7)	(0.7)

[a] Values in parentheses represent the fit of the logistic-normal model.

Source: A. Agresti, *Biometrics* **50**: 494–500 (1994).

unrealistic. One way to allow heterogeneous capture probabilities uses a logit model having subject random effects. For subject i, $i = 1, \ldots, N$ with N unknown, let $\mathbf{y}_i' = (y_{i1}, \ldots, y_{iT})$, where $y_{it} = 1$ denotes capture in sample t and $y_{it} = 0$ denotes noncapture. Lacking explanatory variables, one might use the Rasch-type model

$$\text{logit}\big[P(Y_{it} = 1 \,|\, u_i)\big] = u_i + \beta_t,$$

where $\{u_i\}$ are independent $N(0, \sigma^2)$. The larger the value of β_t, the greater the capture probability at occasion t. The larger is σ, the more heterogeneous are the capture probabilities. When $\sigma = 0$ this logistic-normal model simplifies to mutual independence [i.e., loglinear model (8.6)] for the 2^T table.

As with other random effects models, integrating the random effect from the probability mass function of $(\mathbf{y}_i \,|\, u_i)$ yields the likelihood function (as discussed in Section 12.6). One can consider this likelihood function and the resulting ML estimates of $\{\beta_t\}$ and σ for all possible counts in the unobserved cell. A profile likelihood function views the maximized likelihood as a function of the unobserved cell count. The ML prediction for that unobserved cell count is the value that maximizes this profile likelihood. Lacking specialized software, one can fit the random effects model repeatedly with various counts in the unobserved cell to determine by trial and error the count that maximizes the likelihood function.

ML fitting of this model to Table 12.6 yields a prediction of 24 for the unobserved cell count. Since the study observed 68 hares, the population size estimate is $\hat{N} = 92$. For this fit, $\hat{\sigma} = 1.0$.

Methods for obtaining a confidence interval for N include using the profile likelihood function or a nonparametric bootstrap method. With the profile likelihood approach, the interval for the missing cell count consists of the possible counts for that cell such that the G^2 fit statistic increases by less than $\chi_1^2(\alpha)$ from its value at the ML estimate. Adding the number of subjects observed in the samples to the endpoints of this interval gives the corresponding interval for N. For the snowshoe hares, a 95% profile-likelihood confidence interval for N is $(75, 154)$. It is common for $\hat{N}$ to be nearer the low end of the interval. See Coull and Agresti (1999) for details.

The greater the heterogeneity, as reflected by larger $\hat{\sigma}$, $\hat{N}$ tends to be larger and the confidence interval tends to be wider. Large $\hat{\sigma}$ causes difficulties in estimation, since it results in a relatively flat likelihood surface. This implies imprecise estimates of N. In particular, the upper limit of the profile-likelihood confidence interval for N is essentially infinite when the likelihood function gets sufficiently flat. Also, the ML estimator is then often unstable, with small changes in the data yielding large changes in $\hat{N}$. Difficulties can also arise when probabilities of capture are small. Evidence of this occurs when most subjects captured appear in only one sample. When this happens or when $\hat{\sigma}$ is large, it is unrealistic to expect narrow confidence intervals for N.

Alternative models are discussed in Section 13.1.3. Models that ignore likely heterogeneity can give unrealistically narrow confidence intervals for N. Although traditionally used for animal populations, capture–recapture applications also include estimating population size for human populations, such as estimating population prevalence of injecting drug use and HIV infection. Darroch et al. (1993) considered census population estimation, and Chao et al. (2001) estimated the number of people infected during a hepatitis outbreak (Problem 12.21). An interesting application is estimating the number of files on the World Wide Web relating to some subject by taking samples using several search engines (Fienberg et al. 1999).

12.4 RANDOM EFFECTS MODELS FOR MULTINOMIAL DATA

Random effects models for binary responses extend to multicategory responses. For the multicategory models of Chapter 7, a multinomial observation with I categories is a vector of $I - 1$ indicators, the jth of which is 1 when the observation falls in category j and 0 otherwise. In Section 7.1.5 we defined a multivariate GLM by applying a vector of link functions to this multivariate response. Adding random effects extends this multivariate GLM and the GLMM (12.1) to a multivariate GLMM (Hartzel et al. 2001b; Tutz and Hennevogl 1996). This class includes models for nominal and ordinal responses.

12.4.1 Cumulative Logit Model with Random Intercept

Modeling is simpler with ordinal than nominal responses, since often the same random effect and the same fixed effect can apply to each logit. With cumulative logits, this is the *proportional odds* structure (Section 7.2.2). Denote the possible outcomes for y_{it}, observation t in cluster i, by $1, 2, \ldots, I$. A GLMM for the cumulative logits has the form

$$\text{logit}\big[P(Y_{it} \le j \,|\, \mathbf{u}_i)\big] = \alpha_j + \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_{it}'\mathbf{u}_i, \quad j = 1, \ldots, I-1. \quad (12.14)$$

Hedeker and Gibbons (1994) discussed model fitting, primarily with $\mathbf{u}_i$ as multivariate normal.

For cumulative logit and probit random intercept models, the same relationship exists between their effects and those in marginal models as presented in Section 12.2.2 for binary-response models. Marginal effects tend to be smaller, increasingly so as σ increases. Also, the same predictor structure as in (12.14) holds with other links for which a common effect for each logit is plausible. For instance, Hartzel et al. (2001a, b) used it with adjacent-categories logits.

12.4.2 Insomnia Study Revisited

Table 11.4 showed results of a clinical trial at two occasions comparing a drug with placebo in treating insomnia patients. In Sections 11.2.3 and 11.3.3 the data were analyzed with marginal models. For y_t = time to fall asleep at occasion t, the marginal model

$$\text{logit}\big[P(Y_t \le j)\big] = \alpha_j + \beta_1 t + \beta_2 x + \beta_3 tx$$

permitted interaction between t = occasion (0 = initial, 1 = follow-up) and x = treatment (1 = active, 0 = placebo). Table 12.7 shows the ML and GEE estimates.

Now, let y_{it} denote the response for subject i at occasion t. Table 12.7 also shows results of fitting the random-intercept model

$$\text{logit}\big[P(Y_{it} \le j \,|\, u_i)\big] = u_i + \alpha_j + \beta_1 t + \beta_2 x + \beta_3 tx.$$

TABLE 12.7 Fits of Cumulative Logit Models to Table 11.4[a]

Effect	Marginal ML	Marginal GEE	Random Effects (GLMM) ML
Treatment	0.046 (0.236)	0.034 (0.238)	0.058 (0.366)
Occasion	1.074 (0.162)	1.038 (0.168)	1.602 (0.283)
Treatment × occasion	0.662 (0.244)	0.708 (0.244)	1.081 (0.380)

[a] Values in parentheses represent standard errors.

Results are substantively similar to the marginal model, but estimates and standard errors are about 50% larger. This reflects the relatively large heterogeneity ($\hat{\sigma} = 1.90$) and the resultant strong association between the responses at the two occasions.

12.4.3 Cluster Sampling

With surveys that use cluster sampling, standard methods based on simple random sampling (e.g., for a single multinomial sample) require adjustment. Ordinary standard errors are too small. The usual chi-squared test statistics no longer have chi-squared null distributions, but rather, weighted sums of chi-squared. Rao and Thomas (1988) surveyed ways of adjusting standard inferences to take into account complex sampling methods in the analysis and modeling of categorical data.

When the sampling scheme randomly samples clusters, one can account for the clustering using cluster random effects. We illustrate using data from Brier (1980), who reported 96 observations taken from 20 neighborhoods (the clusters) on Y = satisfaction with home and X = satisfaction with neighborhood as a whole. Each variable was measured with the ordinal scale (unsatisfied, satisfied, very satisfied). Brier's analysis adjusted for clustering by reducing the Pearson statistic for testing independence in the 3×3 contingency table relating X and Y from 17.9 to 15.7 (df = 4).

Consider the model for y_{it}, observation t in cluster i,

$$\text{logit}\big[P(Y_{it} \le j \,|\, u_i)\big] = u_i + \alpha_j + x_{it}\beta, \tag{12.15}$$

with scores $(1, 2, 3)$ for the satisfaction levels of x_{it}. With a $N(0, \sigma^2)$ distribution assumed for u_i, the ML effect estimate is $\hat{\beta} = -1.201$ (SE = 0.407), with $\hat{\sigma} = 0.92$. By contrast, treating the 96 observations as a random sample corresponds to fitting this model with $\sigma = 0$. It has $\hat{\beta} = -1.226$ (SE = 0.370). A slight reduction in significance results from adjusting for clustering.

12.4.4 Baseline-Category Logit Models with Random Effects

For nominal response variables, one can formulate a binary model that pairs each category with a baseline and fit these models simultaneously while allowing separate effects. This requires using a vector of cluster-specific random effects $\mathbf{u}_{ij}$, one for each logit. The general form of the baseline-category logit model with random effects is

$$\log\frac{P(Y_{it} = j)}{P(Y_{it} = I)} = \alpha_j + \mathbf{x}'_{it}\boldsymbol{\beta}_j + \mathbf{z}'_{it}\mathbf{u}_{ij}, \quad j = 1, \ldots, I-1.$$

The fixed effects $\boldsymbol{\beta}_j$ and the random effects $\mathbf{u}_{ij}$ depend on j, since the baseline category is arbitrary. With nominal responses there is no reason to expect effects to be similar for different j.

Cluster i has a vector $\mathbf{u}_i' = (\mathbf{u}_{i1}', \ldots, \mathbf{u}_{i,I-1}')$ of random effects. The usual approach treats $\{\mathbf{u}_i\}$ as independent multivariate normal variates. We recommend an unspecified covariance matrix $\boldsymbol{\Sigma}$ for $\mathbf{u}_i$. For instance, it is sensible to allow different variances for random effects that apply to different logits. With a common variance, that variance would not be the same as that for the implied random effect for a logit for an arbitrary pair of categories, $\log[P(Y_{it} = j)/P(Y_{it} = k)]$. With unspecified covariance the model is structurally the same regardless of the choice of baseline category. See Hartzel et al. (2001b) for an example.

12.5 MULTIVARIATE RANDOM EFFECTS MODELS FOR BINARY DATA

In practice, random effects are often univariate, taking the form of random intercepts. However, we've seen that nominal responses require multivariate random effects and that bivariate random effects are helpful for describing heterogeneity in multicenter clinical trials. In this section we present other examples in which multivariate random effects are natural.

12.5.1 Matched Pairs with a Bivariate Binary Response

Leo Goodman analyzed Table 12.8 in several articles (e.g., Goodman 1974). A sample of schoolboys were interviewed twice, several months apart, and asked about their self-perceived membership in the "leading crowd" and about whether they sometimes needed to go against their principles to belong to that group. Thus, there are two binary response variables, which we refer to as membership and attitude, measured at two interview times for each subject. Table 12.8 labels the categories for attitude as (positive, negative), where "positive" refers to disagreeing with the statement that one must go against his principles.

TABLE 12.8 Membership and Attitude Toward the "Leading Crowd"

(M, A) for First Interview	(M, A) for Second Interview[a]			
	(Yes, Positive)	(Yes, Negative)	(No, Positive)	(No, Negative)
Yes, positive	458	140	110	49
Yes, negative	171	182	56	87
No, positive	184	75	531	281
No, negative	85	97	338	554

[a] M, membership; A, attitude.

Source: J. S. Coleman, *Introduction to Mathematical Sociology* (London: Free Press of Glencoe, 1964), p. 170.

For subject i, let y_{itv} be the response at interview time t on variable v, where $v = M$ for membership and $v = A$ for attitude. The logit model

$$\text{logit}\big[P(Y_{itv} = 1 \,|\, u_{iv})\big] = \beta_{tv} + u_{iv} \tag{12.16}$$

is a multivariate form of the Rasch-type model (12.4). It has additive item and subject effects for each variable v. Here, (u_{iM}, u_{iA}) is a bivariate random effect that describes subject heterogeneity for (membership, attitude). We assume that the $\{(u_{iM}, u_{iA})\}$ are independent from a bivariate normal distribution, $N(\mathbf{0}, \boldsymbol{\Sigma})$, with possibly different variances and nonzero correlation.

The ML fit yields $\hat{\beta}_{2M} - \hat{\beta}_{1M} = 0.379$ (SE = 0.075) and $\hat{\beta}_{2A} - \hat{\beta}_{1A} = 0.176$ (SE = 0.058). For both variables, the probability of the first outcome category is higher at the second interview. For instance, for a given subject the odds of self-perceived membership in the leading crowd at interview 2 are estimated to be exp(0.379) = 1.46 times the odds at interview 1.

The estimated correlation between the random effects is 0.30. Their estimated standard deviations are $\hat{\sigma}_1 = 3.1$ for $\{u_{iM}\}$ and $\hat{\sigma}_2 = 1.5$ for $\{u_{iA}\}$. Since these are quite different, the relative sizes of membership and attitude effects differ for marginal and conditional models (recall the caveat in Section 12.2.3). The marginal effect is attenuated more for membership. For this conditional model, the ratio of estimated odds ratios is $\exp(0.379)/\exp(0.176) = 1.46/1.19 = 1.22$. For the marginal model, the estimated odds ratios use the marginal distributions of each variable at each time [e.g., this is $(1392/2006)/(1253/2145) = 1.188$ for membership], and the ratio of estimated odds ratios is $1.188/1.133 = 1.05$.

Integrating over the estimated random effects distribution yields fitted values for the 16 possible sequences of responses in Table 12.8. The deviance of $G^2 = 5.5$ (df = 8) compares the 16 observed counts to their fitted values. The model, which describes 15 multinomial probabilities with seven parameters, fits well. The model constraining the random effects to be uncorrelated fits poorly ($G^2 = 97.5$, df = 9). The model constraining the random effects to be perfectly correlated is equivalent to having a single random effect u_i for each subject. The model is then a Rasch-type model with four items that are the combinations of interviews and variables. That model fits very poorly ($G^2 = 655.5$, df = 10). Agresti et al. (2000) gave further details.

12.5.2 Continuation-Ratio Logits for Clustered Ordinal Outcomes: Toxicity Study

For continuation-ratio logit models with ordinal responses, the logits refer to independent binomial variates (Section 7.4.3). Thus, binary logit random effects models apply to clustered ordinal responses using continuation-ratio logits (Ten Have and Uttal 1994). For observation t in cluster i, let $\omega_{ij} = P(Y_{it} = j \,|\, Y_{it} \geq j, u_{ij})$. (More generally, this probability could also depend on t, but this generality is not needed for the example below.) The continuation-ratio logits are $\{\text{logit}(\omega_{ij}),\ j = 1, \ldots, I-1\}$.

Let n_{ij} be the number of subjects in cluster i making response j. Let $n_i = \sum_{j=1}^{I} n_{ij}$. For a given cluster in a continuation-ratio logit model, treating $(n_{i1}, \ldots, n_{i,I-1})$ as multinomial is equivalent to treating them as a sequential set of independent binomial variates, where n_{ij} is $\text{bin}(n_i - \sum_{h<j} n_{ih}, \omega_{ij})$, $j = 1, \ldots, I-1$.

We illustrate with a developmental toxicity study conducted under the U.S. National Toxicology Program. This study examined the developmental effects of ethylene glycol (EG) by administering one of four dosages (0, 0.75, 1.50, 3.00 g/kg) to pregnant rodents. The four dose groups had (25, 24, 22, 23) pregnant rodents. The clusters are litters of mice. The three possible outcomes (dead/resorption, malformation, normal) for each fetus are ordered, normal being the most desirable result. Table 12.9 shows the data. The continuation-ratio logit is natural here since categories are hierarchically related; an animal must survive before a malformation can take place. The following analyses are from Coull and Agresti (2000).

For litter i in dose group d, let $\text{logit}(\omega_{i(d)1})$ be the continuation-ratio logit for the probability of death and $\text{logit}(\omega_{i(d)2})$ the continuation-ratio logit for the conditional probability of malformation, given survival. [The notation $i(d)$ represents litter i nested within dose d.] Let x_d be the dosage for group d. We account for the litter effect using litter-specific random effects $\mathbf{u}_{i(d)} = (u_{i(d)1}, u_{i(d)2})$ sampled from $N(\mathbf{0}, \boldsymbol{\Sigma}_d)$. This bivariate random effect allows for differing amounts of overdispersion for the probability of death and for the probability of malformation, given survival. A model also permitting different fixed effects for each is

$$\text{logit}(\omega_{i(d)j}) = u_{i(d)j} + \alpha_j + \beta_j x_d. \tag{12.17}$$

TABLE 12.9 Response Counts for 94 Litters of Mice on (Number Dead, Number Malformed, Number Normal)

Dose = 0.00 g/kg	Dose = 0.75 g/kg	Dose = 1.50 g/kg	Dose = 3.00 g/kg
(1, 0, 7), (0, 0, 14)	(0, 3, 7), (1, 3, 11)	(0, 8, 2), (0, 6, 5)	(0, 4, 3), (1, 9, 1)
(0, 0, 13), (0, 0, 10)	(0, 2, 9), (0, 0, 12)	(0, 5, 7), (0, 11, 2)	(0, 4, 8), (1, 11, 0)
(0, 1, 15), (1, 0, 14)	(0, 1, 11), (0, 3, 10)	(1, 6, 3), (0, 7, 6)	(0, 7, 3), (0, 9, 1)
(1, 0, 10), (0, 0, 12)	(0, 0, 15), (0, 0, 11)	(0, 0, 1), (0, 3, 8)	(0, 3, 1), (0, 7, 0)
(0, 0, 11), (0, 0, 8)	(2, 0, 8), (0, 1, 10)	(0, 8, 3), (0, 2, 12)	(0, 1, 3), (0, 12, 0)
(1, 0, 6), (0, 0, 15)	(0, 0, 10), (0, 1, 13)	(0, 1, 12), (0, 10, 5)	(2, 12, 0), (0, 11, 3)
(0, 0, 12), (0, 0, 12)	(0, 1, 9), (0, 0, 14)	(0, 5, 6), (0, 1, 11)	(0, 5, 6), (0, 4, 8)
(0, 0, 13), (0, 0, 10)	(1, 1, 11), (0, 1, 9)	(0, 3, 10), (0, 0, 13)	(0, 5, 7), (2, 3, 9)
(0, 0, 10), (1, 0, 11)	(0, 1, 10), (0, 0, 15)	(0, 6, 1), (0, 2, 6)	(0, 9, 1), (0, 0, 9)
(0, 0, 12), (0, 0, 13)	(0, 0, 15), (0, 3, 10)	(0, 1, 2), (0, 0, 7)	(0, 5, 4), (0, 2, 5)
(1, 0, 14), (0, 0, 13)	(0, 2, 5), (0, 1, 11)	(0, 4, 6), (0, 0, 12)	(1, 3, 9), (0, 2, 5)
(0, 0, 13), (1, 0, 14)	(0, 1, 6), (1, 1, 8)		(0, 1, 11)
(0, 0, 14)			

Source: Study described by C. J. Price, C. A. Kimmel, R. W. Tyl, and M. C. Marr, *Toxicol. Appl. Pharmacol.* **81**: 113–127 (1985).

TABLE 12.10 Comparisons of Log Likelihoods for Multivariate Random Effects Models for Developmental Toxicity Study

Model	Number of Parameters	Change in Parameters	Change in Log Likelihood
Dose-specific $\boldsymbol{\Sigma}_i$	16	—	—
$\boldsymbol{\Sigma}_i$, Common α, β	14	2	28.4
Common $\boldsymbol{\Sigma}$	7	9	7.4
Common $\boldsymbol{\Sigma}$, $\rho = 0$	6	10	7.4
Univariate σ^2	5	11	16.7

Table 12.10 reports the change in the maximized log likelihood from fitting four special cases of this model:

1. Common intercept and slope for the two logits: $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$
2. Common covariance matrix for the four doses: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \boldsymbol{\Sigma}_4$
3. Common covariance matrix and uncorrelated random effects
4. Univariate common variance component across dose: $u_{i(d)1} = u_{i(d)2}$ and $\sigma_d = \sigma$

Tests of the first three special cases against the general model (12.17) can use ordinary likelihood-ratio tests. Little seems to be lost by using the simpler model having uncorrelated random effects with homogeneous covariance structure (i.e., the fourth model listed in Table 12.10), as the likelihood-ratio statistic comparing this to model (12.17) equals 2(7.4) = 14.8 (df = 10). The model provides a separate univariate logistic-normal model for each conditional binomial outcome, specifying that the proportion of dead pups and the proportion of malformed pups (given survival) are independent, both within litter and marginally.

The univariate model in Table 12.10 is the special case of the third model listed in which the variances are common for the two logits and the random effects are perfectly correlated. Hence, it reduces to a univariate random effects model. Comparing the univariate model to a multivariate counterpart involves testing that correlation parameters fall on the boundary. Ordinary chi-squared asymptotic theory for likelihood-ratio tests applies only when the parameter falls in the interior of the parameter space. Tests when a null model has a correlation of 1 or a variance component of 0 are complex and beyond our scope here (see Section 12.6.6). However, an informal analysis of change in log likelihoods suggests that the univariate model is inadequate.

The ML estimated effects for the separate univariate logistic-normal model for each conditional binomial outcome are $\hat{\beta}_1 = 0.08$ (SE = 0.21), $\hat{\beta}_2 = 1.79$ (SE = 0.22). For a given cluster, there is no evidence of a dose effect on the death rate, but the estimated odds of malformation, given survival, multiply by exp(1.79) = 6.0 for every additional g/kg of ethylene

glycol. The variance component estimates suggest a stronger litter effect for the malformation outcome given survival ($\hat{\sigma}_2 = 1.6$) than for death ($\hat{\sigma}_1 = 0.5$).

12.5.3 Hierarchical (Multilevel) Modeling

Hierarchical data structures, with units grouped at different levels, are common in education. A statewide study of factors that affect student performance might measure students' scores on a battery of exams but use a model that takes into account the student, the school or school district, and the county. Just as two observations on the same student might tend to be more alike than observations on different students, so might two students in the same school tend to be more alike than students from different schools. Student, school, and county terms might be treated as random effects, with different ones referring to different *levels* of the model. For instance, a model might have students at level 1, schools at level 2, and counties at level 3. GLMMs for data having a hierarchical grouping of this sort are called *multilevel models*. Random effects enter the model at each level of the hierarchy.

We illustrate with a two-level model. Let $\pi_{i(j)t}$ denote the probability that student i in school j passes test t in a battery of tests. A multilevel model with random effects for student and school and fixed effects for explanatory variables has the form

$$\text{logit}[\pi_{i(j)t}] = \mathbf{x}'_{i(j)t}\boldsymbol{\beta} + u_j + v_{i(j)}.$$

Here, the explanatory variables $\mathbf{x}$ might include one that identifies the test in the battery. The random effects u_j for schools and $v_{i(j)}$ for students within schools are independent with different variance components. The level 1 random effects $\{v_{i(j)}\}$ account for variability among students in ability or parents' socioeconomic status or other characteristics not measured by $\mathbf{x}$. When they have a relatively large variance component, there is a strong correlation among the test results for students. The level 2 random effects $\{u_j\}$ account for variability among schools due to possibly unmeasured factors such as per-capita expenditure in the school's budget.

For examples of the use of multivariate random effects in multilevel modeling, see Aitkin et al. (1981), Anderson and Aitkin (1985), Gibbons and Hedeker (1997), Goldstein (1995), Goldstein and Rasbash (1996), and Longford (1993).

12.6 GLMM FITTING, INFERENCE, AND PREDICTION

Model fitting is rather complex for GLMMs. The main difficulty is that the likelihood function does not have a closed form. Numerical methods for approximating it can be computationally intensive for models with multivari-

ate random effects. In this section we outline the basic ideas of ML fitting of GLMMs. Some ML methods are available in software (e.g., PROC NLMIXED in SAS).

12.6.1 Marginal Likelihood and Maximum Likelihood Fitting

The GLMM is a two-stage model. At the first stage, conditional on the random effects, observations are assumed to follow a GLM. That is, observation y_{it} in cluster i has distribution in the exponential family with expected value μ_{it} linked to a linear predictor,

$$g(\mu_{it}) = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{u}_i .$$

Then, $\mathbf{z}'_{it}\mathbf{u}_i$ is a known offset and observations in a cluster are independent. At the second stage, the random effects $\{\mathbf{u}_i\}$ are assumed independent from a $N(\mathbf{0}, \boldsymbol{\Sigma})$ distribution.

For a discrete variable, denote the vector of all the observations by $\mathbf{y}$ and the vector of all the random effects by $\mathbf{u}$. Let $f(\mathbf{y}\,|\,\mathbf{u};\,\boldsymbol{\beta})$ denote the conditional mass function of $\mathbf{y}$, given $\mathbf{u}$. Let $f(\mathbf{u};\boldsymbol{\Sigma})$ denote the normal density function for $\mathbf{u}$. The likelihood function $\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma};\ \mathbf{y})$ for a GLMM is the probability mass function $f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\Sigma})$ of $\mathbf{y}$, viewed as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. This mass function refers to the marginal distribution of $\mathbf{y}$ after integrating out the random effects,

$$\ell(\boldsymbol{\beta},\boldsymbol{\Sigma};\mathbf{y}) = f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\Sigma}) = \int f(\mathbf{y}\,|\,\mathbf{u};\boldsymbol{\beta})f(\mathbf{u};\boldsymbol{\Sigma})\,d\mathbf{u}. \qquad (12.18)$$

It is often called a *marginal likelihood*. For example, the likelihood function $\ell(\boldsymbol{\beta}, \sigma^2;\mathbf{y})$ for the logistic-normal model (12.5) (absorbing α into $\boldsymbol{\beta}$) is

$$\prod_i \left(\int_{-\infty}^{\infty} \prod_t \left[\frac{\exp(\mathbf{x}'_{it}\boldsymbol{\beta} + u_i)}{1 + \exp(\mathbf{x}'_{it}\boldsymbol{\beta} + u_i)} \right]^{y_{it}} \left[\frac{1}{1 + \exp(\mathbf{x}'_{it}\boldsymbol{\beta} + u_i)} \right]^{1-y_{it}} f(u_i;\sigma^2)\,du_i \right).$$

The likelihood function is evaluated numerically and maximized as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. Many methods have been developed to do this. We next discuss a few of the most popular.

12.6.2 Gauss–Hermite Quadrature Methods

The integral determining the likelihood function has dimension that depends on the random effects structure. When the dimension is small, as in the one-dimensional integral above for the logistic-normal model (12.5), standard numerical integration methods can approximate the likelihood function.

Gauss–Hermite quadrature is a method for approximating the integral of a function $f(\cdot)$ multiplied by another function having the shape of a normal density. The approximation is a finite weighted sum that evaluates the function at certain points. In the univariate normal random effects case, the approximation has the form

$$\int_{-\infty}^{\infty} f(u)\exp(-u^2)\,du \approx \sum_{k=1}^{q} c_k f(s_k)\,,$$

with *weights* $\{c_k\}$ and *quadrature points* $\{s_k\}$ that are tabulated. The approximation improves as q, the number of quadrature points, increases.

The approximated likelihood can be maximized with standard algorithms such as Newton–Raphson, yielding ML estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$. Inverting an approximation for the observed information matrix provides standard errors for the ML estimates. For complex models, second partial derivatives for the Hessian may be computed numerically rather than analytically. Adequate approximation usually requires larger q for standard errors than for $\hat{\boldsymbol{\beta}}$. We recommend sequentially increasing q until the changes are negligible in both the estimates and standard errors.

An adaptive version of Gauss–Hermite quadrature (e.g., Liu and Pierce 1994) centers the quadrature points with respect to the mode of the function being integrated and scales them according to the estimated curvature at the mode. This improves efficiency, dramatically reducing the number of quadrature points needed to approximate the integrals effectively. Lesaffre and Spiessens (2001) showed comparisons and warned against using too few points.

12.6.3 Monte Carlo Methods

Multivariate forms of Gauss–Hermite quadrature handle multivariate, correlated random effects. Adequate approximation becomes more difficult, however, when the dimension of the integral exceeds roughly 5. Then, Monte Carlo methods are more feasible computationally than numerical integration. Various Monte Carlo approaches have been studied (e.g., McCulloch 1997), including Monte Carlo in combination with Newton–Raphson, Monte Carlo in combination with the EM algorithm, and simulating the likelihood directly. Here, we briefly describe a Monte Carlo EM (MCEM) algorithm.

The EM algorithm is a popular iterative method of finding ML estimates when data are missing or when filling in some "missing" data simplifies a likelihood (Dempster et al. 1977) [see Laird (1998) for a useful review]. In each cycle an *E*-step takes an expectation over the missing data to approximate the likelihood function and an *M*-step maximizes the likelihood given the working values of the parameter estimates. In GLMMs, one regards the random effects $\mathbf{u}$ as missing data. Then, $h(\mathbf{y},\mathbf{u};\boldsymbol{\beta},\boldsymbol{\Sigma}) = f(\mathbf{y}\,|\,\mathbf{u};\boldsymbol{\beta})f(\mathbf{u};\boldsymbol{\Sigma})$ specifies the joint distribution of the complete data. The *E*-step in iteration r

of the EM algorithm calculates

$$E\{\log h(\mathbf{y}, \mathbf{u}; \boldsymbol{\beta}, \boldsymbol{\Sigma}) \mid \mathbf{y}; \boldsymbol{\beta}^{(r)}, \boldsymbol{\Sigma}^{(r)}\}.$$

The expectation is with respect to the distribution of $(\mathbf{u} \mid \mathbf{y})$ with parameter values equal to $\boldsymbol{\beta}^{(r)}$ and $\boldsymbol{\Sigma}^{(r)}$, the working estimates for iteration r. The distribution of $(\mathbf{u} \mid \mathbf{y})$ follows from those of $(\mathbf{y} \mid \mathbf{u})$ and $\mathbf{u}$ in the GLMM via Bayes' theorem. The *M*-step then maximizes the result with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ to obtain $\boldsymbol{\beta}^{(r+1)}$ and $\boldsymbol{\Sigma}^{(r+1)}$.

The MCEM algorithm approximates the expectation in the *E*-step using Monte Carlo methods. Possible ways of doing this include using independent simulations from the distribution of $(\mathbf{u} \mid \mathbf{y})$, at the current estimate of parameters, or using Markov chain Monte Carlo (MCMC). For details, including the issue of choosing an appropriate Monte Carlo sample size, see Booth and Hobert (1999), Chan and Kuk (1997), and McCulloch (1994, 1997).

12.6.4 Penalized Quasi-likelihood Approximation

The Gauss–Hermite and Monte Carlo integration methods provide likelihood approximations such that resulting parameter estimates converge to the ML estimates as they are applied more finely (i.e., as the number of quadrature points increases for numerical integration and as the Monte Carlo sample size increases in the MCEM method). This contrasts with other approximate methods that are simpler but need not yield estimates near the ML estimates. These methods maximize an analytical approximation of the likelihood function.

Recall that the likelihood function (12.18) results from integrating out the random effects $\mathbf{u}$ from the joint distribution of $\mathbf{y}$ and $\mathbf{u}$. Using the exponential family representation of each component of that joint distribution, the integrand of (12.18) is an exponential function of $\mathbf{u}$. One approach approximates that function using a second-order Taylor series expansion of its exponent around a point $\tilde{\mathbf{u}}$ at which the first-order term equals 0. [That point $\tilde{\mathbf{u}} \approx E(\mathbf{u} \mid \mathbf{y})$.] The approximating function for the integrand is then exponential with quadratic exponent in $(\mathbf{u} - \tilde{\mathbf{u}})$ and has the form of a constant multiple of a multivariate normal density. Thus, its integral has closed form. This type of integral approximation is called a *Laplace approximation.* The approximation for integral (12.18) is then treated as a likelihood and maximized with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$.

For one such method (Breslow and Clayton 1993), the integral approximation yields a function approximating the log likelihood that has the form

$$q(\boldsymbol{\beta}, \mathbf{y}) - (1/2)\tilde{\mathbf{u}}' \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{u}},$$

where $q(\boldsymbol{\beta}, \mathbf{y})$ resembles a quasi-log-likelihood function for the GLM conditional on $\mathbf{u} = \tilde{\mathbf{u}}$. Thus, the approximation results in a penalty for the quasi-log likelihood, with the penalty increasing as elements of $\tilde{\mathbf{u}}$ increase in absolute value. This approach is called *penalized quasi-likelihood* (PQL). The calculations for maximizing the penalized quasi-likelihood use methods for linear mixed models with a normal response. This treats a linearization of the logit as a working response and entails iterative solution of sets of likelihood-like equations in $\boldsymbol{\beta}$ and $\mathbf{u}$. PQL methods do not require numerical or Monte Carlo integration and so are simpler than ML methods. They are computationally feasible for large data sets and models with complex random effects structure.

Unfortunately, PQL methods can perform poorly relative to ML (McCulloch 1997). For instance, for the abortion example in Section 12.3.2, the PQL approximations to the ML estimates (obtained using the GLIMMIX macro in SAS) are decent for $\{\beta_t\}$, but the standard errors and the estimate of σ are only about half what they should be (e.g., PQL gives $\hat{\sigma} = 4.3$, compared to the ML estimate of 8.6). When true variance components are large, ordinarily PQL tends to produce variance component estimates with substantial negative bias (Breslow and Lin 1995). The PQL estimators also behave poorly when the response distribution is far from normal (e.g., binary). Adjustments have been developed for some cases to lessen the bias (e.g., Goldstein and Rasbash 1996), but where possible we recommend using ML rather than PQL.

12.6.5 Bayesian Approaches

Another approach to fitting of GLMMs is Bayesian. With it, the distinction between fixed and random effects no longer occurs, as every effect has a probability distribution. Use of a flat prior distribution yields a posterior that is a constant multiple of the likelihood function. Then, Markov chain Monte Carlo (MCMC) methods for approximating intractable posterior distributions can approximate the likelihood function (Zeger and Karim 1991). For instance, an approximation for the mode of the posterior distribution approximates the ML estimate.

A danger is that improper prior distributions have improper posteriors for many models for categorical data (Natarajan and McCulloch 1995). In using MCMC, one may fail to realize that the posterior is improper. It is safer to use a proper but relatively diffuse prior. However, the posterior mode need not be close to the ML estimate, and Markov chains may converge slowly (Natarajan and McCulloch 1998). This is currently an active area of research, not just as a way of approximating ML results but also as an approach preferred over ML by those who adopt the Bayesian paradigm. See, for instance, Daniels and Gatsonis (1999) for multilevel modeling of geographic and temporal trends with clustered longitudinal binary data, which built on earlier hierarchical modeling by Wong and Mason (1985).

12.6.6 Inference for Model Parameters

After fitting the model, inference about fixed effects proceeds in the usual way. For instance, likelihood-ratio tests can compare nested models. Asymptotics for GLMMs apply as the number of clusters increases, rather than as the numbers of observations within the clusters increase. Similarly, resampling methods such as the bootstrap using a large number of clusters should sample clusters rather than individual observations within clusters, to preserve the within-cluster dependence.

Inference about random effects (e.g., their variance components) is more complex. For instance, sometimes one model is a special case of another in which a variance component equals 0. The simpler model then falls on the boundary of the parameter space relative to the more complex model, so ordinary likelihood-based inference does not apply. The asymptotic distribution of the likelihood-ratio statistic is known for the most common situation, testing H_0: $\sigma^2 = 0$ against H_a: $\sigma^2 > 0$ for a model containing a single variance component. The null distribution is an equal mixture of χ_0^2 (i.e., degenerate at 0) and χ_1^2 random variables (Self and Liang 1987). The value of 0 occurs when $\hat{\sigma} = 0$, in which case the maximized likelihoods are identical under H_0 and H_a. When $\hat{\sigma} > 0$ and the observed test statistic equals t, the P-value for this large-sample test is $\frac{1}{2}P(\chi_1^2 > t)$, half the P-value that applies for χ_1^2 asymptotic tests. For testing more than one variance component, the mixture distribution becomes more complex, and it is simpler to use a score test (Lin 1997).

12.6.7 Prediction Using Random Effects

The use of random effects in a model implies heterogeneity of certain effects of interest, such as odds ratios. Estimated effects of interest are often then linear combinations of fixed and random effects. For example, in the clinical trial comparing two treatments with random effects for centers (Section 12.3.4), one can predict the probability of success for each treatment in each center and odds ratios in those centers.

Given the data, the conditional distribution of $(\mathbf{u}\,|\,\mathbf{y})$ contains the information about the random effects $\mathbf{u}$. A prediction for $\mathbf{u}$ is $E(\mathbf{u}\,|\,\mathbf{y})$, its *posterior mean* given the data. Calculation of $E(\mathbf{u}\,|\,\mathbf{y})$ itself requires numerical integration or Monte Carlo approximation. The expectation depends on $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, so in practice one substitutes $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$ in the approximation. The standard error of the predictor of the random effect u_i is the standard deviation of the distribution of $(u_i\,|\,\mathbf{y})$. When one substitutes $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$ in $E(\mathbf{u}\,|\,\mathbf{y})$, however, the standard error does not account for the sampling variability in those estimates. Hence, the true standard error tends to be underestimated (Booth and Hobert 1998).

This approach to prediction using posterior means of random effects provides effect estimates that exhibit shrinkage relative to estimates using

only data in the specific cluster. In this sense the results are similar to those using an *empirical Bayes* approach (Ten Have and Localio 1999). This adapts an ordinary Bayesian analysis by using the sample data to estimate parameters of the prior distribution. For a vector of mean parameters, this approach yields an estimate of a particular mean that is a weighted average of the sample mean and the overall mean of the sample means. Thus, it shrinks the sample mean toward the overall mean. Shrinkage estimators can be far superior to sample values when the sample size for estimating each parameter is small, when there are many parameters to estimate, or when the true parameter values are roughly equal. The empirical Bayes paradigm has been in use for some time: for instance, for estimating a vector of means or binomial proportions (Efron and Morris 1975).

Although random effects models are natural in many applications, further work is needed. Work continues on the development of methodology for model-fitting and inference with complex GLMMs. In addition, research is needed on model checking and diagnostics. Nonetheless, we believe that GLMMs provide a very useful extension of ordinary GLMs.

NOTES

Section 12.1: Random Effects Modeling of Clustered Categorical Data

12.1. For further discussion of the Rasch model and ways of estimating its parameters, see Andersen (1980, Sec. 6.4) and Fischer and Molenaar (1995). Haberman (1977b) showed ML estimators can achieve consistency when both n and T grow at suitable rates. For multinomial Rasch extensions, see Andersen (1980, pp. 272–284; 1995) and Conaway (1989). Early work on random effects models for a categorical response includes Anderson and Aitkin (1985), Bartholomew (1980), Bock and Aitkin (1981), Chamberlain (1980), Gilmour et al. (1985), Pierce and Sands (1975), and Stiratelli et al. (1984).

12.2. In models with covariates, Neuhaus and Lesperance (1996) noted that conditional ML may lose considerable efficiency compared to the random effects approach when cluster sizes are small and covariates have strong positive within-cluster correlation. As that correlation approaches $+1$, the covariate effect resembles a between-cluster one, which the conditional ML approach cannot estimate. The matched-pairs case referred to in Section 12.1.2 in which the conditional ML estimate equals the random effects estimate has within-cluster covariate correlation $= -1$, as depending on the order of viewing the observations, x_t changes from 0 to 1 or from 1 to 0; then, no efficiency loss occurs.

Section 12.3: Examples of Random Effects Models for Binary Data

12.3. For further discussion of modeling capture–recapture data, see Bishop et al. (1975, Chap. 6), Chao et al. (2001), Cormack (1989), Coull and Agresti (1999), Darroch et al. (1993), Fienberg et al. (1999), and Hook and Regal (1995). Similarities exist between this problem and the related problem of estimating the binomial index n when observing independent $\text{bin}(n, \pi)$ counts with unknown n and π; see Aitkin and Stasinopoulos (1989) and references therein. Relatively flat log likelihoods also occur with other models that permit capture heterogeneity (Burnham and Overton 1978), such as a beta-binomial model.

12.4. King (1997) used random effects models as part of a solution for analyzing aggregated categorical data, the problem of *ecological inference*. Chambers and Steel (2001) discussed early work by Leo Goodman on this problem and proposed a simpler semiparametric approach.

Section 12.4: Random Effects Models for Multinomial Data

12.5. With the complementary log-log link, the likelihood function has closed form with a log gamma random effects distribution (Crouchley 1995, Farewell 1982, Ten Have 1996).

12.6. Chen and Kuo (2001) discussed nominal responses, including discrete choice models (Sec. 7.6) with random effects. See also Brownstone and Train (1999) for discrete choice GLMMs.

Section 12.5: Multivariate Random Effects Models for Binary Data

12.7. Rabe-Hesketh and Skrondal (2001) showed that careful attention must be paid to parameter identification in models with multivariate random effects. Their factor model contains many multivariate random effects models as special cases.

12.8. For longitudinal bivariate binary responses, Ten Have and Morabia (1999) simultaneously modeled bivariate log odds ratios and univariate logits. Multivariate responses sometimes have both continuous and categorical components. For random effects modeling of such data, see Catalano and Ryan (1992) and Gueorguieva and Agresti (2001).

Section 12.6: GLMM Fitting, Inference, and Prediction

12.9. See Fahrmeir and Tutz (2001, Chap. 7) and McCulloch and Searle (2001) for more details on the fitting of GLMMs. Just as the likelihood function for a GLMM is an integral, so do likelihood equations have the form of integral equations (McCulloch and Searle 2001, p. 227). Wolfinger and O'Connell (1993) described a fitting method related to PQL, also motivated by a Laplace approximation.

12.10. A GLMM determines the marginal relationship (averaged over random effects) between the mean response and explanatory variables. Conversely, Heagerty (1999) noted that a marginal model for the mean implicitly determines the form of the fixed portion of the linear predictor in a conditional model. The conditional GLMM (12.1) has linear predictor, $\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_{it}'\mathbf{u}_i$. A more general form $\Delta_{it} + \mathbf{z}_{it}'\mathbf{u}_i$ implies a particular marginal model. Here, Δ_{it} is a function of the marginal linear predictor and the random effects distribution. It is implicitly defined by the integral equation that links the marginal and conditional means.

PROBLEMS

Applications

12.1 Refer to the matched-pairs data of Table 10.14 and Problem 10.1.

a. Fit model (12.3). Interpret $\hat{\beta}$. If your software uses numerical integration, report $\hat{\beta}$, $\hat{\sigma}$, and their standard errors for 5, 10, 25, 100, and 200 quadrature points, and comment on convergence.

b. Compare $\hat{\beta}$ and its SE for this approach to the conditional ML approach.

12.2 Refer to Table 4.8 on the free-throw shooting of Shaq O'Neal. In game i, suppose that $y_i =$ number made out of n_i attempts is a bin(n_i, π_i) variate and $\{y_i\}$ are independent.

a. Fit the model, logit$(\pi_i) = \alpha$. Find and interpret $\hat{\pi}_i$. Does the model appear to fit adequately?

b. Fit the model, logit$(\pi_i) = \alpha + u_i$, where $\{u_i\}$ are independent $N(0, \sigma^2)$. Use $\hat{\alpha}$ and $\hat{\sigma}$ to summarize O'Neal's free-throw shooting.

c. Explain how the model in part (a) is a special case of that in part (b). Is there evidence that the one in part (b) fits better?

12.3 For Table 8.3, let $y_{it} = 1$ when subject i used substance t. Table 12.11 shows output for the logistic-normal model

$$\text{logit}\left[P\left(Y_{it} = 1 \,|\, u_i\right)\right] = u_i + \beta_t .$$

Interpret. Illustrate by comparing use of cigarettes and marijuana.

TABLE 12.11 Output for Problem 12.3

Description	Value			Std	
Subjects	2276	Parameter	Estimate	Error	t Value
Max Obs Per Subject	3	beta1	4.2227	0.1824	23.15
Parameters	4	beta2	1.6209	0.1207	13.43
Quadrature Points	200	beta3	−0.7751	0.1061	−7.31
Log Likelihood	−3311	sigma	3.5496	0.1627	21.82

12.4 How is the focus different for the model in Problem 12.3 than for the loglinear model (AC, AM, CM) used in Section 8.2.4? If $\hat{\sigma} = 0$, which loglinear model has the same fit as the GLMM?

12.5 For the student survey in Table 9.1, (**a**) analyze using GLMMs, and (**b**) compare results and interpretations to those with marginal models in Problem 11.2.

12.6 Fit model (12.10) to the responses on abortion. If your software uses Gauss–Hermite quadrature, report the approximate number of quadrature points needed for parameter estimates to converge and the number needed for standard error estimates to converge. (This example has large $\hat{\sigma}$ and requires many points.)

12.7 For the crossover study in Table 11.10 (Problem 11.6), fit the model

$$\text{logit}\left[P\left(Y_{i(k)t} = 1 \,|\, u_{i(k)}\right)\right] = \alpha_k + \beta_t + u_{i(k)}, \qquad (12.19)$$

where $\{u_{i(k)}\}$ are independent $N(0, \sigma^2)$. Interpret $\{\hat{\beta}_t\}$ and $\hat{\sigma}$.

12.8 For Problem 12.7, compare estimates of $\beta_B - \beta_A$ and $\beta_C - \beta_A$ and SE values to those using (**a**) a marginal model (Problem 11.6), and (**b**) conditional logistic regression (Section 10.2), treating subject terms in model (12.19) as fixed effects.

12.9 For Problem 12.7, fit the more general GLMM having treatment effects $\{\beta_{tk}\}$ that vary by sequence. Test whether the fit is better. One could also consider period or carryover effects. Add two period effects to model (12.19) (e.g., the first-period-effect parameter adds to the model when $t = A$ and $k = 1, 2$, $t = B$ and $k = 3$, 4, and $t = C$ and $k = 5, 6$). Check whether the fit improves. Interpret.

12.10 Consider the logistic-normal model (12.10) for the abortion opinion data, under the constraint $\sigma = 0$.

a. Explain why the fit is the same as an ordinary logit model treating the three responses for each subject as if they were independent responses for three separate subjects.

b. Explain why the model fit is the same as an ordinary loglinear model (GI_1, GI_2, GI_3) of mutual independence of responses on the three items (I_1, I_2, I_3), given G = gender.

c. Fit the model. Interpret, and explain why $\{\hat{\beta}_t - \hat{\beta}_u\}$ are quite different from those in Section 12.3.2 allowing $\sigma > 0$.

12.11 For Table 6.7 on admissions decisions for graduate school applicants, let $y_{ig} = 1$ denote a subject in department i of gender g (1 = females, 0 = males) being admitted.

a. For the fixed effects model, $\text{logit}[P(Y_{ig} = 1)] = \alpha + \beta g + \beta_i^D$, $\hat{\beta} = 0.173$ (SE = 0.112). Interpret.

b. The corresponding model (12.12) in which departments are a normal random effect has $\hat{\beta} = 0.163$ (SE = 0.111). Interpret.

c. The model of form (12.12) allowing the gender effect to vary by department has $\hat{\beta} = 0.176$ (SE = 0.132), with $\hat{\sigma}_b = 0.20$. Interpret. Explain why the standard error of $\hat{\beta}$ is slightly larger than with the other analyses.

d. The marginal sample log odds ratio between gender and whether admitted equals -0.07. How could this take different sign from $\hat{\beta}$ in these models?

e. The sample conditional odds ratios between gender and whether admitted vary between 0 and ∞. By contrast, predicted odds ratios for the interaction random effects model do not vary much. Explain why results can be so different.

12.12 For the clinical trial in Table 9.16, let $\pi_{it} = P(Y_{it} = 1 \mid u_i)$ denote the probability of success for treatment t in center i.

a. The random intercept model (12.11) has $\hat{\beta} = 1.52$ (SE = 0.70) and $\hat{\sigma} = 1.9$. Interpret.

b. From Section 9.8.3, the fixed effects analog of this model (replacing $\alpha + u_i$ by α_i) has $\hat{\alpha}_1 = \hat{\alpha}_3 = -\infty$, corresponding to $\hat{\pi}_{1t} = \hat{\pi}_{3t} = 0$ for each treatment. By contrast, the random effects model has $\hat{\alpha} + \hat{u}_1 = -3.78$ (using NLMIXED in SAS) and $\hat{\pi}_{11} = 0.047$ and $\hat{\pi}_{12} = 0.011$ in center 1. Explain how this model can have $\hat{\pi}_{it} > 0$ in centers having no successes.

12.13 Refer to the subject-specific model in Section 12.3.3. Verify that the estimated difference in time effect slopes between the new and standard drugs for treating depression are (**a**) 1.018 (SE = 0.192) with the GLMM approach, and (**b**) 1.156 (SE = 0.222) with conditional ML.

12.14 For marginal model (10.14) for Table 10.5 on premarital and extramarital sex, Table 12.12 shows results of fitting a corresponding random intercept model. Interpret $\hat{\beta}$. Compare estimates of and inferences about β to those in Section 10.3.2 for the marginal model.

TABLE 12.12 Output for Problem 12.14

				Std	
Subjects	475	Parameter	Estimate	Error	t Value
Max Obs Per Subject	2	inter1	−1.5422	0.1826	−8.45
Parameters	5	inter2	−0.6682	0.1578	−4.24
Quadrature Points	100	inter3	0.9273	0.1673	5.54
Log Likelihood	−890.1	beta	4.1342	0.3296	12.54
		sigma	2.0757	0.2487	8.35

12.15 A data set from the 1994 General Social Survey on subjects' opinions on four items (the environment, health, law enforcement, education) related to whether they believed government spending on each item should increase, stay the same, or decrease. Subjects were also classified by their gender and race. For subject i, let $G_i = 1$ for females and 0 for males, let $R_{1i} = 1$ for whites and 0 otherwise,

$R_{2i} = 1$ for blacks and 0 otherwise, and $R_{1i} = R_{2i} = 0$ for the other category of race. Let y_{it} denote the response for subject i on spending item t, where outcomes (1, 2, 3) represent (increase, stay the same, decrease).

a. With constraint $\beta_4 = 0$, the random-intercept model

$$\text{logit}\left[P(Y_{it} \le j | u_i)\right] = \alpha_j + \beta_t + \beta_g G_i + \beta_{r1} R_{i1} + \beta_{r2} R_{2i} + u_i, \quad j = 1, 2,$$

has $\hat{\beta}_1 = -0.55$, $\hat{\beta}_2 = -0.60$, $\hat{\beta}_3 = -0.49$, with $\hat{\sigma} = 1.03$. These estimates are greater than five standard errors in absolute value. Interpret.

b. Table 12.13 shows results with a race-by-item interaction. Interpret.

TABLE 12.13 Results for Problem 12.15[a]

Variable	Estimate	SE
Intercept-1	1.065	0.391
Intercept-2	1.919	0.051
Gender	0.409	0.088
Race1-w	−0.055	0.397
Race2-b	0.434	0.452
Item1-envir	−0.357	0.539
Item2-health	−0.319	0.493
Item3-crime	−0.585	0.480
Race1 * Item1	−0.170	0.549
Race1 * Item2	−0.387	0.503
Race1 * Item3	0.197	0.491
Race2 * Item1	−0.452	0.606
Race2 * Item2	0.454	0.598
Race2 * Item3	−0.518	0.560

[a] Coding 0 for item 4 (education) and race 3 (other).

12.16 Refer to Problem 11.12 for Table 8.19 on government spending. Analyze these data using a cumulative logit model with random effects. Interpret. Compare results to those with a marginal model (Problem 11.12).

12.17 For the insomnia example in Section 12.4.2, according to SAS the maximized log likelihood equals −593.0, compared to −621.0 for the simpler model forcing $\sigma = 0$. Compare models, using either a likelihood-ratio test or AIC. What do you conclude?

TABLE 12.14 Results for Problem 12.18

Observer Effect	GEE	Random Effects
A	−0.451 (0.108)	−1.201 (0.300)
B	−0.391 (0.093)	−0.919 (0.299)
C	0.319 (0.118)	0.558 (0.301)
D	0.632 (0.105)	1.545 (0.313)
E	−0.491 (0.098)	−1.379 (0.300)
F	1.252 (0.161)	2.907 (0.344)

12.18 Landis and Koch (1977) showed ratings by seven pathologists who separately classified 118 slides regarding the presence and extent of carcinoma of the uterine cervix, using a five-point ordinal scale. (Table 13.1 is a collapsing of their table that combines the first two categories and the last three categories.) For slide i with rater t, Table 12.14 shows results of fitting model

$$\text{logit}\left[P(Y_{it} \le j \mid u_i)\right] = u_i + \alpha_j + \beta_t$$

to the ordinal table (with $\hat{\beta}_G = 0$), assuming that the $\{u_i\}$ are independent $N(0, \sigma^2)$. It also shows GEE estimates, using independence working equations, for the corresponding marginal model. Interpret $\hat{\beta}_F$ for each model. Explain why estimates using the random effects model, for which $\hat{\sigma} = 3.8$, tend to be much larger in absolute value. Discuss the differences in assumptions and interpretations for the two models.

12.19 Refer to Section 12.5.1 on boys' attitudes toward the leading crowd. Table 12.15 shows results for a sample of schoolgirls. Fit model (12.16) and interpret. Summarize the estimated variability and correlation of random effects.

TABLE 12.15 Data for Problem 12.19

(M, A) for First Interview	(M, A) for Second Interview[a]			
	(Yes, Positive)	(Yes, Negative)	(No, Positive)	(No, Negative)
Yes, positive	484	93	107	32
Yes, negative	112	110	30	46
No, positive	129	40	768	321
No, negative	74	75	303	536

[a] M, membership; A, attitude.

Source: J. S. Coleman, *Introduction to Mathematical Sociology* (London: Free Press of Glencoe, 1964), p. 168.

12.20 Generalize model (12.16) to apply simultaneously to Tables 12.8 and 12.15, using a gender main effect but the same membership effect and the same attitude effect for each gender. Fit the model. Use the maximized log likelihood to compare with a more general model having different membership effects and different attitude effects for each gender. Interpret.

12.21 Table 12.16 reports results from a study to estimate the number N of people infected during a 1995 hepatitis A outbreak in Taiwan. The 271 observed cases were reported from records based on a serum test taken by the Institute of Preventive Medicine of Taiwan (P), records reported by the National Quarantine Service (Q), and records based on questionnaires administered by epidemiologists (E). Estimating N is difficult, because many subjects had only one capture.

a. Find $\hat{N}$ if you observed only **(i)** P and Q, **(ii)** P and E, **(iii)** Q and E.

b. Find $\hat{\mathrm{N}}$ using the model of mutual independence with P, Q, and E.

c. Find a 95% profile likelihood interval for N using the model in part (b).

d. The random effects model of Section 12.3.6 has fit shown in Table 12.16, for which $\hat{\sigma} = 2.9$. The log-likelihood is relatively flat, and $\hat{\mathrm{N}} = 4551$ with a 95% profile likelihood interval of $(758, \infty)$ (Coull and Agresti 1999). Explain why this model may provide imprecise estimates of N. Since the interval in part (c) is much narrower, is it necessarily more reliable?

TABLE 12.16 Data for Problem 12.21

P Q E	Observed Count	Logistic-Normal ML Fit
0 0 0	—	$(487, \infty)$
0 0 1	63	61.0
0 1 0	55	58.0
0 1 1	18	17.0
1 0 0	69	68.0
1 0 1	17	20.0
1 1 0	21	19.0
1 1 1	28	28.0

Source: Data from Chao et al. (2001).

12.22 Analyze the crossover data of Table 11.1 using a random effects approach. Interpret, and compare results to those in Section 11.1.2.

12.23 The analyses in Section 12.3.2 comparing opinions on some topic extend to ordinal responses. Using an ordinal random effects model, analyze the 4^3 table in Agresti (1993), found also at the book's Web site, *www.stat.ufl.edu/~ aa/cda/cda.html*.

12.24 The analyses in Section 12.3.4 describing heterogeneity in multicenter clinical trials extend to ordinal responses. Using random effects models, analyze the $2 \times 3 \times 8$ table in Hartzel et al. (2001a).

12.25 You are a statistical consultant asked to analyze Table 4 in B. Efron, *Statistical Science* **13**: 95–122 (1998), which shows 2×2 tables from a clinical trial in 41 cities. Analyze, and write a report summarizing your analysis.

12.26 Analyze Table 11.9 with age and maternal smoking as predictors using a **(a)** logistic-normal model, **(b)** marginal model, and **(c)** transitional model. Explain how the interpretation of the maternal smoking effect differs for the three approaches.

Theory and Methods

12.27 Refer to Section 12.3.1. Using supplementary information improves predictions. Let q_i denote the true proportion of votes for Clinton in state i in the 1992 election, conditional on voting for him or Bush. Consider the model

$$\text{logit}\big[P(Y_{it} = 1 \,|\, u_i)\big] = \text{logit}(q_i) + \alpha + u_i,$$

where $\{q_i\}$ are known and $\{u_i\}$ are independent $N(0, \sigma^2)$. When $\hat{\sigma} = 0$, show $\hat{\pi}_i = q_i \exp(\hat{\alpha})/[1 - q_i + q_i \exp(\hat{\alpha})]$. Compared to $\{q_i\}$, explain how $\hat{\pi}_i$ then shifts up or down depending on how the overall Democratic vote compares in the current poll to the previous election (i.e., depending on $\hat{\alpha}$). When also $\hat{\alpha} = 0$, show $\hat{\pi}_i = q_i$.

12.28 For a binary response, consider the random effects model

$$\text{logit}\big[P(Y_{it} = 1 \,|\, u_i)\big] = \alpha + \beta_t + u_i, \qquad t = 1, \ldots, T,$$

where $\{u_i\}$ are independent $N(0, \sigma^2)$, and the marginal model

$$\text{logit}\big[P(Y_t = 1)\big] = \alpha + \beta_t^*, \qquad t = 1, \ldots, T.$$

For identifiability, $\beta_T = \beta_T^* = 0$. Explain why all $\beta_t = 0$ implies that all $\beta_t^* = 0$. Is the converse true?

12.29 The GLMM for binary data using probit link function is

$$\Phi^{-1}\big[P(Y_{it}=1\,|\,\mathbf{u}_i)\big] = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{u}_i,$$

where Φ is the $N(0,1)$ cdf and $\mathbf{u}_i$ has $N(\mathbf{0}, \boldsymbol{\Sigma})$ pdf, $f(\mathbf{u}_i; \boldsymbol{\Sigma})$.

a. Show that the marginal mean is

$$P(Y_t = 1) = \int P(Z - \mathbf{z}'_{it}\mathbf{u}_i \le \mathbf{x}'_{it}\boldsymbol{\beta}) f(\mathbf{u}_i; \boldsymbol{\Sigma}) d\mathbf{u}_i,$$

where Z is a standard normal variate that is independent of $\mathbf{u}_i$.

b. Since $Z - \mathbf{z}'_{it}\mathbf{u}_i$ has a $N(0, 1 + \mathbf{z}'_{it}\boldsymbol{\Sigma}\mathbf{z}_{it})$ distribution, deduce that

$$\Phi^{-1}\big[P(Y_t = 1)\big] = \mathbf{x}'_{it}\boldsymbol{\beta}\big[1 + \mathbf{z}'_{it}\boldsymbol{\Sigma}\mathbf{z}_{it}\big]^{-1/2}.$$

Hence, the marginal model is a probit model with attenuated effect. In the univariate random intercept case, show the marginal effect equals that from the GLMM divided by $\sqrt{1+\sigma^2}$.

12.30 In the Rasch model, $\text{logit}[P(Y_{it}=1)] = \alpha_i + \beta_t$, α_i is a fixed effect.

a. Assuming independence of responses for different subjects and for different observations on the same subject, show that the log likelihood is

$$\sum_i \sum_t \alpha_i y_{it} + \sum_i \sum_t \beta_t y_{it} - \sum_i \sum_t \log\big[1 + \exp(\alpha_i + \beta_t)\big].$$

b. Show that the likelihood equations are $y_{+t} = \sum_i P(Y_{it}=1)$ and $y_{i+} = \sum_t P(Y_{it}=1)$ for all i and t. Explain why conditioning on $\{y_{i+}\}$ yields a distribution that does not depend on $\{\alpha_i\}$.

c. Discuss advantages and disadvantages of, instead, treating α_i as random.

12.31 Consider the matched-pairs random effects model (12.3). For given β_0, let δ_0 be such that $\hat{\mu}_{12} = n_{12} + \delta_0$ and $\hat{\mu}_{21} = n_{21} - \delta_0$ satisfies $\log(\hat{\mu}_{21}/\hat{\mu}_{12}) = \beta_0$. Suppose $\{\hat{\mu}_{ij}\}$ has nonnegative log odds ratio. Explain why:

a. This is the fit of the model assuming $\beta = \beta_0$.

b. The likelihood-ratio statistic for testing H_0: $\beta = \beta_0$ in this model equals

$$2\left(n_{12}\log\frac{n_{12}}{n_{12}+\delta_0} + n_{21}\log\frac{n_{21}}{n_{21}-\delta_0}\right).$$

c. The likelihood-ratio test of H_0: $\beta = 0$ is the test of symmetry.

12.32 Explain why the logistic-normal model is not helpful for capture–recapture experiments with only two captures.

12.33 Refer to the crossover study in Problem 12.7. Kenward and Jones (1991) reported results using the ordinal response scale (none, moderate, complete) for relief. Explain how to formulate an ordinal logit random effects model for these data analogous to model (12.19).

12.34 Formulate a model using adjacent-categories logits that is analogous to model (12.14) for cumulative logits. Interpret parameters.

12.35 For ordinal square $I \times I$ tables of counts $\{n_{ab}\}$, model (12.3) for binary matched-pairs responses (Y_{i1}, Y_{i2}) for subject i extends to

$$\text{logit}\big[P(Y_{it} \le j | u_i)\big] = \alpha_j + \beta x_t + u_i$$

with $\{u_i\}$ independent $N(0, \sigma^2)$ variates and $x_1 = 0$ and $x_2 = 1$.

a. Explain how to interpret β, and compare to the interpretation of β in the corresponding marginal model (10.14).

b. This model implies model (12.3) for each 2×2 collapsing that combines categories 1 through j for one outcome and categories $j + 1$ through I for the other. Use the form of the conditional ML (or random effects ML) estimator for binary matched pairs to explain why

$$\log\left[\left(\sum_{a>j}\sum_{b<j} n_{ab}\right) \Big/ \left(\sum_{a<j}\sum_{b>j} n_{ab}\right)\right]$$

is a consistent estimator of β.

c. Treat these $(I - 1)$ collapsed 2×2 tables naively as if they are independent samples. Show that adding the numerators and adding the denominators of the separate estimates of e^{β} motivates the summary estimator of β,

$$\tilde{\beta} = \log\left\{\left[\sum_{a>b}(a-b)n_{ab}\right] \Big/ \left[\sum_{b>a}(b-a)n_{ab}\right]\right\}.$$

Explain why $\tilde{\beta}$ is consistent for β even recognizing the actual dependence.

d. A standard error for $\tilde{\beta}$ that treats the collapsed tables in part (c) as independent is inappropriate. Treating $\{n_{ab}\}$ as a multinomial sample, show that an estimated asymptotic variance of $\tilde{\beta}$ is (Agresti

and Lang 1993a)

$$\left\{ \sum_{b>a} (b-a)^2 n_{ab} \Big/ \left[\sum_{b>a} (b-a) n_{ab} \right]^2 \right\}$$
$$+ \left\{ \sum_{a>b} (a-b)^2 n_{ab} \Big/ \left[\sum_{a>b} (a-b) n_{ab} \right]^2 \right\}.$$

12.36 Summarize advantages and disadvantages of using a GLMM approach compared to a marginal model approach. Describe conditions under which parameter estimators are consistent for (**a**) marginal models using GEE, (**b**) marginal models using ML, (**c**) GLMM using PQL, and (**d**) GLMM using ML.

Categorical Data Analysis, Second Edition. Alan Agresti

ISBN: 0-471-36093-7

CHAPTER 13

Other Mixture Models for Categorical Data*

In Chapters 10 through 12 we introduced ways of handling correlated observations due to repeated measurement and other forms of clustering. The generalized linear mixed models (GLMMs) of Chapter 12 assume normal random effects. They describe heterogeneity by replacing the linear predictor by a normally distributed mixture of linear predictors. In this chapter we present additional models having connections with GLMMs. Except for one case, these models use nonnormal mixture distributions.

In Section 13.1 we present latent class models. These treat a contingency table as a mixture of unobserved tables at categories of a qualitative latent (unobserved) variable. In Section 13.2 we discuss a related nonparametric approach to fitting GLMMs that uses an unspecified discrete quantitative distribution for the random effects distribution.

In Section 13.3 we model clustered binomial responses using the beta distribution to describe heterogeneity of binomial parameters. The resulting beta-binomial distribution has variance function for which quasi-likelihood methods are also available. In Section 13.4 we model count responses using the gamma distribution to describe heterogeneity of Poisson parameters. The resulting negative binomial regression model corresponds to a Poisson GLMM having a log-gamma distributed random effect. It is an alternative to the GLMM for Poisson responses with normal random effects, a model discussed in Section 13.5.

13.1 LATENT CLASS MODELS

GLMMs create a mixture of linear predictor values using a latent variable, the unobserved random effect vector, having a normal distribution. By contrast, latent class models use a mixture distribution that is qualitative rather than quantitative. The basic model assumes existence of a latent

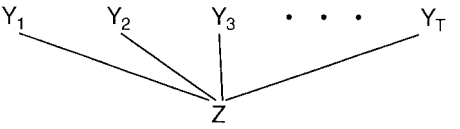


FIGURE 13.1 Association graph for latent class model.

categorical variable such that the observed response variables are conditionally independent, given that variable.

For categorical response variables $(Y_1, Y_2, \ldots, Y_T)$, the latent class model assumes a latent categorical variable Z such that for each possible sequence of response outcomes $(y_1, \ldots, y_T)$ and each category z of Z,

$$P(Y_1 = y_1, \ldots, Y_T = y_T \mid Z = z) = P(Y_1 = y_1 \mid Z = z) \cdots P(Y_T = y_T \mid Z = z).$$

Figure 13.1 shows the association graph for the model. A latent class model summarizes probabilities of classification $P(Z = z)$ in the latent classes as well as conditional probabilities $P(Y_t = y_t \mid Z = z)$ of outcomes for each Y_t within each latent class. These are the model parameters. More generally, the latent variable $\mathbf{Z}$ can be multivariate. The model is an analog for categorical responses and latent variables of the factor analysis model for multivariate normal responses.

The latent class model is sometimes plausible when the observed variables are several indicators of some concept, such as prejudice, religiosity, or opinion about an issue. An example is Table 10.13, in which subjects gave their opinions about whether abortion should be legal in various situations. Perhaps an underlying latent variable describes one's basic attitude toward legalized abortion, such that given the value of that latent variable, responses on the observed variables are conditionally independent. For instance, the latent variable may be a qualitative variable with three categories: One class for those who always oppose legalized abortion regardless of the situation, one for those who always favor it, and one for those whose response depends on the situation.

The T-dimensional contingency table cross classifying $(Y_1, \ldots, Y_T)$ is observed. The $(T + 1)$-dimensional table that cross-classifies it with the latent variable is an unobserved table. Denote the number of categories of each Y_t by I and the number of latent classes of Z by q. For the observed table, let $\pi_{y_1, \ldots, y_T} = P(Y_1 = y_1, \ldots, Y_T = y_T)$. The model assumes a multinomial distribution over its I^T cells. For a given cell,

$$\pi_{y_1, \ldots, y_T} = \sum_{z=1}^{q} P(Y_1 = y_1, \ldots, Y_T = y_T \mid Z = z) P(Z = z).$$

The conditional independence factorization for the latent class model states that

$$\pi_{y_1,\ldots,y_T} = \sum_{z=1}^{q}\left[\prod_{t=1}^{T} P(Y_t = y_t \mid Z = z)\right]P(Z = z). \tag{13.1}$$

This is a nonlinear model for the I^T multinomial probabilities.

13.1.1 Fitting Latent Class Models

Denote the counts in the observed table by $\{n_{y_1,\ldots,y_T}\}$. Summing over the I^T cells in that table, the kernel of the multinomial log likelihood is

$$\sum n_{y_1,\ldots,y_T} \log \pi_{y_1,\ldots,y_T}. \tag{13.2}$$

Substituting parameters from (13.1), one can maximize (13.2) with respect to those parameters using Newton–Raphson (Haberman 1979, Chap. 10) or the EM algorithm (Goodman 1974). It is helpful to note that the latent class model states that the loglinear model symbolized by $(Y_1Z, Y_2Z, \ldots, Y_TZ)$ holds for the unobserved table. The model makes no assumption about the $\{Y_tZ\}$ associations but assumes that the $\{Y_t\}$ are mutually independent within each category of Z.

The EM algorithm has two steps in each iteration. The E (expectation) step in iteration s calculates pseudo-counts $\{n^{(s)}_{y_1,\ldots,y_T,z}\}$ for the unobserved table using $\{n_{y_1,\ldots,y_T}\}$ and a working conditional distribution for $(Z \mid Y_1, \ldots, Y_T)$ described shortly. The M (maximization) step treats $\{n^{(s)}_{y_1,\ldots,y_T,z}\}$ as data and applies an algorithm such as iterative reweighted least squares or IPF for fitting the model (i.e., the loglinear model $(Y_1Z, Y_2Z, \ldots, Y_TZ)$). The fit $\{\mu^{(s)}_{y_1,\ldots,y_T,z}\}$ of that model in the unobserved table then determines the new working conditional distribution of $(Z \mid Y_1, \ldots, Y_T)$ to apply to $\{n_{y_1,\ldots,y_T}\}$ for the E-step of the next iteration. This allocates the observed data to pseudo-counts in the unobserved cells in proportion to this fit, using

$$n^{(s+1)}_{y_1,\ldots,y_T,z} = n_{y_1,\ldots,y_T}\frac{\mu^{(s)}_{y_1,\ldots,y_T,z}}{\sum_{k=1}^{q}\mu^{(s)}_{y_1,\ldots,y_T,k}}.$$

These are entries in the unobserved table for iteration $(s + 1)$. They are used as pseudo-data for the M-step of iteration $(s + 1)$.

Eventually, the algorithm converges to fitted values for the unobserved table that provide fitted probabilities that satisfy mutual independence within each latent class, and such that the corresponding fitted probabilities in the observed table (i.e., added over the latent categories) maximize the likelihood (13.2). The fitted probabilities in the unobserved table are an estimated joint

distribution for $(Y_1, \ldots, Y_T, Z)$. One can use them to calculate the ML estimates of the latent class model parameters $\{P(Y_t = y_t \mid Z = z)\}$ and $\{P(Z = z)\}$.

The EM algorithm is computationally simple and relatively stable. Each iteration increases the likelihood. However, its convergence can be slow. See Laird (1998) for a review. The log likelihood for a latent class model may have local maxima. Thus, with either the Newton–Raphson or EM algorithm, it is advisable to perform the fitting process a few times with different starting guesses for the parameter values. The EM algorithm tends to be less sensitive to the choice of starting values. Thus, some software begins with the EM algorithm and then switches to the Newton–Raphson algorithm as it approaches the ML estimates to speed the process. As q increases, multiple local maxima are more likely and the danger increases of a lack of identifiability.

Standard errors for model parameter estimates result from inverting the model's estimated information matrix. This is a by-product of the Newton–Raphson algorithm but not the EM algorithm. One way to obtain standard errors with it applies a useful formula of Louis (1982) for the observed information when using the EM algorithm. It equals the expected value of the observed information for the loglinear model for the unobserved table minus the expected value of the information for the conditional distribution of Z given the observed data. Baker (1992) and Lang (1992) gave related results.

Chi-squared statistics comparing observed cell counts to fitted values test the model fit. The residual df $= I^T - qT(I - 1) - q$. This follows since multinomial model (13.1) describes $I^T - 1$ multinomial probabilities using $(I - 1)$ parameters $\{P(Y_t = y_t \mid Z = z), y_t = 1, \ldots, I - 1\}$ at each of qT combinations of z and t, and $q - 1$ parameters $\{P(Z = z)\}$. Often, the nature of the variables suggests a value for q, usually quite small (2 to 4). Otherwise, the usual procedure starts with $q = 2$; if the fit is inadequate, it increases by steps of 1 as long as the fit shows substantive improvement. Specialized software exists for such models (Appendix A).

13.1.2 Latent Class Model for Rater Agreement

Table 13.1 is an expanded data set of the example in Section 10.5. Seven pathologists classified each of 118 slides on the presence or absence of carcinoma in the uterine cervix. For modeling interobserver agreement, the conditional independence assumption of the latent class model is often plausible. With a blind rating scheme, ratings of a given subject or unit by different pathologists are independent. If subjects having true rating in a given category are relatively homogeneous, then ratings by different pathologists may be nearly independent within a given true rating class. Thus, one might posit a latent class model with $q = 2$ classes, one for subjects whose true rating is positive and one for subjects whose true rating is negative. This

TABLE 13.1 Diagnoses of Carcinoma and Fits of Latent Class Models[a]

Pathologist								Fit		
A	B	C	D	E	F	G	Count	$q = 1$	$q = 2$	$q = 3$
0	0	0	0	0	0	0	34	1.1	23.0	33.8
0	0	0	0	1	0	0	2	1.6	6.6	2.0
0	1	0	0	0	0	0	6	2.2	12.7	6.3
0	1	0	0	0	0	1	1	2.8	1.7	1.5
0	1	0	0	1	0	0	4	3.3	3.6	3.0
0	1	0	0	1	0	1	5	4.2	0.5	4.7
1	0	0	0	0	0	0	2	1.4	3.0	2.1
1	0	1	0	1	0	1	1	1.6	0.2	0.2
1	1	0	0	0	0	0	2	2.8	1.7	1.3
1	1	0	0	0	0	1	1	3.5	0.3	1.6
1	1	0	0	1	0	0	2	4.2	0.5	2.9
1	1	0	0	1	0	1	7	5.3	3.7	6.5
1	1	0	0	1	1	1	1	1.4	2.6	1.4
1	1	0	1	0	0	1	1	1.3	0.1	0.1
1	1	0	1	1	0	1	2	2.0	4.3	2.6
1	1	0	1	1	1	1	3	0.5	3.1	2.0
1	1	1	0	1	0	1	13	3.3	11.5	9.6
1	1	1	0	1	1	1	5	0.9	8.4	8.7
1	1	1	1	1	0	1	10	1.2	13.5	13.6
1	1	1	1	1	1	1	16	0.3	9.9	12.3

[a] Fits obtained with Latent Gold (Statistical Innovations, Belmont MA). 1, yes; 0, no.

Source: Based on data in Landis and Koch (1977), not showing empty cells.

model expresses the 2^7 joint distribution of the seven ratings as a mixture of two 2^7 distributions, one for each true rating class.

Table 13.2 shows results of fitting some latent class models (including a mixture model studied in Section 13.2.4). Because the observed table is sparse, the deviance is mainly useful for comparing models. This is an informal comparison, though, since the chi-squared distribution does not apply for comparing deviances of models with different numbers of latent classes. A model with q classes is a special case of a model with $q^* > q$ classes in which $P(Z = z) = 0$ for $z > q$ and hence falls on the boundary of the parameter space. Ordinary chi-squared likelihood-ratio tests require parameters to fall in the interior of the parameter space (i.e., $0 < P(Z = z) < 1$ for $z = 1, \ldots, q^*$).

Table 13.1 also shows the fitted values for latent class models with $q = 1, 2, 3$, for the cells having positive counts. (Each empty cell also has a fitted value, not shown here). The model with $q = 1$ latent class is the model of mutual independence of the seven ratings. Equivalently, it is the loglinear model $(Y_1, Y_2, \ldots, Y_7)$. It fits poorly, as one would expect. With $q = 2$, considerable evidence remains of lack of fit. For instance, the fitted count for

TABLE 13.2 Likelihood-Ratio Statistics for Latent Class Models Fitted to Table 13.1[a]

Number of Latent Classes	Model	Deviance (G^2) Statistic	df
1	Mutual independence	476.8	120
2	Latent class	62.4	112
	Rasch mixture	67.6	118
3	Latent class	15.3	104
	Rasch mixture	27.5	116
4	Latent class	6.4	96
	Rasch mixture (quasi-symmetry)	23.7	114

[a]Models fitted with Latent Gold (Statistical Innovations, Belmont, MA).

a negative rating by each pathologist is 23.0, compared to an observed count of 34. (The small G^2 that Table 13.2 reports for this model does not imply a good fit; in Section 9.8.4 we noted that G^2 tends to be highly conservative when most fitted values are very close to 0.) The model with $q = 3$ seems to fit adequately.

Studying the estimated probability $P(Y_t = 1 \mid Z = z)$ of a carcinoma diagnosis for each pathologist, conditional on a given latent class z, helps illuminate the nature of these classes. Table 13.3 reports these for the three-class model. They suggest that (1) the first latent class refers to cases that all pathologists (except occasionally B) agree show no carcinoma; (2) the third latent class refers to cases in which A, B, E, and G agree show carcinoma and C and D usually agree; and (3) the second latent class refers to cases of strong disagreement, whereby C, D, and F rarely diagnose carcinoma but B, E, and G usually do. The estimated proportions in the three latent classes are $\hat{P}(Z = 1) = 0.37$, $\hat{P}(Z = 2) = 0.18$, and $\hat{P}(Z = 3) = 0.45$. The model estimates that 18% of the cases fall in the problematic class.

TABLE 13.3 Estimated Probabilities of Diagnosing Carcinoma, for Latent Class Model and Rasch Mixture Model with Three Classes[a]

Model	Latent Class	Pathologist						
		A	B	C	D	E	F	G
Latent Class	1	0.057	0.138	0.000	0.000	0.055	0.000	0.000
	2	0.513	1.00	0.000	0.058	0.751	0.000	0.631
	3	1.000	0.981	0.858	0.586	1.000	0.476	1.000
Rasch Mixture	1	0.022	0.150	0.001	0.000	0.047	0.000	0.022
	2	0.611	0.923	0.052	0.015	0.774	0.009	0.611
	3	0.994	0.999	0.853	0.617	0.997	0.483	0.994

[a]Results obtained with Latent Gold (Statistical Innovations, Belmont, MA).

A danger with latent variable models, shared by factor analysis for continuous responses, is the temptation to interpret latent variables too literally. In this example it is tempting to treat latent class 1 (latent class 3) as cases truly without carcinoma (with carcinoma). Thus, it is tempting to treat a rating of no carcinoma (a rating of carcinoma) given that the subject falls in latent class 1 (latent level 3) as necessarily being a correct judgment. One should realize the tentative nature of the latent variable. Be careful not to make the error of reification—treating an abstract construction as if it has actual existence (Gould 1981).

Using model parameter estimates and Bayes' theorem, one can also estimate $P(Z = z \mid Y_t = y_t)$ and $P(Z = z \mid Y_1 = y_1, \ldots, Y_T = y_T)$. If a pathologist makes a "yes" rating, for instance, what is the estimated probability that the subject is in the latent class for which agreement on a positive rating usually occurs? We perform further analysis in Section 13.2.5 after studying a simpler model.

Espeland and Handelman (1989), Uebersax (1993), Uebersax and Grove (1990, 1993), and Yang and Becker (1997) presented various latent variable models for rater agreement and diagnostic accuracy. One could also use methods of Chapters 11 and 12, such as a model with a continuous rather than qualitative latent variable. A logistic-normal random intercept model, for instance, yields subject-specific comparisons of $P(Y_t = 1)$ for various t.

13.1.3 Latent Class Models for Capture–Recapture

We next apply latent class models to capture–recapture modeling for estimating population size. In Section 12.3.6 a logistic-normal GLMM was used for this. With T sampling occasions, a 2^T contingency table displays the data, with scale (captured, not captured) at each occasion. A prediction of the population size equals the prediction for the missing cell count, representing subjects not captured at every occasion, added to the counts in other cells.

With two classes, the latent class model treats the population as a mixture of two types, perhaps determined by genetic or environmental factors. Homogeneity of capture probabilities occurs for subjects within each type, but the type of any given subject is unknown. This model represents a compromise between the mutual independence model, which assumes a single latent class and complete homogeneity, and the logistic-normal GLMM, which assumes a continuous mixture of capture probabilities rather than two classes.

We illustrate with the $T = 6$-capture data set on snowshoe hares in Table 12.6. The model of mutual independence predicts that $\hat{N} = 75$. Its 95% profile-likelihood confidence interval for N is (70, 83). The latent class model with two classes has $\hat{N} = 85$ and a profile-likelihood confidence interval of (74, 106). The latent class model with three classes gives similar results. Since the logistic-normal GLMM in Section 12.3.6 gave the interval (75, 154), these seem too short to be trusted. This simple latent class model may not capture

all the existing heterogeneity. It is more plausible to assume a continuous latent variable than a discrete one with a couple of classes. We'll analyze these data further with related models in the next section.

13.2 NONPARAMETRIC RANDOM EFFECTS MODELS

In spite of its popularity and attractive features, the normality assumption for random effects in ordinary GLMMs can rarely be closely checked. For instance, in studying normal GLMMs, Verbeke and Lesaffre (1996) noted that under a normality assumption for random effects, their predicted values often appear normally distributed even when the true values are generated from a highly nonnormal distribution. An obvious concern of this or any parametric assumption for the random effects is possibly harmful effects of misspecification. To check sensitivity to this assumption, one can fit GLMMs using alternative or more general random effects assumptions.

13.2.1 Logit Models with Unspecified Random Effects Distribution

A nonparametric approach (e.g., Aitkin 1999) guards against possibly harmful misspecification effects. This uses an unspecified random effects distribution on a finite set of mass points. The location of the mass points and their probabilities are parameters. The number of mass points can be fixed. When this number is itself unknown, one treats it as fixed in the estimation process but increases it sequentially until the likelihood is maximized. The maximization usually requires relatively few mass points. Even allowing a continuous mixture distribution, the nonparametric estimate of that distribution takes a finite number of points (e.g., Lindsay et al. 1991). In fact, fitting a model having only two mass points often results in fixed effects estimates quite similar to those with the full maximization. This approach is useful primarily when the random effects distribution is not itself of direct interest, since the nonparametric estimate of that distribution tends to be poor even for very large samples.

Model fitting is actually simpler than for models with normal random effects, since the integral that determines the likelihood function simplifies to a finite sum. In Section 13.2.4 we discuss this point with a Rasch-type model. Specialized software can fit nonparametric mixture models (Appendix A). However, this approach also has disadvantages. For instance, with multivariate random effects it cannot provide simple correlation structure as the normal can. Standard inference does not apply for comparing models with different numbers of mass points, since one model is on the boundary of the parameter space compared to the other. Also, the ML estimate of the random effects distribution often places some weight at $\pm\infty$. Although this can be useful with binary data for identifying a subsample for which the estimated response probability equals 1 or equals 0 for all observations in a

cluster, it is not then possible to describe heterogeneity with an estimated variance component.

To illustrate this approach, we reanalyze Table 10.13 on attitudes about legalized abortion. In Section 12.3.2 we fitted the logistic-normal model (12.10),

$$\text{logit}\big[P(Y_{it} = 1 \mid u_i)\big] = u_i + \beta_t + \gamma x, \tag{13.3}$$

with x = gender and parameters $\{\beta_t\}$ representing three conditions under which abortion might be legal. Treating u_i instead nonparametrically, the likelihood maximizes with a two-point mixture distribution. Estimated abortion item effects are $\hat{\beta}_1 - \hat{\beta}_3 = 0.83$ (SE = 0.16), $\hat{\beta}_2 - \hat{\beta}_3 = 0.30$ (SE = 0.16), and $\hat{\beta}_1 - \hat{\beta}_2 = 0.52$ (SE = 0.16). Results are similar to those that Table 12.3 shows for the normal random effects approach (Section 12.3.2).

13.2.2 Nonparametric Mixing of Logistic Regression

Follman and Lambert (1989) presented an example with a prespecified number of mass points. They analyzed the effect of the dosage of a poison on the probability of death of a protozoan of a particular genus. Table 13.4 shows the data. They assumed two unobserved types of that genus.

Let $\pi_i(x)$ denote the probability of death at log dose level x for genus type i, $i = 1, 2$. Let ρ denote the probability a protozoan belongs to genus type 1. Their model specifies

$$\pi(x) = \rho\pi_1(x) + (1 - \rho)\pi_2(x), \qquad \text{where} \quad \text{logit}\big[\pi_i(x)\big] = \alpha_i + \beta x,$$

with unknown ρ. The curve for $\pi(x)$ is a weighted average of two curves having the same shapes but different intercepts.

The ordinary logistic regression model is the special case $\rho = 1$. Its fit, $\text{logit}[\hat{\pi}(x)] = -68.4 + 42.1x$ (with SE = 3.8 for $\hat{\beta} = 42.1$), is poor, with deviance $G^2 = 24.7$ (df = 6). The fit of the mixture model is

$$\hat{\pi}(x) = 0.34\hat{\pi}_1(x) + 0.66\hat{\pi}_2(x),$$

with

$$\text{logit}\big[\hat{\pi}_1(x)\big] = -196.2 + 124.8x, \qquad \text{logit}\big[\hat{\pi}_2(x)\big] = -205.7 + 124.8x,$$

TABLE 13.4 Number of Protozoa Exposed to Poison Dose and Number That Died

Poison Dose	Exposed	Dead	Poison Dose	Exposed	Dead
4.7	55	0	5.1	53	22
4.8	49	8	5.2	53	37
4.9	60	18	5.3	51	47
5.0	55	18	5.4	50	50

Source: Follman and Lambert (1989). Reprinted with permission from the *Journal of the American Statistical Association*.

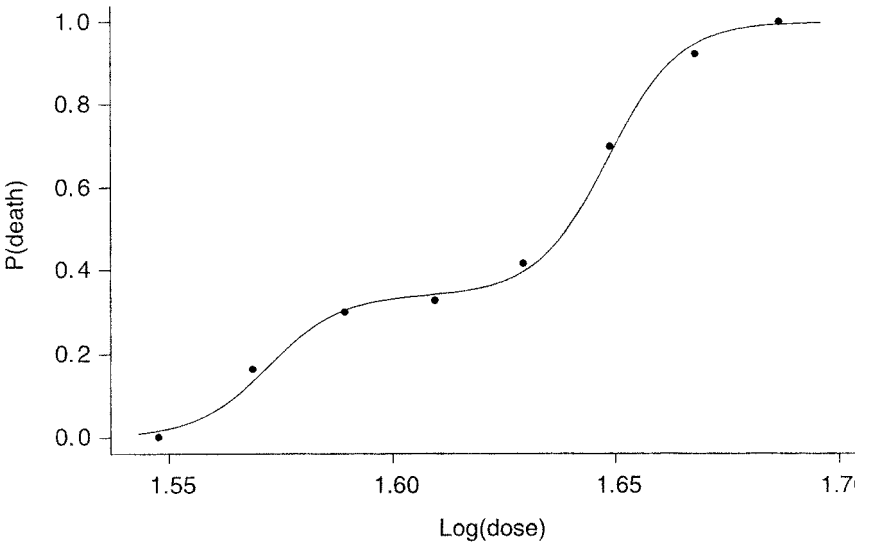


FIGURE 13.2 Fit of binary mixture of logistic regressions to Table 13.4 [model fitted using Latent Gold (Statistical Innovations, Belmont, MA)].

and SE = 25.2 for $\hat{\beta} = 124.8$. Figure 13.2 shows the fit. This is much better, with $G^2 = 3.4$ (df = 4); that is, double the maximized log-likelihood increases by $24.7 - 3.4 = 21.3$ by adding two parameters: an additional intercept and the probability for the mixture. Follman and Lambert noted that with eight dose levels, at most two mixture points are identifiable for this model.

The ordinary GLMM assumes a normal mixture of logistic curves. It gives a deviance reduction of only 1.7 compared to the ordinary logistic model with $\rho = 1$.

13.2.3 Is Misspecification a Serious Problem?

Is it worth the trouble to consider alternatives to the normality assumption for random effects in GLMMs, whether they be parametric or nonparametric? Not much work exists on investigating misspecification effects. For logistic random intercept models, different assumptions for the random effects distribution often provide similar results for estimating the regression effects. Choosing an incorrect random effects distribution does not tend to bias estimators of those effects. The true distribution for the random effects being skewed can result in some bias for the normal intercept estimator (Neuhaus et al. 1992). The choice of random effects distribution also usually has little impact on efficiency of estimation.

When the true random effects distribution is dramatically far from normal, there can be some efficiency loss for the logistic-normal estimator. This can

happen when the true distribution is a two-point mixture with large variance component. B. Caffo and I studied this with various models, such as a simple one-way random effects model. In cluster i, let y_{it} be a Bernoulli variate satisfying

$$\text{logit}\big[P(Y_{it} = 1 \,|\, u_i)\big] = \alpha + u_i, \qquad i = 1, \ldots, n, \quad t = 1, \ldots, T, \quad (13.4)$$

where $\text{var}(u_i) = \sigma^2$. Simulated samples from this model used various n, T, α, and σ, and various true distributions for u_i including normal, uniform, exponential, and binary. Usually, assuming normality does not hurt when the true distribution is nonnormal. Also, using a nonparametric approach when the true distribution is normal does not result in much efficiency loss [Neuhaus and Lesperance (1996) noted this for a related model.] However, when the true distribution is a two-point mixture, the normal approach loses efficiency in estimating $\{\mu_i = P(Y_{it} = 1 \,|\, u_i)\}$ as σ and T increase. For example, when $n = T = 30$, $\alpha = 0$, and the mixture has probability 0.5 at each point, the expected value of $|\hat{\mu}_i - \mu_i|$ is $(0.06, 0.05)$ for the (normal, nonparametric) approach when $\sigma = 0.5$, $(0.06, 0.02)$ when $\sigma = 1.0$, and $(0.04, 0.01)$ when $\sigma = 2.0$. Differences for estimating α are less dramatic.

The example from Follman and Lambert (1989) discussed in Section 13.2.2, which has a covariate but $T = 1$, illustrates the potential efficiency loss with the logistic-normal GLMM. The two-point mixture model has $\hat{\beta} = 124.8$ with SE $= 25.2$, for which $\hat{\beta}/\text{SE} = 4.9$. The normal mixture model has $\hat{\beta} = 65.5$ with SE $= 19.5$, for which $\hat{\beta}/\text{SE} = 3.4$.

Our study suggested that the random effects distribution has to be rather extremely nonnormal for the normal GLMM to suffer in bias or efficiency. However, Heagerty and Zeger (2000) (see also McCulloch 1997) noted that other types of misspecification can be more crucial. Regarding bias, they argued that sensitivity to the random effects assumption is greater for estimating regression parameters in random effects models than estimating their counterparts in corresponding marginal models. They illustrated this with a model violation by which the variance of the random effects depends on values of covariates. They concluded that between-cluster effects may be more sensitive to correct specification of the random effects distribution than within-cluster effects. This is an advantage of using marginal models for between-cluster effects.

13.2.4 Rasch Mixture Model

From Section 12.1.4, for subject i with item t the Rasch model for a binary response is

$$\text{logit}\big[P(Y_{it} = 1 \,|\, u_i)\big] = u_i + \beta_t, \quad t = 1, \ldots, T. \qquad (13.5)$$

The GLMM treats $\{u_i\}$ as normal random effects. Lindsay et al. (1991) studied this model when u_i instead can assume only a finite number q of values. Denote the distribution of the latent variable u_i, which is the same for all i, by

$$P(U = a_k) = \rho_k, \quad k = 1, \ldots, q,$$

for unknown $\{a_k\}$ and $\{\rho_k\}$. For identifiability one can either place a constraint on this distribution, such as $\sum_k \rho_k a_k = 0$, or on $\{\beta_t\}$. This model is called a *Rasch mixture model*.

Like other random effects models, the Rasch mixture model is a latent variable model. The random effect u_i is unobserved, and the T responses are assumed conditionally independent at each fixed u_i value. It differs from the ordinary latent class model for binary responses having q latent classes (Section 13.1), since it assumes structure (13.5) for $P(Y_{it} = 1 \mid u_i)$ whereas latent class model (13.1) assumes no structure for $P(Y_t = y_t \mid Z = z)$.

This model is simpler to fit than GLMMs with normal random effects because the GLMM's intractable integral that determines the likelihood function is replaced by a finite sum. The marginal probability of a sequence of responses $(y_1, \ldots, y_T)$ is

$$\pi_{y_1, \ldots, y_T} = \sum_{k=1}^{q} \rho_k \left[\prod_{t=1}^{T} \frac{\exp[y_t(a_k + \beta_t)]}{1 + \exp(a_k + \beta_t)} \right].$$

Substituting this in the multinomial log likelihood (13.2), ML estimation of $\{a_k, \rho_k\}$ and $\{\beta_t\}$ can proceed using Newton–Raphson or EM algorithms. As q increases, the maximized likelihood increases and the fit improves. However, Lindsay et al. (1991) showed that with T items, the likelihood no longer changes once $q = (T + 1)/2$. Then, the model gives the same fit to the 2^T observed table as the quasi-symmetry model (10.33). Thus, this simpler latent class model has a symmetric conditional association structure among the observed variables. Arminger et al. (2000) extended the Rasch mixture model to incorporate covariates.

13.2.5 Modeling Rater Agreement

For the ratings of carcinoma by seven pathologists (Table 13.1), Table 13.2 also summarizes the fit of Rasch mixture models. Here, $P(Y_{it} = 1 \mid u_i)$ in (13.5) denotes the probability of a carcinoma diagnosis for pathologist t evaluating slide i. With $q = 3$ (i.e., u_i can take 3 values), it does not fit significantly more poorly than the latent class model. With $T = 7$ raters, the discrete mixture can take at most $(T + 1)/2 = 4$ points. The model with $q = 4$ is equivalently the quasi-symmetry model. It does not seem to fit better than with $q = 3$.

Pathologist	F	D	C	A	G	E	B
Estimate	–3.70	–3.15	–1.87	1.48	1.48	2.26	3.52
Comparison							

FIGURE 13.3 Pathologist estimates for Rasch mixture model and results of 90% Bonferroni simultaneous comparison.

Figure 13.3 shows $\{\hat{\beta}_t\}$ for the Rasch mixture model with $q = 3$, setting $\sum_t \hat{\beta}_t = 0$. These describe variation among the pathologists' response distributions at each latent level. For a given latent class, for instance, the estimated odds of a carcinoma diagnosis for pathologist B are $\exp(3.52 - 1.48) = 7.7$ times the estimated odds for pathologist A. Pathologist B tends to make a carcinoma diagnosis most often, and D and F the least. The figure also shows results of a 90% Bonferroni comparison of the 21 pairs of pathologists, based on standard errors of pairwise differences $\hat{\beta}_t - \hat{\beta}_s$.

For pathologist t, conditional on latent level k for a slide,

$$\exp\left(\hat{a}_k + \hat{\beta}_t\right) \Big/ \left[1 + \exp\left(\hat{a}_k + \hat{\beta}_t\right)\right]$$

estimates the probability of a carcinoma diagnosis. Table 13.3 reports these, which use $\hat{a}_1 = -5.25$, $\hat{a}_2 = -1.02$, and $\hat{a}_3 = 3.63$. They are similar to the estimates for the ordinary latent class model but a bit smoother, with fewer estimates at the boundary. Again, at latent level 1 pathologists tend not to diagnose carcinoma, at level 2 many disagreements occur, and at level 3 pathologists tend to diagnose carcinoma. The estimated latent class proportions are $\hat{\rho}_1 = 0.37$, $\hat{\rho}_2 = 0.19$, and $\hat{\rho}_3 = 0.43$, with 19% of cases falling in the problematic class.

Model (13.5) implies that the association between each Y_t and U has log odds ratio $(a_k - a_\ell)$ for levels k and ℓ of U. For instance, in the third latent class the estimated odds that a pathologist diagnoses carcinoma are $\exp[3.63 - (-5.25)] > 7000$ times those in the first latent class. In terms of the estimated probabilities in Table 13.3, using pathologist A this is $\exp[(0.994/0.006)/(0.022/0.978)]$. The large $\{\hat{a}_k - \hat{a}_\ell\}$ suggest strong association between each pathologist's rating and the latent variable. This induces strong association between pairs of pathologist ratings. (The model-fitted odds ratios between pairs of raters vary between about 7 and 400.) However, the quite varied $\{\hat{\beta}_t\}$ suggest that substantial marginal heterogeneity exists among the seven ratings. This causes heterogeneity in pairwise levels of agreement.

The mutual independence model is the special case of the Rasch mixture model with $q = 1$; that is, $\rho_1 = 1$. For Table 13.1 the Rasch mixture model with $q = 3$ has only four more parameters than the mutual independence

model (i.e., ρ_k and a_k, $k = 1,2$). Yet it fits well and has simple interpretations. See Agresti and Lang (1993b) for further details and a simpler model that sets $a_1 - a_2 = a_2 - a_3$.

13.2.6 Other Models for Capture–Recapture

In Section 13.1.3 latent-class models were used for capture–recapture experiments. Alternatively, one could use the Rasch mixture model. Model (13.5) with two classes gives $\hat{N} = 77$ and a 95% profile-likelihood confidence interval of (71, 87). This seems too short to trust. It is more realistic to allow a continuous distribution for capture probabilities. Model (13.5) treating u_i as normal rather than binary does this, and in Section 12.3.6 we used it for these data.

So, which models might be used other than a parametric random effects model? One possibility is a loglinear model (Cormack 1989). This is a marginal model, applying to probabilities averaged over subjects. Let Y_t denote the binary capture variable for a randomly selected subject at occasion t, with categories (captured, not captured). The simplest model, denoted by $(Y_1, Y_2, \ldots, Y_T)$, assumes that capture events are mutually independent. This is equivalent to the logistic-normal model (13.5) with $\sigma = 0$ and latent class model (13.1) with $q = 1$. A more plausible model allows an association between pairs of capture variables. This is equivalently the loglinear model denoted $(Y_1Y_2, Y_1Y_3, \ldots, Y_{T-1}Y_T)$. Alternatively, a model with Markov structure such as $(Y_1Y_2, Y_2Y_3, \ldots, Y_{T-1}Y_T)$ may be useful. Usually, insufficient data exists to warrant using very complex loglinear models. For any such model, its fit for the $2^T - 1$ observed cells projects to the remaining cell to predict the number unobserved at every occasion.

A connection exists between nonparametric random effects and loglinear approaches. In Section 13.2.7 we show that assuming model (13.5) but using a nonparametric treatment of u_i implies a loglinear model of quasi-symmetric form for the marginal model. The quasi-symmetry model (10.33) itself is not useful for this problem, because *any* count in the missing cell is consistent with it. The model has an interaction parameter pertaining to that cell alone, which results in a likelihood equation equating that cell count to its fitted value. So, information in other cells does not help in the estimation of the expected frequency in that cell. However, special cases of quasi-symmetry are useful (Darroch et al. 1993). An example is the loglinear model with the same association for each pair of occasions. Like the logistic-normal model, this model of *exchangeable association* has only one more parameter than the mutual independence model.

For the snowshoe hare data of Table 12.6, the model with exchangeable two-factor association has $\hat{N} = 90.5$ and a confidence interval of (75, 125). This interval and the one of (71, 87) for the Rasch mixture model with $q = 2$ are substantially narrower than the interval (75, 154) for the logistic-normal model (Section 12.3.6). In capture–recapture experiments, $\hat{N}$ and the confi-

dence interval for N depend strongly on the choice of model. The problem is inherently one of prediction. Estimating N requires extrapolating from the observed numbers of subjects having $1, 2, \ldots, T$ captures to the number of subjects with 0 captures. Standard goodness-of-fit criteria are of limited help. Two models can fit the data well, yet yield quite different estimates for the unobserved count. For instance, for the snowshoe hare data, the loglinear models of mutual independence and of two-factor association both fit the observed cells relatively well ($G^2 = 58.3$, df = 56 for mutual independence and $G^2 = 32.4$, df = 41 for the two-factor model); however, their $\hat{N}$ values are 75 and 105.

Simpler models usually give narrower confidence intervals for N, through the usual benefits of model parsimony. This is not necessarily good. A narrow confidence interval for N is desirable, but not at the expense of severe sacrifice in the actual confidence level. Intervals based on a possibly unrealistic assumption of subject homogeneity may be overly optimistic. Simulations suggest that actual coverage probabilities are often well below nominal levels when even slight model misspecification occurs. Allowance for heterogeneity among subjects results in wider intervals. Severe population heterogeneity makes reaching useful conclusions difficult, as intervals can be very wide (Burnham and Overton 1978, Coull and Agresti 1999).

13.2.7 Nonparametric Mixtures and Quasi-symmetry

A distribution-free approach for u_i with the Rasch form of model (13.5) implies the quasi-symmetry loglinear model marginally (Darroch 1981; Tjur 1982). We now show this result, to which we alluded in Section 10.4.2.

Let $\mathbf{Y}_i$ denote the sequence of T responses for subject i. For possible outcomes $\mathbf{y} = (y_1, \ldots, y_T)$, where each $y_t = 1$ or 0,

$$P(\mathbf{Y}_i = \mathbf{y} \mid u_i) = \prod_t \left[\frac{\exp(u_i + \beta_t)}{1 + \exp(u_i + \beta_t)}\right]^{y_t} \left[\frac{1}{1 + \exp(u_i + \beta_t)}\right]^{1-y_t}$$

$$= \frac{\exp\left[u_i(\Sigma_t y_t) + \Sigma_t y_t \beta_t\right]}{\Pi_t\left[1 + \exp(u_i + \beta_t)\right]}.$$

Let F denote the cdf of u_i. The marginal probability of sequence $\mathbf{y}$ for a randomly selected subject is (suppressing the subject label)

$$\pi_{y_1, \ldots, y_T} = E_U P(\mathbf{Y} = \mathbf{y} \mid U) = \exp\left(\sum_t y_t \beta_t\right) \int \frac{\exp\left[u(\Sigma_t y_t)\right]}{\Pi_t\left[1 + \exp(u + \beta_t)\right]} dF(u) .$$

This probability contributes to the log likelihood, which is (13.2) for a multinomial distribution over the 2^T cells for possible $\mathbf{y}$. Regardless of the choice for F, the integral is complex. However, it depends on the data only

through $\sum_t y_t$. A more general model replaces this integral by a separate parameter for each value of $\sum_t y_t$. This model has form

$$\log \pi_{y_1,\ldots,y_T} = \sum_t y_t \beta_t + \lambda_{y_1+\cdots+y_t}. \tag{13.6}$$

The final term represents a separate parameter at each value of $\sum_t y_t$.

The implied marginal model (13.6) has interaction term that is invariant to a permutation of the response outcomes $\mathbf{y}$, since each such permutation yields the same sum, $\sum_t y_t$. Thus, it is the loglinear model of quasi-symmetry (10.33). No matter what form F takes, the marginal model has the same main effect structure, and it has an interaction term that is a special case of the one in (13.6). Thus, one can consistently estimate $\{\beta_t\}$ using the ordinary ML estimates for the loglinear model. In fact, Tjur (1982) showed that these estimates are also the conditional ML estimates, treating $\{u_i\}$ as fixed effects and conditioning on their sufficient statistics. The interaction parameters in model (13.6) result from the dependence in responses among variables, due to heterogeneity in $\{u_i\}$.

We illustrate for the opinions about legalized abortion analyzed in Sections 10.7.2 and 12.3.2 and with a nonparametric random effects approach in Section 13.2.1. For model (13.3), estimated within-subject comparisons $\beta_t - \beta_s$ of items result from fitting a quasi-symmetric loglinear model. Let $\mu_g(y_1, y_2, y_3)$ denote the expected frequency for gender g making response y_t to item t, $t = 1, 2, 3$, where for item t, $y_t = 1$ for approval of legalized abortion and 0 for disapproval. The loglinear model is

$$\log \mu_g(y_1, y_2, y_3) = \beta_1 y_1 + \beta_2 y_2 + \beta_3 y_3 + \gamma g + \lambda_{y_1+y_2+y_3}. \tag{13.7}$$

For $y_1 + y_2 + y_3 = k$, λ_k refers to all cells in which subjects voiced approval for k of the three items, $k = 0, 1, 2, 3$. The ML fit, which has $G^2 = 10.2$ with df = 9, yields $\hat{\beta}_1 - \hat{\beta}_2 = 0.521$ (SE = 0.154), $\hat{\beta}_1 - \hat{\beta}_3 = 0.828$ (SE = 0.160), and $\hat{\beta}_2 - \hat{\beta}_3 = 0.307$ (SE = 0.161). These are similar to the normal random effects estimates (Table 12.3) and nonparametric random effects estimates in Section 13.2.1. They also are the conditional ML estimates for model (13.3), treating $\{u_i\}$ as fixed. With this approach or conditional ML, however, one cannot estimate between-groups effects, such as the gender effect in model (13.7). [The γ parameter in model (13.7) refers to relative sample sizes of males and females and is not the same as the gender effect in (13.3).]

13.3 BETA-BINOMIAL MODELS

The beta-binomial model is a parametric mixture model that is another alternative to binary GLMMs with normal random effects. As with other

mixture models that assume a binomial distribution at a fixed parameter value, the marginal distribution permits more variation than the binomial. Thus, a model using the beta-binomial is a way to handle overdispersion occurring with ordinary binomial models.

13.3.1 Beta-Binomial Distribution

The beta-binomial distribution results from a beta distribution mixture of binomials. Suppose that (a) given π, Y has a binomial distribution, bin(n, π), and (b) π has a beta distribution.

The beta probability density function is

$$f(\pi; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1 - \pi)^{\beta-1}, \qquad 0 \le \pi \le 1, \quad (13.8)$$

with parameters $\alpha > 0$ and $\beta > 0$, for the gamma function $\Gamma(\cdot)$. Let

$$\mu = \frac{\alpha}{\alpha + \beta}, \quad \theta = 1/(\alpha + \beta).$$

The beta distribution for π has mean and variance

$$E(\pi) = \mu, \qquad \text{var}(\pi) = \mu(1 - \mu)\theta/(1 + \theta).$$

When α and β exceed 1.0, the distribution is unimodal, with skew to the right when $\alpha < \beta$, skew to the left with $\alpha > \beta$, and symmetry when $\alpha = \beta$. It simplifies to the uniform distribution when $\alpha = \beta = 1$.

Marginally, averaging with respect to the beta distribution for π, Y has the *beta-binomial distribution*. Its mass function is

$$p(y; \alpha, \beta) = \binom{n}{y} \frac{B(\alpha + y, n + \beta - y)}{B(\alpha, \beta)}, \qquad y = 0, 1, \ldots, n.$$

In terms of μ and θ, the beta-binomial mass function is

$$p(y; \mu, \theta) = \binom{n}{y} \frac{\left[\prod_{k=0}^{y-1}(\mu + k\theta)\right]\left[\prod_{k=0}^{n-y-1}(1 - \mu + k\theta)\right]}{\prod_{k=0}^{n-1}(1 + k\theta)}. \quad (13.9)$$

It is easier to understand the nature of this distribution from its moments than from its mass function. The first two moments are

$$E(Y) = n\mu, \qquad \text{var}(Y) = n\mu(1 - \mu)[1 + (n - 1)\theta/(1 + \theta)].$$

As $\theta \to 0$ in the beta distribution, $\text{var}(\pi) \to 0$ and that distribution converges to a degenerate distribution at μ. Then $\text{var}(Y) \to n\mu(1-\mu)$ and the beta-binomial distribution converges to the bin(n, μ).

13.3.2 Models Using the Beta-Binomial Distribution

Models using the beta-binomial distribution permit μ [and hence $E(Y)$] to depend on explanatory variables. The simplest models let θ be the same unknown constant for all observations. [Prentice (1986) considered extensions where it could also depend on covariates.] Like GLMs, models can use various link functions, but the logit is most common. For observation i with n_i trials, assuming that y_i has a beta-binomial distribution with index n_i and parameters (μ_i, θ), the model links μ_i to predictors by

$$\text{logit}(\mu_i) = \alpha + \boldsymbol{\beta}'\mathbf{x}_i.$$

The beta-binomial is not in the natural exponential family, even for known θ. Articles using beta-binomial models have employed a variety of fitting methods (Note 13.4). Crowder (1978) discussed the likelihood behavior for an ANOVA-type model. Hinde and Demétrio (1998) obtained the ML fit by iterating between solving the likelihood equations for the regression parameters $\boldsymbol{\beta}$, for fixed θ, and solving the likelihood equation for θ for fixed $\boldsymbol{\beta}$. Each part can use Newton–Raphson. McCulloch and Searle (2001, p. 61) showed the asymptotic covariance matrix of $(\hat{\mu}, \hat{\theta})$ and of $(\hat{\alpha}, \hat{\beta})$ for independent observations from a single beta-binomial distribution.

A related but simpler approach for overdispersed binary counts uses quasi-likelihood with similar variance function as the beta-binomial. The quasi-likelihood variance function is

$$v(\mu_i) = n_i\,\mu_i(1-\mu_i)\left[1 + (n_i - 1)\,\rho\right] \qquad (13.10)$$

with $|\rho| \leq 1$. Although motivated by the beta-binomial model, this variance function results merely from assuming that π_i has a distribution with $\text{var}(\pi_i) = \rho\mu_i(1-\mu_i)$. It also results from assuming a common correlation ρ between each pair of the n_i individual binary random variables that sum to y_i (Altham 1978). The ordinary binomial variance results when $\rho = 0$. Overdispersion occurs when $\rho > 0$.

For this quasi-likelihood approach, Williams (1982) gave an iterative routine for estimating $\boldsymbol{\beta}$ and the overdispersion parameter ρ. He let $\hat{\rho}$ be such that the resulting Pearson X^2 that sums the squared Pearson residuals for this variance function equals the residual df for the model. This requires an iterative two-step process of (1) solving the quasi-likelihood equations for $\boldsymbol{\beta}$ for a given $\hat{\rho}$, and then (2) using the updated $\hat{\boldsymbol{\beta}}$, solving for $\hat{\rho}$ in the equation that equates X^2 (which depends on $\hat{\boldsymbol{\beta}}$ and $\hat{\rho}$) to its df.

An alternative quasi-likelihood approach uses the simpler variance function

$$v(\mu_i) = \phi n_i \mu_i (1 - \mu_i) \tag{13.11}$$

introduced in Section 4.7.3. The ordinary binomial variance has $\phi = 1.0$ and overdispersion has $\phi > 1$. With this approach, $\hat{\boldsymbol{\beta}}$ is the same as its ML estimate for the ordinary binomial model. Commonly, $\hat{\phi} = X^2/\text{df}$, where X^2 is the Pearson fit statistic for the binomial model (Finney 1947). The standard errors for the overdispersion approach multiply those for the binomial model by $\hat{\phi}^{1/2}$.

Liang and McCullagh (1993) showed several examples using these two variance functions. A plot of the standardized residuals for the ordinary binomial model against the indices $\{n_i\}$ can provide insight about which is more appropriate. When the residuals show an increasing trend in their spread as n_i increases, the beta-binomial-type variance function may be more appropriate. This is because when the beta-binomial variance holds, the residuals from an ordinary binomial model have denominator that is progressively too small as n_i increases. The two quasi-likelihood approaches are equivalent when $\{n_i\}$ are identical. Only when the indices vary considerably might results differ much. Because the variance function $v(\mu_i) = \phi n_i \mu_i (1 - \mu_i)$ has a structural problem when $n_i = 1$ (Problem 13.33) and has less direct motivation, we prefer quasi-likelihood with the beta-binomial variance function.

13.3.3 Teratology Overdispersion Example Revisited

Refer back to Table 4.5 on results of a teratology experiment analyzed by Liang and McCullagh (1993) and Moore and Tsiatis (1991). Female rats on iron-deficient diets were assigned to four groups. Group 1 was given only placebo injections. The other groups were given injections of an iron supplement according to various schedules. The rats were made pregnant and then sacrificed after 3 weeks. For each fetus in each rat's litter, the response was whether the fetus was dead. Because of unmeasured covariates, it is natural to permit the probability of death to vary from litter to litter within a particular treatment group.

Let y_i denote the number dead out of the n_i fetuses in litter i. Let π_{it} denote the probability of death for fetus t in litter i. First, suppose that y_i is a $\text{bin}(n_i, \pi_{it})$ variate, with

$$\text{logit}(\pi_{it}) = \alpha + \beta_2 z_{2i} + \beta_3 z_{3i} + \beta_4 z_{4i},$$

where $z_{gi} = 1$ if litter i is in group g and 0 otherwise. This model treats all litters in a group g as having the same probability of death, $\exp(\alpha + \beta_g)/[1 + \exp(\alpha + \beta_g)]$, where $\beta_1 = 0$. However, it has evidence of overdispersion,

TABLE 13.5 Estimates for Several Logit Models Fitted to Table 4.5

	Type of Logit Model[a]				
Parameter	Binomial ML	QL(1)	QL(2)	GEE	GLMM
Intercept	1.144 (0.129)	1.212 (0.223)	1.144 (0.219)	1.144 (0.276)	1.802 (0.362)
Group 2	−3.322 (0.331)	−3.370 (0.563)	−3.322 (0.560)	−3.322 (0.440)	−4.515 (0.736)
Group 3	−4.476 (0.731)	−4.585 (1.303)	−4.476 (1.238)	−4.476 (0.610)	−5.855 (1.190)
Group 4	−4.130 (0.476)	−4.250 (0.848)	−4.130 (0.806)	−4.130 (0.576)	−5.594 (0.919)
Overdispersion	None	$\hat{\rho} = 0.192$	$\hat{\phi} = 2.86$	$\hat{\rho} = 0.185$	$\hat{\sigma} = 1.53$

[a] Binomial ML assumes no overdispersion, QL(1) is quasi-likelihood with beta-binomial-type variance, QL(2) is quasi-likelihood with inflated binomial variance; QL(2) and GEE (independence working equations) estimates are the same as binomial ML estimates. Values in parentheses are standard errors.

with $X^2 = 154.7$ and $G^2 = 173.5$ (df = 54). Table 13.5 shows ML estimates and standard errors.

Table 13.5 also shows results for the two quasi-likelihood approaches. Estimates and standard errors are qualitatively similar for each. For variance function $v(\mu_i) = \phi n_i \mu_i(1 - \mu_i)$, the estimates equal the binomial ML estimates but standard errors are multiplied by $\hat{\phi}^{1/2} = (X^2/\text{df})^{1/2} = \sqrt{154.7/54} = 1.69$. For the beta-binomial-type variance function, $\hat{\rho} = 0.192$. This fit treats the variance of Y_i as

$$n_i \mu_i(1 - \mu_i)\big[1 + 0.192(n_i - 1)\big].$$

This corresponds roughly to a doubling of the variance relative to the binomial with a litter size of 6 and a tripling with $n_i = 11$. Even with these adjustments for overdispersion, Table 13.5 shows that strong evidence remains that the probability of death is substantially lower for each treatment group than the placebo group.

Figure 13.4 plots the standardized Pearson residuals against litter size for the binomial logit model. The apparent increase in their variability as litter size increases suggests that the beta-binomial variance function is plausible. The term ρ in that variance function corresponds to $\theta/(1 + \theta)$ in the variance of the beta-binomial distribution. For that distribution or more generally, $\hat{\rho} = 0.192$ means that the probabilities of death for litters of a particular group have estimated standard deviation $\sqrt{0.192\mu_i(1 - \mu_i)}$. This equals 0.22 when the mean is 0.5 and 0.13 when the mean is 0.1 or 0.9, which is considerable heterogeneity. More generally, a model could let ρ vary by treatment group or be different for the placebo group than the others. We leave this to the reader.

For comparison, Table 13.5 also shows results with the GEE approach to fitting the logit model, assuming an independence working correlation structure for observations within a litter. The estimates are the same as the ML

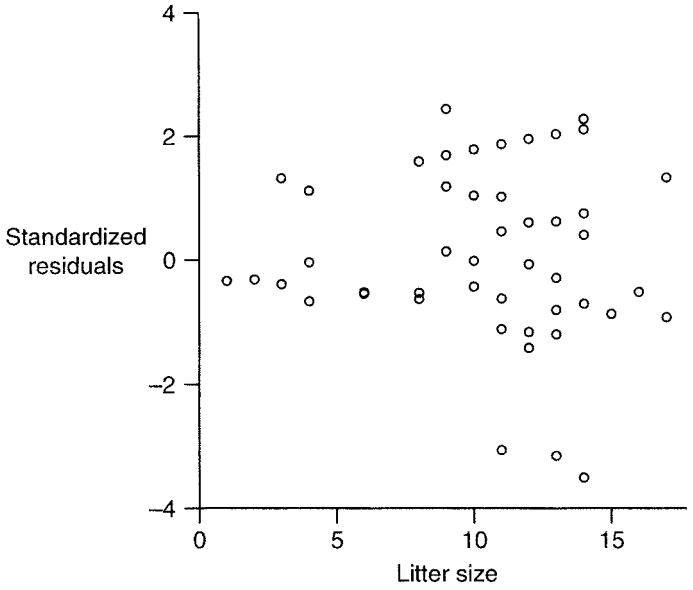


FIGURE 13.4 Standardized Pearson residuals for binomial logit model fitted to Table 4.5.

estimates for the binomial logit model, but the empirical adjustment increases the standard errors. Similar results occur with an exchangeable working correlation structure. For it, the estimated within-litter correlation between the binary responses is 0.185. This is comparable to the value of 0.192 that yields the quasi-likelihood results with beta-binomial variance function. The GEE standard errors are somewhat different from those with the quasi-likelihood approach. It may be that the sample size is insufficient for the GEE sandwich adjustment, which tends to underestimate standard errors unless the number of clusters is quite large. Or, this may simply reflect the different variance function for the GEE approach.

Finally, Table 13.5 also shows results for the GLMM that adds a normal random intercept u_i for litter i to the binomial logit model. Results are also similar in terms of significance of the treatment groups relative to placebo. Estimated effects are larger for this logistic-normal model, since they are subject-specific (i.e., litter-specific) rather than population-averaged.

13.3.4 Conjugate Mixture Models

The beta-binomial model is an example of a *conjugate mixture model*. These are models for which the marginal distribution has closed form. The data have a particular distribution, conditional on a parameter, and then the parameter has its own distribution such that the marginal distribution has closed form.

Similarly, in Bayesian methods the conjugate prior distribution is a distribution that when combined with the likelihood, gives a closed form for the posterior distribution. For instance, for observations from a binomial distribution with beta prior distribution for the binomial parameter, the posterior distribution of that parameter is also beta. Conjugate models were the primary method of conducting Bayesian analysis before the development of computationally intensive methods, such as Markov chain Monte Carlo, for evaluating the integral that determines the posterior distribution.

The beta-binomial conjugate mixture model applies with totals from binary trials. In the next section we study a conjugate mixture model for count data. It uses a gamma distribution to mix the Poisson parameter. A disadvantage of the conjugate mixture approach is the lack of generality and flexibility, requiring a different mixture distribution for each type of problem. In addition, the extra variability need not enter on the same scale as the ordinary predictors, and it can be difficult to have multivariate random effects structure. Lee and Nelder (1996) discussed this approach and considered a variety of hierarchical models of GLMM form in which the random effect need not be normal.

13.4 NEGATIVE BINOMIAL REGRESSION

The negative binomial is a conjugate mixture distribution for count data. It is useful when overdispersion occurs with Poisson GLMs.

13.4.1 Negative Binomial as Gamma Mixture of Poisson Distributions

In Section 4.3.3 we noted that a severe limitation of Poisson models is that the variance of Y must equal the mean. Hence, at a fixed mean the variance cannot decrease as additional predictors enter the model. Count data often show overdispersion, with the variance exceeding the mean. This might happen, for instance, because some relevant explanatory variables are not in the model. A mixture model is a flexible way to account for overdispersion. At a fixed setting of the predictors used, given the mean the distribution of Y is Poisson, but the mean itself varies according to some distribution.

Suppose that (1) given λ, Y has a Poisson distribution with mean λ, and (2) λ has a gamma distribution, $G(k, \mu)$. The gamma probability density function for λ is

$$f(\lambda; k, \mu) = \frac{(k/\mu)^k}{\Gamma(k)} \exp(-k\lambda/\mu)\lambda^{k-1}, \quad \lambda \geq 0. \tag{13.12}$$

This gamma distribution has

$$E(\lambda) = \mu, \quad \text{var}(\lambda) = \mu^2/k.$$

The parameter $k > 0$ describes the shape. The density is skewed to the right, but the degree of skewness decreases as k increases.

Marginally, the gamma mixture of the Poisson distributions yields the negative binomial distribution for Y. Its probability mass function is

$$p(y; k, \mu) = \frac{\Gamma(y + k)}{\Gamma(k)\Gamma(y + 1)} \left(\frac{k}{\mu + k}\right)^k \left(1 - \frac{k}{\mu + k}\right)^y, \qquad y = 0, 1, 2, \ldots . \tag{13.13}$$

This negative binomial distribution has

$$E(Y) = \mu, \qquad \text{var}(Y) = \mu + \mu^2/k .$$

The index k^{-1} is called the *dispersion parameter*. As $k^{-1} \to 0$, the gamma distribution has $\text{var}(\lambda) \to 0$ and it converges to a degenerate distribution at μ; similarly, the negative binomial distribution then has $\text{var}(Y) \to \mu$ and it converges to the Poisson distribution with mean μ.

For given k^{-1}, the negative binomial is in the natural exponential family. The natural parameter is $\log[\mu/(\mu + k)]$. Usually, though, the dispersion parameter k^{-1} is itself unknown. Estimating it helps to summarize the extent of overdispersion. The greater k^{-1}, the greater the overdispersion compared to the ordinary Poisson GLM. For independent observations, the ML estimate of μ is the sample mean, but ML estimation for k^{-1} requires iterative methods (R. A. Fisher showed this in an appendix of a 1953 *Biometrics* article by C. Bliss). Problem 13.40 shows an alternative gamma parameterization that implies a linear rather than quadratic variance function for the negative binomial.

13.4.2 Negative Binomial Regression Modeling

Negative binomial models for counts permit μ to depend on explanatory variables (Lawless 1987). Such models normally take k^{-1} to be the same for all observations. This corresponds to a constant coefficient of variation in the gamma mixing distribution, $\sqrt{\text{var}(\lambda)}\,/\text{E}(\lambda) = 1/\sqrt{k}$. with the standard deviation increasing as the mean does. Most common is the log link, as in Poisson loglinear models. Sometimes the identity link is adequate. One such case is with a single predictor that is a factor.

For k fixed, a negative binomial model is a GLM. Thus, the likelihood equations for the regression parameters $\boldsymbol{\beta}$ are special cases of those [see (4.22)] for an ordinary GLM with variance function $v(\mu) = \mu + \mu^2/k$. The usual iterative reweighted least squares algorithm applies for ML model fitting. When k is unknown, ML fitting can use a Newton–Raphson routine on all the parameters simultaneously. Or, one can evaluate the profile likelihood for various fixed k (Lawless 1987). Another approach alternates

between (1) using iterative reweighted least squares to solve the equations for $\boldsymbol{\beta}$, for fixed k, and (2) for fixed $\hat{\boldsymbol{\beta}}$, using Newton–Raphson to estimate k, iterating between them until convergence.

The full log likelihood $L(\boldsymbol{\beta}, k; \mathbf{y})$ for a negative binomial model satisfies

$$\frac{\partial^2 L}{\partial \beta_j \, \partial k} = \sum_i \frac{y_i - \mu_i}{(k + \mu_i)^2 \, g'(\mu_i)} x_{it} .$$

Thus, $E(\partial^2 L / \partial \beta_j \partial k) = 0$ for each j. Similarly, the inverse of the expected information matrix has 0 elements connecting k with each β_j. Since this is the asymptotic covariance matrix, $\hat{\boldsymbol{\beta}}$ and $\hat{\mathrm{k}}$ are asymptotically independent. It follows that standard errors for $\hat{\boldsymbol{\beta}}$ obtained from part (1) of the iterative scheme above are correct. Cameron and Trivedi (1998, p. 72) showed the asymptotic covariance matrix. They [and Lawless (1987)] considered a moment estimator for k^{-1} and studied robustness properties of estimators. They noted that $\hat{\boldsymbol{\beta}}$ from this model is consistent if the model for the mean is correctly specified, even if the true distribution is not negative binomial.

13.4.3 Frequency of Knowing Homicide Victims Example

Table 13.6 summarizes responses of 1308 subjects to the question: Within the past 12 months, how many people have you known personally that were victims of homicide? The table shows responses by race, for those who identified their race as white or as black. The sample mean for the 159 blacks was 0.522, with a variance of 1.150. The sample mean for the 1149 whites was 0.092, with a variance of 0.155.

A natural first choice for modeling count data is a Poisson GLM, such as a loglinear model with a dummy predictor for race. Let y_{it} denote the response for subject t of race i. For $\mu_{it} = E(Y_{it})$, this model is

$$\log \mu_{it} = \alpha + \beta x_{it} ,$$

TABLE 13.6 Number of Victims of Murder Known in Past Year, by Race, with Fit of Poisson and Negative Binomial Models

	Data		Poisson GLM		Neg. Bin. GLM		Poisson GLMM	
Response	Black	White	Black	White	Black	White	Black	White
0	119	1070	94.3	1047.7	122.8	1064.9	116.7	1068.3
1	16	60	49.2	96.7	17.9	67.5	24.5	65.3
2	12	14	12.9	4.5	7.8	12.7	8.1	10.1
3	7	4	2.2	0.1	4.1	2.9	3.6	2.8
4	3	0	0.3	0.0	2.4	0.7	1.9	1.1
5	2	0	0.0	0.0	1.4	0.2	1.1	0.5
6	0	1	0.0	0.0	0.9	0.1	0.7	0.3

Source: 1990 General Social Survey, National Opinion Research Center.

with $x_{1t} = 1$ (blacks) and $x_{2t} = 0$ (whites). This model has fit $\log \hat{\mu}_{it} = -2.38 + 1.733x_{it}$. The estimated expected responses are $\exp(-2.38 + 1.733) = 0.522$ for blacks and $\exp(-2.38) = 0.092$ for whites, the sample means. For any link function for this model, the likelihood equations imply that the fitted means equal the sample means. Since $\hat{\beta} = 1.733$ (SE = 0.147) is the difference between the log means for blacks and whites, the ratio of sample means is $\exp(1.733) = 5.7 = 0.522/0.092$. However, for each race the sample variance is roughly double the mean. Table 13.6 also shows the fit of this model. The evidence of overdispersion is reflected by the higher observed counts at $y = 0$ and at large y values than the Poisson GLM predicts.

An alternative is the same model form but assuming a negative binomial response. A mixture model does seem plausible. Due to various demographic factors, heterogeneity probably occurs among subjects of a given race in the distribution of Y. For ML fitting, the deviance decreases by 122.2 compared to the ordinary Poisson GLM that is the special case with $k^{-1} = 0$. Table 13.6 also shows this model fit. It is dramatically better at $y = 0$ and 1.

Table 13.7 shows parameter estimates for the negative binomial and Poisson GLMs. For both, $\hat{\beta} = 1.733$ since both models provide fitted means equal to the sample means. However the estimated standard error of $\hat{\beta}$ increases from 0.147 for the Poisson GLM to 0.238 for the negative binomial model. The Wald 95% confidence interval for the ratio of means for blacks and whites goes from $\exp[1.733 \pm 1.96(0.147)] = (4.2, 7.5)$ for the Poisson GLM to $\exp[1.733 \pm 1.96(0.238)] = (3.5, 9.0)$ for the negative binomial. In accounting for the overdispersion, we obtain results that are not as precise as the more naive model suggests.

The negative binomial model has $\hat{k}^{-1} = 4.94$ (SE = 1.00). This shows strong evidence that $k^{-1} > 0$, indicating that the negative binomial model is more appropriate than the Poisson GLM. The estimated variance of Y is $\hat{\mu} + \hat{\mu}^2/\hat{k} = \hat{\mu} + 4.94\hat{\mu}^2$, which is 0.13 for whites and 1.87 for blacks, much closer to the sample values than the Poisson model provides.

Table 13.7 also shows results for negative binomial and Poisson models using the identity link. The fits $\hat{\mu}_{it} = 0.092 + 0.430x_{it}$ reproduce the sample means. Now $\hat{\beta}$ refers to the difference in means rather than their log ratio. The estimated difference $\hat{\beta} = 0.430$ has SE = 0.058 for the Poisson model and SE = 0.109 for the negative binomial. Results are more imprecise but

TABLE 13.7 Parameter Estimates for Models Fitted to Homicide Data

	Models with Log Link			Models with Identity Link	
Term	Neg. Binom. GLM	Poisson GLM	Poisson GLMM	Neg. Binom. GLM	Poisson GLM
α	−2.38	−2.38	−3.69	0.092	0.092
β	1.733	1.733	1.897	0.430	0.430
SE($\hat{\beta}$)	0.238	0.147	0.246	0.109	0.058

more realistic with the negative binomial model. For this link also the estimated dispersion parameter is $\hat{k}^{-1} = 4.94$.

13.5 POISSON REGRESSION WITH RANDOM EFFECTS

The GLMMs introduced in Chapter 12 referred to categorical responses. GLMMs are also useful for other types of discrete responses, such as counts. This section illustrates with Poisson regression modeling of count data.

We've seen that a flexible way to account for overdispersion is with a mixture model. In Section 13.4 we mixed the Poisson using the gamma distribution, yielding the negative binomial marginally. Breslow (1984) and Hinde (1982) suggested the GLMM structure (12.1) with the log link and normal random intercept. The model for the mean for observation t in cluster i is

$$\log\big[E(Y_{it}\,|\,u_i)\big] = \mathbf{x}_{it}'\boldsymbol{\beta} + u_i, \tag{13.14}$$

where $\{u_i\}$ are independent $N(0, \sigma^2)$. Conditional on u_i, y_{it} has a Poisson distribution. Marginally, the distribution has variance greater than the mean whenever $\sigma > 0$.

Applications of Poisson GLMMs include the analysis of maps of cancer rates in epidemiology (Breslow and Clayton 1993) and modeling variability in bacteria counts (Aitchison and Ho 1989). Although links other than the log are possible, the identity link (and any other link having range only the positive real line) has a structural problem. With a normal random effect with $\sigma > 0$, a positive probability exists that the linear predictor is negative, but the Poisson mean must be nonnegative.

The negative binomial model (for fixed k) is a GLMM with nonnormal random effect. With the log link, it results from a loglinear model of form (13.14) with random intercept, where $\exp(u_i)$ has a gamma distribution with mean 1 and variance k^{-1}. With identity link, negative binomial models usually work better than Poisson GLMMs. Regardless of the gamma mixture distribution, the resulting marginal mean is nonnegative for the negative binomial.

13.5.1 Marginal Model Implied by Poisson GLMM

The Poisson GLMM (13.14) implies a relatively simple marginal model, averaging out the random effect. The mean of the marginal distribution is

$$E(Y_{it}) = E\big[E(Y_{it}\,|\,u_i)\big] = E\big[e^{\mathbf{x}_{it}'\boldsymbol{\beta}+u_i}\big] = e^{\mathbf{x}_{it}'\boldsymbol{\beta}+\sigma^2/2}.$$

Here $E[\exp(u_i)] = \exp(\sigma^2/2)$ because a $N(0, \sigma^2)$ variate u_i has moment generating function $E[\exp(tu_i)] = \exp(t^2\sigma^2/2)$. So, for the Poisson GLMM

the log of the mean conditionally equals $\mathbf{x}_{it}'\boldsymbol{\beta} + u_i$ and marginally equals $\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma^2/2$. A loglinear model still applies. The marginal effects of the explanatory variables are the same as the cluster-specific effects. Thus, the *ratio* of means at two different settings of $\mathbf{x}_{it}'$ is the same conditionally and marginally. However, marginally the intercept is offset. (Note that Jensen's inequality applies, since the link is not linear.)

The variance of the marginal distribution is

$$\begin{aligned}\operatorname{var}(Y_{it}) &= E\big[\operatorname{var}(Y_{it}\,|\,u_i)\big] + \operatorname{var}\big[E(Y_{it}\,|\,u_i)\big] = E\big[e^{\mathbf{x}_{it}'\boldsymbol{\beta}+u_i}\big] + e^{2\mathbf{x}_{it}'\boldsymbol{\beta}}\operatorname{var}(e^{u_i}) \\ &= e^{\mathbf{x}_{it}'\boldsymbol{\beta}+\sigma^2/2} + e^{2\mathbf{x}_{it}'\boldsymbol{\beta}}\big(e^{2\sigma^2} - e^{\sigma^2}\big) = E(Y_{it}) + \big[E(Y_{it})\big]^2\big(e^{\sigma^2} - 1\big).\end{aligned}$$

Here, $\operatorname{var}(e^{u_i}) = E(e^{2u_i}) - [E(e^{u_i})]^2 = e^{2\sigma^2} - e^{\sigma^2}$ by evaluating the moment generating function at $t = 2$ and $t = 1$. As in the negative binomial model, the marginal variance is a quadratic function of the marginal mean. It exceeds the marginal mean when $\sigma > 0$. The ordinary Poisson model results when $\sigma = 0$. When $\sigma > 0$ the marginal distribution is not Poisson, and the extent to which the variance exceeds the mean increases as σ increases.

As in binary GLMMs, Y_{it} and Y_{is} are independent given u_i but are marginally nonnegatively correlated. For $t \neq s$,

$$\begin{aligned}\operatorname{cov}(Y_{it}, Y_{is}) &= E\big[\operatorname{cov}(Y_{it}, Y_{is}\,|\,u_i)\big] + \operatorname{cov}\big[E(Y_{it}\,|\,u_i), E(Y_{is}\,|\,u_i)\big] \\ &= 0 + \operatorname{cov}\big[\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i), \exp(\mathbf{x}_{is}'\boldsymbol{\beta} + u_i)\big]. \qquad (13.15)\end{aligned}$$

The functions in the last covariance term are both monotone increasing functions of u_i, and hence are nonnegatively correlated (Problem 13.44).

13.5.2 Frequency of Knowing Homicide Victims Example

We now return to Table 13.6 on responses, classified by race, of the number of victims of homicide within the past 12 months that subjects knew personally. Models permitting subject heterogeneity are sensible. For the response y_{it} for subject t of race i, the Poisson GLMM is

$$\log\big[E(Y_{it}\,|\,u_{it})\big] = \alpha + \beta x_{it} + u_{it},$$

where $\{u_{it}\}$ are independent $N(0, \sigma^2)$. The log means vary according to a $N(\alpha, \sigma^2)$ distribution for whites and a $N(\alpha + \beta, \sigma^2)$ distribution for blacks. Given u_{it}, y_{it} has a Poisson distribution.

Table 13.6 also shows this model fit, and Table 13.7 shows estimates. The random effects have $\hat{\sigma} = 1.63$ (SE = 0.15). The deviance decreases by 116.6 compared to the Poisson GLM, indicating a better fit by allowing heterogeneity. For subjects at the means of the random effects distributions ($u_{it} = 0$) the estimated expected responses are $\exp(-3.69 + 1.90) = 0.167$ for blacks and

$\exp(-3.69) = 0.025$ for whites. The fitted marginal mean is $\exp(\hat{\alpha} + \hat{\beta}x_{it} + \hat{\sigma}^2/2)$, or 0.63 for blacks and 0.09 for whites. The fitted marginal variances are 0.21 for blacks and 5.78 for whites. These are somewhat larger than the sample means and variances, perhaps because the fitted distribution has nonnegligible mass above the largest observed response of 6.

13.5.3 Negative Binomial Models versus Poisson GLMMs

The Poisson GLMM with normal random effects has the advantage, relative to the negative binomial GLM, of easily permitting multivariate random effects and multilevel models. However, the negative binomial has properties that can make interpretation simpler. We've seen that the identity link is valid for it, which is useful for simple examples such as the preceding one with a factor predictor. With any link and a factor predictor, its ML fitted means equal the sample means. This is not the case for the Poisson GLMM.

Besides the Poisson GLMM and the negative binomial model, an alternative way of accounting for overdispersion with count data is quasi-likelihood with variance function

$$v(\mu_i) = \phi\mu_i,$$

for some constant ϕ. This is often adequate for exploratory analyses.

NOTES

Section 13.1: Latent Class Models

13.1. Aitkin et al. (1981), Bartholomew and Knott (1999), Clogg (1995), Clogg and Goodman (1984), Goodman (1974), Haberman (1979, Chap. 10), Hagenaars (1998), Heinen (1996), and Lazarsfeld and Henry (1968) discussed fitting and intrepretation of latent class and related latent variable models.

13.2. Rudas et al. (1994) proposed a clever mixture method for summarizing goodness of fit. For a model M for a contingency table with true probabilities $\boldsymbol{\pi}$, they used the mixture $\boldsymbol{\pi} = (1 - \rho)\boldsymbol{\pi}_1 + \rho\boldsymbol{\pi}_2$, with $\boldsymbol{\pi}_1$ the model-based probabilities and $\boldsymbol{\pi}_2$ unconstrained. Their index of lack of fit is the smallest such ρ possible for which this holds. It is the fraction of the population that cannot be described by the model. This recognizes that any given model does not truly hold but is useful if ρ is close to 0. The mixture contrasts with the latent class model in which both $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ correspond to independence.

Section 13.2: Nonparametric Random Effects Models

13.3. For connections between Rasch-type models and quasi-symmetry models, see Agresti (1993), Conaway (1989), Darroch (1981), Darroch et al. (1993), Hatzinger (1989), and Kelderman (1984). For the matched-pairs random effects model (12.16), a nonparametric or conditional ML treatment of (u_{i1}, u_{i2}) implies a multivariate quasi-symmetry model (Agresti 1997). Model (12.16) with correlated normal random effects is a

continuous analog to discrete latent class models that Goodman (1974) proposed, based on two associated binary latent variables.

Section 13.3: Beta-Binomial Models

13.4. Skellam (1948) introduced the beta-binomial distribution and discussed parameter estimation. For modeling using this distribution or related quasi-likelihood approaches, see Brooks et al. (1997), Crowder (1978), Hinde (1996), Lee and Nelder (1996), Liang and Hanfelt (1994), Liang and McCullagh (1993), Lindsey and Altham (1998), Moore (1986a), Moore and Tsiatis (1991), Nelder and Pregibon (1987), Prentice (1986), Rosner (1984, 1989) [with critique by Neuhaus and Jewell (1990a)], Slaton et al. (2000), and Williams (1975, 1982). For beta-binomial type variance, Ryan (1995) and Williams (1988) showed advantages of the quasi-likelihood approach over ML. Often, it helps to permit the quasi-likelihood scale parameter ρ (or the related parameter θ in the beta-binomial) to vary among groups.

The beta-binomial generalizes to a Dirichlet-multinomial. Conditional on the probabilities, the distribution is multinomial. The probabilities themselves have a Dirichlet distribution, which is a generalization of the beta defined on vectors of probabilities that sum to 1. See Mosimann (1962) and Paul et al. (1989).

13.5. Kupper et al. (1986) and Ryan (1992) discussed modeling overdispersion caused by litter effects in developmental toxicity studies. See Follman and Lambert (1989), Kupper and Haseman (1978), and Lefkopoulou et al. (1989) for related material.

Section 13.4: Negative Binomial Regression

13.6. Greenwood and Yule (1920) derived the negative binomial as a gamma mixture of Poissons. Johnson et al. (1992) summarized its properties. Biggeri (1998), Cameron and Trivedi (1998), Hinde and Demétrio (1998), and Lawless (1987) discussed modeling using it.

PROBLEMS

Applications

13.1 For the 2^3 table of opinions about legalized abortion (Table 10.13) collapsed over gender, fit a latent class model with two classes. Show that it is saturated. For each latent class, report the estimated probability of supporting legalized abortion in each of the three situations. Give a tentative interpretation for the classes.

13.2 Analyze Table 8.3 using a latent class model with $q = 2$.

a. For a subject in the first latent class, estimate the probability of having used **(i)** marijuana, **(ii)** alcohol, **(iii)** cigarettes, **(iv)** all three, and **(v)** none of them.

b. Estimate the probability a subject is in the first latent class, given they have used **(i)** marijuana, **(ii)** alcohol, **(iii)** cigarettes, **(iv)** all three, and **(v)** none of them.

13.3 Analyze Table 8.19 on government spending using latent class models.

13.4 For capture–recapture experiments, Coull and Agresti (1999) used a loglinear model with exchangeable association and no higher-order terms. Explain why the model expected frequencies satisfy

$$\log \mu(y_1, \ldots, y_T) = \lambda + \beta_1 y_1 + \cdots + \beta_T y_T + \beta(y_1 y_2 + y_1 y_3 + \cdots + y_{T-1} y_T) .$$

Show that the fit of this model to Table 12.6 yields $\hat{N} = 90.5$ and a 95% profile-likelihood confidence interval for N of (75, 125).

13.5 Use or write software to replicate the analyses of the opinions about abortion data in Section 13.2 using **(a)** nonparametric random effects fitting of logit model (13.3), and **(b)** the quasi-symmetry model.

13.6 A data set on pregnancy rates among girls under 18 years of age in 13 north central Florida counties has information on a 3-year total for each county i on n_i = number of births and y_i = number of those for which mother had age under 18 (see J. Booth, in *Statistical Modelling: Lecture Notes in Statistics*, *104*, Springer, 43–52, 1995).

a. A beta-binomial model states that given $\{\pi_i\}$, $\{Y_i\}$ are independent $\{\text{bin}(n_i, \pi_i)\}$ variates, and $\{\pi_i\}$ are independent from a beta(α, β) distribution. The ML estimated parameters are $\hat{\alpha} = 9.9$ and $\hat{\beta} = 240.8$ (thanks to J. Booth for this analysis). Use the mean and variance to describe the estimated beta distribution and the estimated marginal distribution of Y_i (as a function of n_i).

b. Quasi-likelihood using variance function (13.10) for the model $\text{logit}(\mu_i) = \alpha$ has $\hat{\alpha} = -3.18$ and $\hat{\rho} = 0.005$. Describe the estimated mean and variance of Y_i.

c. Quasi-likelihood using variance (13.11) for the model $\text{logit}(\mu_i) = \alpha$ has $\hat{\alpha} = -3.35$ and $\hat{\phi} = 8.3$. Describe the estimated mean and variance of Y_i.

d. The logistic-normal GLMM, $\text{logit}(\pi_i) = \alpha + u_i$, yields $\hat{\alpha} = -3.24$ and $\hat{\sigma} = 0.33$. Describe the estimated mean of Y_i [Recall (12.8)].

13.7 In Problem 12.2 about Shaq O′Neal's free-throw shooting, the simple binomial model, $\pi_i = \alpha$, has lack of fit. Fit the beta-binomial model, or use the quasi-likelihood approach with that variance structure. Use the fit to summarize his free-throw shooting, by giving an estimated mean and standard deviation for π_i.

13.8 For the toxicity study of Table 12.9, collapsing to a binary response, consider linear logit models for the probability a fetus is normal.

a. Does the ordinary binomial model show evidence of overdispersion?

b. Fit the linear logit model using the quasi-likelihood approach with inflated binomial variance. How do the standard errors change?

c. Fit the linear logit model using quasi-likelihood with beta-binomial variance. Interpret and compare with previous results.

d. Fit the linear logit model using a GEE approach with exchangeable working correlation among fetuses in the same litter. Interpret and compare with previous results, including comparing the estimated GEE correlation with the estimate $\hat{\rho}$ from part (c).

e. Fit the linear logit GLMM after adding a litter-specific normal random effect. Interpret and compare with previous results.

13.9 Extend the various analyses of the teratology data (Table 4.5) in Section 13.3.3 as follows:

a. Include a predictor for litter size (as well as group). Interpret, and compare results to those without this predictor.

b. Fit a model with beta-binomial variance (13.10) in which ρ varies by treatment group. Use results to motivate a model that allows overdispersion only in the placebo group. Interpret and compare results to those with common ρ for each group.

13.10 Table 13.8 reports the results of a study of fish hatching under three environments. Eggs from seven clutches were randomly assigned to three treatments, and the response was whether an egg hatched by day 10. The three treatments were (1) carbon dioxide and oxygen removed, (2) carbon dioxide only removed, and (3) neither removed.

TABLE 13.8 Data for Problem 13.10

	Treatment 1		Treatment 2		Treatment 3	
Clutch	Number Hatched	Total	Number Hatched	Total	Number Hatched	Total
1	0	6	3	6	0	6
2	0	13	0	13	0	13
3	0	10	8	10	6	9
4	0	16	10	16	9	16
5	0	32	25	28	23	30
6	0	7	7	7	5	7
7	0	21	10	20	4	20

Source: Data courtesy of Becca Hale, Zoology Department, University of Florida.

a. Let π_{it} denote the probability of hatching for an egg from clutch i in treatment t. Assuming independent binomial observations, fit the model

$$\text{logit}(\pi_{it}) = \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3,$$

where $z_t = 1$ for treatment t and 0 otherwise. What does your software report for $\hat{\beta}_1$, and what should it be? (*Hint:* Note that treatment 1 has no successes.)

b. Analyze these data using an approach that allows overdispersion. Interpret. Indicate whether evidence of overdispersion occurs for treatments 2 and 3.

13.11 For the train accidents in Problem 9.19, a negative binomial model assuming constant log rate over the 14-year period has estimate -4.177 (SE = 0.153) and estimated dispersion parameter 0.012. Interpret.

13.12 One question in the 1990 General Social Survey asked subjects how many times they had sexual intercourse in the preceding month. Table 13.9 shows responses, classified by gender.

a. The sample means were 5.9 for males and 4.3 for females; the sample variances were 54.8 and 34.4. The mode for each gender was 0. Does an ordinary Poisson GLM seem appropriate? Explain.

b. The Poisson GLM with log link and a dummy variable for gender (1 = males, 0 = females) has gender estimate 0.308 (SE = 0.038). Explain why this implies a ratio of 1.36 for the fitted means. (This is also the ratio of sample means, since this model has fitted means equal to sample means.) Show that the Wald 95% confidence interval for the ratio of means for males and females is (1.26, 1.47).

TABLE 13.9 Data for Problem 13.12

Response	Male	Female	Response	Male	Female	Response	Male	Female
0	65	128	9	2	2	20	7	6
1	11	17	10	24	13	22	0	1
2	13	23	12	6	10	23	0	1
3	14	16	13	3	3	24	1	0
4	26	19	14	0	1	25	1	3
5	13	17	15	3	10	27	0	1
6	15	17	16	3	1	30	3	1
7	7	3	17	0	1	50	1	0
8	21	15	18	0	1	60	1	0

Source: 1990 General Social Survey, National Opinion Research Center.

c. For the negative binomial model, the log likelihood increases by 248.7 (deviance decreases by 497.3). The estimated difference between the log means is also 0.308, but now SE = 0.127. Show that the 95% confidence interval for the ratio of means is (1.06, 1.75). Compare to the Poisson GLM, and interpret.

d. The mode for the Poisson distribution is the integer part of the mean, rather than 0. Argue that a possibly more realistic mixture model assumes for gender i a proportion ρ_i that has a Poisson distribution with mean 0 and a proportion $1 - \rho_i$ that has distribution that is a gamma mixture of Poissons. Explain why the corresponding marginal distribution for each gender is a mixture of a degenerate distribution at 0 and a negative binomial distribution.

13.13 Refer to Problem 13.12. Fit the Poisson and negative binomial GLMs using identity link. Show that the estimated differences in means between males and females are identical for the two GLMs but the SE values are very different. Explain why. Use the more appropriate one to form a confidence interval for the true difference in means.

13.14 For the counts of horseshoe-crab satellites in Table 4.3, Table 13.10 shows the results of ML fitting of the negative binomial model using width as the predictor, with the identity link.

a. State and interpret the prediction equation.

b. Show that at a predicted $\hat{\mu}$, the estimated variance is roughly $\hat{\mu} + \hat{\mu}^2$.

c. The corresponding Poisson GLM has fit $\hat{\mu} = -11.53 + 0.55x$ (SE = 0.06). Compare 95% confidence intervals for the slopes for the two models. Interpret, and indicate whether overdispersion seems to exist relative to the Poisson GLM.

TABLE 13.10 Results for Problem 13.14

Parameter	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square
Intercept	−11.1471	2.8275	−16.6890	−5.6052	15.54
width	0.5308	0.1132	0.3089	0.7528	21.97
Dispersion	0.9843	0.1822	0.6847	1.4149	

13.15 Refer to Problem 13.14.

a. Fit a negative binomial model with log link. Interpret. Plot the counts against width and indicate which link seems more appropriate.

b. Fit a Poisson GLMM with log link, using width predictor. Interpret.

c. Compare results for the various models, including those in Section 4.3.2 for a Poisson GLM. Indicate your preferred model. Justify.

13.16 Refer to Problems 13.14 and 13.15. Using width and qualitative color as predictors, fit a (**a**) negative binomial GLM, and (**b**) Poisson GLMM, checking for interaction and interpreting the final model.

13.17 Refer to Table 13.6. For those with race classified as "other," the sample counts for $(0, 1, 2, 3, 4, 5, 6)$ homicides were $(55, 5, 1, 0, 1, 0, 0)$. Fit an appropriate model simultaneously to these data and those for white and black race categories. Interpret by making pairwise comparisons of the three pairs of means.

13.18 Use a quasi-likelihood approach to analyze Table 13.6 on counts of murder victims.

13.19 Conduct the analyses of Problem 4.6 on defects in the fabrication of computer chips, but use a negative binomial GLM. Compare results to those for the Poisson GLM. Indicate why results are similar.

13.20 With data at the book's Web site (*www.stat.ufl.edu/~aa/cda/cda.html*), use methods of this chapter to analyze how the countywide vote for the Reform Party candidate Pat Buchanan in the 2000 presidential election related to the vote for Reform Party candidate Ross Perot in the 1996 presidential election. Note that Palm Beach County is an enormous outlier (apparently mainly reflecting votes intended for Al Gore but cast for Buchanan because of a confusing ballot). Model with and without that observation and compare results.

13.21 Conduct a latent class analysis of the data in Espeland and Handelman (1989).

13.22 Refer to the teratology study in Liang and Hanfelt (1994). Analyze these data using at least two different approaches for overdispersed binary data. Compare results and interpret.

13.23 Refer to Problem 13.14. Using an appropriate subset of width, weight, color, and spine condition as predictors, find and interpret a reasonable model for predicting the number of satellites.

Theory and Methods

13.24 Derive residual df for a latent class model with q latent classes. When $I = 2$, for $q \geq 2$ show one needs $T \geq 4$ for the model to be unsaturated. Then, find the maximum value for q when $T = 4, 5$. For an I^2 table, show one needs $q < I^2/(2I - 1)$.

13.25 Express the log likelihood for latent class model (13.1) in terms of the model parameters. Derive likelihood equations (Goodman 1974, Haberman 1979).

13.26 Let $\mathbf{\Pi}$ denote an $I \times J$ matrix of cell probabilities for the joint distribution of X and Y. Suppose that there exist $I \times 1$ column vectors $\boldsymbol{\pi}_{1k}$ and $J \times 1$ column vectors $\boldsymbol{\pi}_{2k}$ of probabilities, $k = 1, \ldots, q$, and a set of probabilities $\{\rho_k\}$ such that

$$\mathbf{\Pi} = \sum_{k=1}^{q} \rho_k \boldsymbol{\pi}_{1k} \boldsymbol{\pi}'_{2k}.$$

Explain why this implies that there is a latent variable Z such that X and Y are conditionally independent, given Z.

13.27 In Section 13.2.2, under the null that the ordinary logistic regression model holds, explain why it is inappropriate to treat the difference between the deviances for that model and the mixture of two logistic regressions as a chi-squared statistic.

13.28 Refer to Problem 12.7. Let $\mu_k(a, b, c)$ denote the expected frequency of outcomes (a, b, c) for treatments (A, B, C) under treatment sequence k, where outcome 1 = relief and 0 = nonrelief. With a nonparametric random effects approach, show that one can estimate treatment effects in model (12.19) by fitting the quasi-symmetry model

$$\log \mu_k(a, b, c) = a\beta_A + b\beta_B + c\beta_C + \lambda_k(a, b, c),$$

where $\lambda_k(a, b, c) = \lambda_k(a, c, b) = \lambda_k(b, a, c) = \lambda_k(b, c, a) = \lambda_k(c, a, b) = \lambda_k(c, b, a)$. Fit the model, and show that $\hat{\beta}_B - \hat{\beta}_A = 1.64$ (SE = 0.34), $\hat{\beta}_C - \hat{\beta}_A = 2.23$ (SE = 0.39), $\hat{\beta}_C - \hat{\beta}_B = 0.59$ (SE = 0.39). Interpret. Compare results with Problem 12.7 for model (12.19).

13.29 Show that the beta-binomial distribution (13.9) simplifies to the binomial when $\theta = 0$.

13.30 Express the numerator of the beta density in terms of μ and θ. Using this, show that it is **(a)** unimodal when $\theta < \min(\mu, 1 - \mu)$, and **(b)** the uniform density when $\mu = \theta = \frac{1}{2}$.

13.31 Suppose that $\pi_i = P(Y_{it} = 1) = 1 - P(Y_{it} = 0)$, for $t = 1, \ldots, n_i$, and $\text{corr}(Y_{it}, Y_{is}) = \rho$ for $t \neq s$. Show that $\text{var}(Y_{it}) = \pi_i(1 - \pi_i)$,

$\text{cov}(Y_{it}, Y_{is}) = \rho\pi_i(1 - \pi_i)$, and

$$\text{var}\left(\sum_t Y_{it}\right) = n_i\pi_i(1 - \pi_i)\left[1 + \rho(n_i - 1)\right].$$

13.32 When $n = 1$, show that the beta-binomial distribution is no different from the binomial (i.e., Bernoulli). Explain why overdispersion cannot occur when $n = 1$.

13.33 When y_i is the sum of n_i binary responses each having mean μ_i, refer to the quasi-likelihood approach with $v(\mu_i) = \phi n_i \mu_i(1 - \mu_i)$. Explain why this variance function has a structural problem, with only $\phi = 1$ making sense when $n_i = 1$.

13.34 Liang and Hanfelt (1994) described a teratology study comparing control and treatment groups in which the ML estimate of the treatment effect in a beta-binomial model differs by a factor of 2 depending on whether one assumes the same overdispersion parameter for each group. By contrast, with variance function (13.11), the quasi-likelihood estimate of the treatment effect is the same whether one assumes the same or different ϕ for the two groups. Explain why, and discuss whether this is an advantage or disadvantage of that method.

13.35 Consider the logistic-normal model, $\text{logit}(\pi_i) = \alpha + \mathbf{x}_i'\boldsymbol{\beta} + u_i$. For small σ, show that it corresponds approximately to a mixture model for which the mixture distribution has $\text{var}(\pi_i) = [\mu_i(1 - \mu_i)]^2\sigma^2$. (*Hint:* See Problem 6.33.)

13.36 Altham (1978) introduced the discrete distribution

$$f(y; \pi, \psi) = c(\pi, \psi)\binom{n}{y}\pi^y(1 - \pi)^{n-y}\exp[\psi y(n - y)],$$
$$y = 0,1,\ldots,n,$$

where $c(\pi, \psi)$ is a normalizing constant. Show that this is in the exponential family. Show that the binomial occurs when $\psi = 0$. [Altham noted that overdispersion occurs when $\psi < 0$. Corcoran et al. (2001) and Lindsey and Altham (1998) used this as the basis of an alternative model to the beta-binomial.]

13.37 When $y_1, \ldots, y_N$ are independent from the negative binomial distribution (13.13) with k fixed, show that $\hat{\mu} = \bar{y}$.

13.38 Using $E(Y) = E[E(Y|X)]$ and $\text{var}(Y) = E[\text{var}(Y|X)] + \text{var}[E(Y|X)]$, derive the mean and variance of the (**a**) beta-binomial distribution,
and (**b**) negative binomial distribution.

13.39 Suppose that given u, Y is Poisson with $E(Y|u) = u\mu$, where μ may depend on predictors. Suppose that u is a positive random variable with $E(u) = 1$ and $\text{var}(u) = \tau$. Show that $E(Y) = \mu$ and $\text{var}(Y) = \mu + \tau\mu^2$. Explain how negative binomial GLMs and Poisson GLMMs with log link can follow as special cases.

13.40 An alternative negative binomial parameterization results from the gamma density formula,

$$f(\lambda; k, \mu) = \frac{(k)^{k\mu}}{\Gamma(k\mu)} \exp(-k\lambda)\lambda^{k\mu-1}, \qquad \lambda \geq 0,$$

for which $\text{E}(\lambda) = \mu$, $\text{var}(\lambda) = \mu/k$. Show that this gamma mixture of Poissons yields a negative binomial with

$$E(Y) = \mu, \quad \text{var}(Y) = \mu(1 + k)/k\,.$$

For what limiting value of k does this reduce to the Poisson? [See Nelder and Lee (1996) for ML model fitting. Cameron and Trivedi (1998, p. 75) pointed out that, unlike with quadratic variance, consistency does not occur for parameter estimators when the model for the mean holds but the true distribution is not negative binomial.]

13.41 The negative binomial distribution is unimodal with a mode at the integer part of $\mu(k - 1)/k$ (Johnson et al. 1992, pp. 208–209). Show that the mode is 0 when $\mu \leq 1$, and that when $\mu > 1$ the mode is still 0 if $k < \mu/(\mu - 1)$. (This gives greater scope than the Poisson, since its mode equals the integer part of the mean.)

13.42 Consider the loglinear random effects model

$$\log\big[E(Y_{it}|\mathbf{u}_i)\big] = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{u}_i,$$

where $\{\mathbf{u}_i\}$ are independent $N(\mathbf{0}, \boldsymbol{\Sigma})$. Show that this implies the marginal loglinear model

$$\log\big[E(Y_{it})\big] - \frac{1}{2}\mathbf{z}'_{it}\boldsymbol{\Sigma}\mathbf{z}_{it} = \mathbf{x}'_{it}\boldsymbol{\beta},$$

with the same fixed effects but with offset term. For the random-intercept case, indicate the role of σ on the size of the offset. Explain what happens when $\sigma = 0$.

13.43 In Section 13.5.1 and Problem 13.42 we saw that for Poisson GLMMs, the marginal effects are the same as the cluster-specific effects. This does not imply that ML estimates of effects are the same for a Poisson GLMM and a Poisson GLM. Explain why. (*Hint*: For the GLMM, is the marginal distribution Poisson?)

13.44 For the Poisson GLMM (13.14), use the normal mgf to show that for $t \neq s$,

$$\text{cov}(Y_{it}, Y_{is}) = \exp\left[(\mathbf{x}'_{it} + \mathbf{x}'_{is})\boldsymbol{\beta}\right]\left[\exp(\sigma^2)\left(\exp(\sigma^2) - 1\right)\right]$$

Hence, find $\text{corr}(Y_{it}, Y_{is})$.

13.45 Consider a Poisson GLMM using the identity link. Relate the marginal mean and variance to the conditional mean and variance. Explain the structural problem that this model has.

Categorical Data Analysis, Second Edition. Alan Agresti

ISBN: 0-471-36093-7

CHAPTER 14

Asymptotic Theory for Parametric Models

This chapter has a more theoretical flavor than others. It presents asymptotic theory for parametric models for categorical data, with emphasis on multinomial models for contingency tables. In Section 14.1 we review and extend the delta method. This is used to derive large-sample normal distributions for many statistics. In Section 14.2 we apply the delta method to ML estimation of parameters in models for contingency tables, later illustrated in Section 14.4 for logit and loglinear models. In Section 14.3 we derive asymptotic distributions of cell residuals and the X^2 and G^2 goodness-of-fit statistics.

The results in this chapter have a long history. Pearson (1900) derived the asymptotic chi-squared distribution of X^2 for testing a specified multinomial distribution. Fisher (1922, 1924) showed the adjustment in degrees of freedom when multinomial probabilities are functions of unknown parameters. Cramér (1946, pp. 424–434) formally proved this result, under the assumption that ML estimators of the parameters are consistent. Rao (1957) proved consistency of the ML estimators under general conditions. He also gave the asymptotic distribution of the ML estimators, although the primary emphasis of his articles was on proving consistency. Birch (1964a) proved these results under weaker conditions. Andersen (1980), Bishop et al. (1975), Cox (1984), Haberman (1974a), and Watson (1959) provided other proofs or considered related cases.

As in Cramér's and Rao's proofs, our derivation regards the ML estimator as a point in the parameter space where the derivative of the log likelihood function is zero. Birch regarded it as a point at which the likelihood takes value arbitrarily near its supremum. Although his approach is more powerful, the proofs are more complex. We avoid a formal "theorem–proof" style of exposition. Instead, we show that powerful results follow from simple mathematical ideas, such as Taylor series expansions.

14.1 DELTA METHOD

Suppose that a statistic used as an estimator of a parameter has a large-sample normal distribution. Then, in this section we show that many functions of that statistic are also asymptotically normal.

14.1.1 *O*, *o* Rates of Convergence

Big O and little o notation is useful for describing limiting behavior of sequences. For real numbers $\{z_n\}$, the *little o* notation $o(z_n)$ represents a term that has *smaller* order than z_n as $n \to \infty$, in the sense that $o(z_n)/z_n \to 0$ as $n \to \infty$. For instance, $\sqrt{n}$ is $o(n)$ as $n \to \infty$, since $\sqrt{n}/n \to 0$ as $n \to \infty$. A sequence that is $o(1)$ satisfies $o(1)/1 = o(1) \to 0$; for instance, $n^{-1/2}$ is $o(1)$ as $n \to \infty$.

The *big O* notation $O(z_n)$ represents terms that have the *same* order of magnitude as z_n, in the sense that $|O(z_n)/z_n|$ is bounded as $n \to \infty$. For instance, $(3/n) + (8/n^2)$ is $O(n^{-1})$ as $n \to \infty$; dividing it by n^{-1} gives a ratio that takes value close to 3 as n increases.

Similar notation applies to sequences of random variables. This notation uses a subscript p to indicate that the sequence has probabilistic rather than deterministic behavior. The symbol $o_p(z_n)$ denotes a random variable of *smaller* order than z_n for large n, in the sense that $o_p(z_n)/z_n$ *converges in probability* to 0; that is, for any fixed $\epsilon > 0$, $P(|o_p(z_n)/z_n| \leq \epsilon) \to 1$ as $n \to \infty$. The notation $O_p(z_n)$ represents a random variable such that for every $\epsilon > 0$, there is a constant K and an integer n_0 such that $P[|O_p(z_n)/z_n| < K] > 1 - \epsilon$ for all $n > n_0$.

To illustrate, let $\bar{Y}_n$ denote the sample mean of n independent observations $Y_1, \ldots, Y_n$ from a distribution having $E(Y_i) = \mu$. Then $(\bar{Y}_n - \mu) = o_p(1)$, since $(\bar{Y}_n - \mu)/1$ converges in probability to zero as $n \to \infty$ by the law of large numbers. By Tchebychev's inequality, the difference between a random variable and its expected value has the same order of magnitude as the standard deviation of that random variable. Since $\bar{Y}_n - \mu$ has standard deviation $\sigma/\sqrt{n}$, $(\bar{Y}_n - \mu) = O_p(n^{-1/2})$.

A random variable that is $O_p(n^{-1/2})$ is also $o_p(1)$. An example is $(\bar{Y}_n - \mu)$. Multiplication affects the order in the way one expects intuitively (Problem 14.1). For instance, $\sqrt{n}(\bar{Y}_n - \mu) = n^{1/2} O_p(n^{-1/2}) = O_p(n^{1/2} n^{-1/2}) = O_p(1)$. If the difference between two random variables is $o_p(1)$ as $n \to \infty$, Slutzky's theorem states that those random variables have the same limiting distribution.

14.1.2 Delta Method for Function of Random Variable

Let T_n denote a statistic, the subscript expressing its dependence on the sample size n. For large samples, suppose that T_n is approximately normally

distributed about θ, with approximate standard error $\sigma/\sqrt{n}$. More precisely, as $n \to \infty$, suppose that the cdf of $\sqrt{n}(T_n - \theta)$ converges to a $N(0, \sigma^2)$ cdf. This limiting behavior is an example of *convergence in distribution*, denoted

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2). \tag{14.1}$$

For a function g, we now derive the limiting distribution of $g(T_n)$. Suppose that g is at least twice differentiable at θ. We use the Taylor series expansion for $g(t)$ in a neighborhood of θ. For some θ^* between t and θ,

$$\begin{aligned} g(t) &= g(\theta) + (t - \theta)g'(\theta) + (t - \theta)^2 g''(\theta^*)/2 \\ &= g(\theta) + (t - \theta)g'(\theta) + O\big(|t - \theta|^2\big). \end{aligned}$$

Substituting the random variable T_n for t, we have

$$\begin{aligned} \sqrt{n}\,[g(T_n) - g(\theta)] &= \sqrt{n}\,(T_n - \theta)g'(\theta) + \sqrt{n}\,O\big(|T_n - \theta|^2\big) \\ &= \sqrt{n}\,(T_n - \theta)g'(\theta) + O_p(n^{-1/2}) \end{aligned} \tag{14.2}$$

since

$$\sqrt{n}\,O\big(|T_n - \theta|^2\big) = \sqrt{n}\,O\big[O_p(n^{-1})\big] = O_p(n^{-1/2}).$$

Since the $O_p(n^{-1/2})$ term is asymptotically negligible, $\sqrt{n}\,[g(T_n) - g(\theta)]$ has the same limiting distribution as $\sqrt{n}\,(T_n - \theta)g'(\theta)$; that is, $g(T_n) - g(\theta)$ behaves like the constant multiple $g'(\theta)$ of $(T_n - \theta)$. Now, $(T_n - \theta)$ is approximately normal with variance σ^2/n. Thus, $g(T_n) - g(\theta)$ is approximately normal with variance $\sigma^2[g'(\theta)]^2/n$. More precisely,

$$\sqrt{n}\,[g(T_n) - g(\theta)] \xrightarrow{d} N\big(0, \sigma^2[g'(\theta)]^2\big). \tag{14.3}$$

Figure 3.1 illustrated this result, and in Section 3.1.6 it was applied to the sample logit.

Result (14.3) is called the *delta method* for obtaining asymptotic distributions. Since $\sigma^2 = \sigma^2(\theta)$ and $g'(\theta)$ usually depends on θ, the asymptotic variance is unknown. Let $\sigma^2(T_n)$ and $g'(T_n)$ denote these terms evaluated at the sample estimator T_n of θ. When $g'(\cdot)$ and $\sigma = \sigma(\cdot)$ are continuous at θ, $\sigma(T_n)g'(T_n)$ is a consistent estimator of $\sigma(\theta)g'(\theta)$. Thus, confidence intervals and tests use the result that $\sqrt{n}\,[g(T_n) - g(\theta)]/\sigma(T_n)\,|g'(T_n)|$ is asymptotically standard normal. For instance,

$$g(T_n) \pm z_{\alpha/2}\,\sigma(T_n)\,|g'(T_n)|/\sqrt{n}$$

is a large-sample $100(1 - \alpha)\%$ confidence interval for $g(\theta)$.

When $g'(\theta) = 0$, (14.3) is uninformative because the limiting variance equals 0. In that case, $\sqrt{n}\,[g(T_n) - g(\theta)] = o_p(1)$, and higher-order terms in the Taylor series expansion yield the asymptotic distribution (see Note 14.1).

14.1.3 Delta Method for Function of Random Vector

The delta method generalizes to functions of random *vectors*. Suppose that $\mathbf{T}_n = (T_{n1}, \ldots, T_{nN})'$ is asymptotically multivariate normal with mean $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)'$ and covariance matrix $\boldsymbol{\Sigma}/n$. Suppose that $g(t_1, \ldots, t_N)$ has a nonzero differential $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N)'$ at $\boldsymbol{\theta}$, where

$$\phi_i = \left.\frac{\partial g}{\partial t_i}\right|_{\mathbf{t}=\boldsymbol{\theta}}.$$

Then,

$$\sqrt{n}\,\big[g(\mathbf{T}_n) - g(\boldsymbol{\theta})\big] \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\phi}'\boldsymbol{\Sigma}\boldsymbol{\phi}). \tag{14.4}$$

For large n, $g(\mathbf{T}_n)$ has distribution similar to the normal with mean $g(\boldsymbol{\theta})$ and variance $\boldsymbol{\phi}'\boldsymbol{\Sigma}\boldsymbol{\phi}/n$.

The proof of (14.4) follows from the expansion

$$g(\mathbf{T}_n) - g(\boldsymbol{\theta}) = (\mathbf{T}_n - \boldsymbol{\theta})'\boldsymbol{\phi} + o(\|\mathbf{T}_n - \boldsymbol{\theta}\|),$$

where $\|\mathbf{z}\| = (\Sigma z_i^2)^{1/2}$ denotes the length of vector $\mathbf{z}$. For large n, $g(\mathbf{T}_n) - g(\boldsymbol{\theta})$ behaves like a linear function of the approximately normal random vector $(\mathbf{T}_n - \boldsymbol{\theta})$. Thus, it itself is approximately normal.

14.1.4 Asymptotic Normality of Functions of Multinomial Counts

The delta method for random vectors implies asymptotic normality of many functions of cell counts in contingency tables. Suppose that cell counts $(n_1, \ldots, n_N)$ have a multinomial distribution with cell probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)'$. Let $n = n_1 + \cdots + n_N$, and let $\mathbf{p} = (p_1, \ldots, p_N)'$ denote the sample proportions, where $p_i = n_i/n$.

Denote observation i of the n cross-classified in the contingency table by $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iN})$, where $Y_{ij} = 1$ if it falls in cell j, and $Y_{ij} = 0$ otherwise, $i = 1, \ldots, n$. For instance, $\mathbf{Y}_6 = (0, 0, 1, 0, 0, \ldots, 0)$ means that observation 6 is in the third cell of the table. Now, since each observation falls in just one cell, $\Sigma_j Y_{ij} = 1$ and $Y_{ij}Y_{ik} = 0$ when $j \neq k$. Also, $p_j = \Sigma_i Y_{ij}/n$, and

$$E(Y_{ij}) = P(Y_{ij} = 1) = \pi_j = E(Y_{ij}^2), \qquad E(Y_{ij}Y_{ik}) = 0 \quad \text{if } j \neq k.$$

It follows that

$$E(\mathbf{Y}_i) = \boldsymbol{\pi} \quad \text{and} \quad \operatorname{cov}(\mathbf{Y}_i) = \boldsymbol{\Sigma}, \quad i = 1, \ldots, n,$$

where $\boldsymbol{\Sigma} = (\sigma_{jk})$ with

$$\sigma_{jj} = \text{var}(Y_{ij}) = E(Y_{ij}^2) - \left[E(Y_{ij})\right]^2 = \pi_j(1 - \pi_j),$$
$$\sigma_{jk} = \text{cov}(Y_{ij}, Y_{ik}) = E(Y_{ij}Y_{ik}) - E(Y_{ij})E(Y_{ik}) = -\pi_j\pi_k \quad \text{for } j \neq k.$$

The matrix $\boldsymbol{\Sigma}$ has form

$$\boldsymbol{\Sigma} = \mathbf{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$$

where $\mathbf{diag}(\boldsymbol{\pi})$ is the diagonal matrix with the elements of $\boldsymbol{\pi}$ on the main diagonal.

Since $\mathbf{p}$ is a sample mean of n independent observations, namely

$$\mathbf{p} = \frac{\sum_{i=1}^n \mathbf{Y}_i}{n},$$
$$\text{cov}(\mathbf{p}) = \left[\mathbf{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'\right]/n. \tag{14.5}$$

This covariance matrix is singular, because of the linear dependence $\Sigma p_i = 1$. The multivariate central limit theorem (Rao 1973, p. 128) implies

$$\sqrt{n}\,(\mathbf{p} - \boldsymbol{\pi}) \xrightarrow{d} N\left[\mathbf{0}, \mathbf{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'\right]. \tag{14.6}$$

By the delta method, functions of $\mathbf{p}$ having nonzero differential at $\boldsymbol{\pi}$ are also asymptotically normal. Let $g(t_1, \ldots, t_N)$ be a differentiable function, and let

$$\phi_i = \partial g/\partial \pi_i, \qquad i = 1, \ldots, N,$$

denote $\partial g/\partial t_i$ evaluated at $\mathbf{t} = \boldsymbol{\pi}$. By the delta method (14.4),

$$\sqrt{n}\left[g(\mathbf{p}) - g(\boldsymbol{\pi})\right] \xrightarrow{d} N\left(0, \boldsymbol{\phi}'\left[\mathbf{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'\right]\boldsymbol{\phi}\right). \tag{14.7}$$

The asymptotic variance equals

$$\boldsymbol{\phi}'\mathbf{diag}(\boldsymbol{\pi})\boldsymbol{\phi} - (\boldsymbol{\phi}'\boldsymbol{\pi})^2 = \sum \pi_i\phi_i^2 - \left(\sum \pi_i\phi_i\right)^2.$$

In Section 3.1.7 we used this formula to derive the large-sample variance of the sample log odds ratio.

14.1.5 Delta Method for Vector Function of Random Vector

The delta method generalizes further to a *vector* of functions of an asymptotically normal random vector. Let $\mathbf{g}(\mathbf{t}) = (g_1(\mathbf{t}), \ldots, g_q(\mathbf{t}))'$ and let $(\partial \mathbf{g}/\partial \boldsymbol{\theta})$ denote the $q \times N$ Jacobian matrix for which the entry in row i and column j

is $\partial g_i(\mathbf{t})/\partial t_j$ evaluated at $\mathbf{t} = \boldsymbol{\theta}$. Then,

$$\sqrt{n}\,[\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\theta})] \xrightarrow{d} N[\mathbf{0}, (\partial\mathbf{g}/\partial\boldsymbol{\theta})\boldsymbol{\Sigma}(\partial\mathbf{g}/\partial\boldsymbol{\theta})'] . \tag{14.8}$$

The rank of the limiting normal distribution equals the rank of the asymptotic covariance matrix.

Expression (14.8) is useful for finding large-sample joint distributions. For instance, from (14.6), (14.7), and (14.8), the asymptotic distribution of several functions of multinomial proportions has covariance matrix of the form

$$\text{asymp. cov}\{\sqrt{n}\,[\mathbf{g}(\mathbf{p}) - \mathbf{g}(\boldsymbol{\pi})]\} = \boldsymbol{\Phi}[\mathbf{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}']\boldsymbol{\Phi}',$$

where $\boldsymbol{\Phi}$ is the Jacobian $(\partial\mathbf{g}/\partial\boldsymbol{\pi})$.

14.1.6 Joint Asymptotic Normality of Log Odds Ratios

We illustrate formula (14.8) by finding the joint asymptotic distribution of a set of log odds ratios in a contingency table. We use the log scale because convergence to normality is more rapid for it.

Let $\mathbf{g}(\boldsymbol{\pi}) = \log(\boldsymbol{\pi})$ denote the vector of natural logs of cell probabilities, for which

$$\partial\mathbf{g}/\partial\boldsymbol{\pi} = \mathbf{diag}(\boldsymbol{\pi})^{-1}.$$

The covariance of the asymptotic distribution of $\sqrt{n}[\log(\mathbf{p}) - \log(\boldsymbol{\pi})]$ is

$$\text{diag}(\boldsymbol{\pi})^{-1}[\mathbf{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}']\mathbf{diag}(\boldsymbol{\pi})^{-1} = \mathbf{diag}(\boldsymbol{\pi})^{-1} - \mathbf{1}\mathbf{1}'$$

where $\mathbf{1}$ is an $N \times 1$ vector of 1 elements.

For a $q \times N$ matrix of constants $\mathbf{C}$, it follows that

$$\sqrt{n}\,\mathbf{C}[\log(\mathbf{p}) - \log(\boldsymbol{\pi})] \xrightarrow{d} N\left[\mathbf{0}, \mathbf{C}\,\mathbf{diag}(\boldsymbol{\pi})^{-1}\mathbf{C}' - \mathbf{C}\mathbf{1}\mathbf{1}'\mathbf{C}'\right]. \tag{14.9}$$

Now, suppose $\mathbf{C}\log(\mathbf{p})$ is a set of sample log odds ratios. Then, each row of $\mathbf{C}$ contains zeros except for two $+1$ elements and two -1 elements in the positions multiplied by the relevant elements of $\log(\mathbf{p})$ to form the given log odds ratio. The second term in the covariance matrix in (14.9) is then zero. If a particular odds ratio uses the cells numbered h, i, j, and k, the variance of the asymptotic distribution is

$$\text{asymp. var}\left[\sqrt{\text{n}}\,(\text{sample log odds ratio})\right] = \pi_h^{-1} + \pi_i^{-1} + \pi_j^{-1} + \pi_k^{-1}.$$

When two log odds ratios have no cells in common, their asymptotic covariance in the limiting normal distribution equals zero.

14.2 ASYMPTOTIC DISTRIBUTIONS OF ESTIMATORS OF MODEL PARAMETERS AND CELL PROBABILITIES

We now derive basic results of large-sample model-based inference for contingency tables. The delta method is the key tool. The derivations apply to a single multinomial distribution. They extend directly to products of multinomials, when the parameter space stays fixed as the sample size increases.

The observations are counts $\mathbf{n} = (n_1, \ldots, n_N)'$ in N cells of a contingency table. The asymptotics regard N as fixed and let $n = \sum n_i \to \infty$. We assume that $\mathbf{n} = n\mathbf{p}$ has a multinomial distribution with probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)'$. The model is

$$\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta}),$$

where $\boldsymbol{\pi}(\boldsymbol{\theta})$ denotes a function that relates $\boldsymbol{\pi}$ to a smaller number of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$.

As $\boldsymbol{\theta}$ ranges over its parameter space, $\boldsymbol{\pi}(\boldsymbol{\theta})$ ranges over a subset of the space of $\boldsymbol{\pi}$ for N probabilities. Adding components to $\boldsymbol{\theta}$, the model becomes more complex and the space of $\boldsymbol{\pi}$ that satisfy the model is larger. We use $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ to denote generic parameter and probability values, and $\boldsymbol{\theta}_0 = (\theta_{10}, \ldots, \theta_{q0})'$ and $\boldsymbol{\pi}_0 = (\pi_{10}, \ldots, \pi_{N0})' = \pi(\boldsymbol{\theta}_0)$ to denote true values for a particular application. When the model does not hold, no $\boldsymbol{\theta}_0$ exists for which $\boldsymbol{\pi}(\boldsymbol{\theta}_0) = \boldsymbol{\pi}_0$; that is, $\boldsymbol{\pi}_0$ falls outside the subset of $\boldsymbol{\pi}$ values that is the range of $\boldsymbol{\pi}(\boldsymbol{\theta})$ for the space of possible $\boldsymbol{\theta}$. We consider this case in Section 14.3.5.

We first derive the asymptotic distribution of the ML estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. We use that to derive the asymptotic distribution of the model-based ML estimator $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})$ of $\boldsymbol{\pi}$. The approach follows Rao (1973, Sec. 5e) and Bishop et al. (1975, Secs. 14.7 and 14.8). The assumed regularity conditions are:

1. $\boldsymbol{\theta}_0$ is not on the boundary of the parameter space.
2. All $\pi_{i0} > 0$.
3. $\boldsymbol{\pi}(\boldsymbol{\theta})$ has continuous first-order partial derivatives in a neighborhood of $\boldsymbol{\theta}_0$.
4. The Jacobian matrix $(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta})$ has full rank q at $\boldsymbol{\theta}_0$.

These conditions ensure that $\boldsymbol{\pi}(\boldsymbol{\theta})$ is locally smooth and one-to-one at $\boldsymbol{\theta}_0$ and Taylor series expansions exist in neighborhoods around $\boldsymbol{\theta}_0$ and $\boldsymbol{\pi}_0$. When the Jacobian does not have full rank, often it does with reformulation of the model using fewer parameters.

14.2.1 Distribution of Model Parameter Estimator

The key to deriving the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ is to express $\hat{\boldsymbol{\theta}}$ as a linearized function of $\mathbf{p}$. Then the delta method applies, using the asymptotic

normality of $\mathbf{p}$. The linearization has two steps, first relating $\mathbf{p}$ to $\hat{\boldsymbol{\pi}}$, and then $\hat{\boldsymbol{\pi}}$ to $\hat{\boldsymbol{\theta}}$.

The kernel of the multinomial log likelihood is

$$L(\boldsymbol{\theta}) = \log \prod_{i=1}^{N} \pi_i(\boldsymbol{\theta})^{n_i} = n \sum_{i=1}^{N} p_i \log \pi_i(\boldsymbol{\theta}).$$

The likelihood equations are

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_j} = n \sum_i \frac{p_i}{\pi_i(\boldsymbol{\theta})} \frac{\partial \pi_i(\boldsymbol{\theta})}{\partial \theta_j} = 0, \quad j = 1, \ldots, q. \tag{14.10}$$

These depend on the functional form $\boldsymbol{\pi}(\boldsymbol{\theta})$ used in the model. Note that

$$\sum_i \frac{\partial \pi_i(\boldsymbol{\theta})}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \left[\sum_i \pi_i(\boldsymbol{\theta}) \right] = \frac{\partial}{\partial \theta_j}(1) = 0. \tag{14.11}$$

Let $\partial \pi_i / \partial \hat{\theta}_j$ represent $\partial \pi_i(\boldsymbol{\theta}) / \partial \theta_j$ evaluated at $\hat{\boldsymbol{\theta}}$. Subtracting a common term from both sides of the jth likelihood equation (14.10),

$$\sum_i \frac{n(p_i - \pi_{i0})}{\hat{\pi}_i} \frac{\partial \pi_i}{\partial \hat{\theta}_j} = \sum_i \frac{n(\hat{\pi}_i - \pi_{i0})}{\hat{\pi}_i} \frac{\partial \pi_i}{\partial \hat{\theta}_j}, \tag{14.12}$$

since the first sum on the right-hand side equals zero from (14.11).

Next we express $\hat{\boldsymbol{\pi}}$ in terms of $\hat{\boldsymbol{\theta}}$ using

$$\hat{\pi}_i - \pi_{i0} = \sum_k (\hat{\theta}_k - \theta_{k0}) \frac{\partial \pi_i}{\partial \bar{\theta}_k}$$

where $\partial \pi_i / \partial \bar{\theta}_k$ represents $\partial \pi_i / \partial \theta_k$ evaluated at some point $\bar{\boldsymbol{\theta}}$ falling between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Substitution of this into the right-hand side of (14.12) and division of both sides by $\sqrt{n}$ yields, for each j,

$$\sum_i \frac{\sqrt{n}(p_i - \pi_{i0})}{\hat{\pi}_i} \frac{\partial \pi_i}{\partial \hat{\theta}_j} = \sum_k \sqrt{n}(\hat{\theta}_k - \theta_{k0}) \left(\sum_i \frac{1}{\hat{\pi}_i} \frac{\partial \pi_i}{\partial \hat{\theta}_j} \frac{\partial \pi_i}{\partial \bar{\theta}_k} \right). \tag{14.13}$$

Some notation lets us express more simply the dependence of $\hat{\boldsymbol{\theta}}$ on $\mathbf{p}$. Let $\mathbf{A}$ denote the $N \times q$ matrix having elements

$$a_{ij} = \pi_{i0}^{-1/2} \frac{\partial \pi_i(\boldsymbol{\theta})}{\partial \theta_{j0}}.$$

The matrix expression for $\mathbf{A}$ is

$$\mathbf{A} = \mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0), \tag{14.14}$$

where $(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0)$ denotes the Jacobian $(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}_0$. As $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$, the term in brackets on the right-hand side of (14.13) converges to the element in row j and column k of $\mathbf{A}'\mathbf{A}$. As $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$, the set of equations (14.13) has the form

$$\mathbf{A}'\mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}\sqrt{n}\,(\mathbf{p} - \boldsymbol{\pi}_0) = (\mathbf{A}'\mathbf{A})\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) + o_p(1).$$

Since the Jacobian has full rank at $\boldsymbol{\theta}_0$, $\mathbf{A}'\mathbf{A}$ is nonsingular. Thus,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}\sqrt{n}\,(\mathbf{p} - \boldsymbol{\pi}_0) + o_p(1). \tag{14.15}$$

Now, the asymptotic distribution of $\mathbf{p}$ determines that of $\hat{\boldsymbol{\theta}}$. From (14.6), $\sqrt{n}\,(\mathbf{p} - \boldsymbol{\pi}_0)$ is asymptotically normal, with covariance matrix $[\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0']$. By the delta method, $\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is also asymptotically normal, with asymptotic covariance matrix

$$(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2} \times \left[\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0'\right] \times \mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}.$$

Using (14.11) and (14.14), the term subtracted in this expression disappears because

$$\begin{aligned}\boldsymbol{\pi}_0'\mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}\mathbf{A} &= \boldsymbol{\pi}_0'\mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}\mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0)\\ &= \mathbf{1}'(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0) = \left(\sum_i \partial\pi_i/\partial\boldsymbol{\theta}_0\right)' = \mathbf{0}'.\end{aligned}$$

Thus, this asymptotic covariance expression for $\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ simplifies to $(\mathbf{A}'\mathbf{A})^{-1}$.

In summary, this argument establishes the general result

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left[\mathbf{0}, (\mathbf{A}'\mathbf{A})^{-1}\right]. \tag{14.16}$$

The asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ depends on $(\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}_0)$ and hence on the function for modeling $\boldsymbol{\pi}$ in terms of $\boldsymbol{\theta}$. Let $\hat{\mathbf{A}}$ denote $\mathbf{A}$ evaluated at the ML estimate $\hat{\boldsymbol{\theta}}$. The estimated covariance matrix is

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\theta}}) = (\hat{\mathbf{A}}'\hat{\mathbf{A}})^{-1}/n.$$

The asymptotic normality and covariance of $\hat{\boldsymbol{\theta}}$ follows more simply from general results for ML estimators. However, those results require stronger

regularity conditions (Rao 1973, p. 364) than the ones assumed here. Suppose that observations are independent from $f(\mathbf{y}; \boldsymbol{\theta})$, some probability mass function. The ML estimator $\hat{\boldsymbol{\theta}}$ is efficient, in the sense that

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}^{-1}),$$

where $\mathcal{I}$ is the information matrix for a single observation. The (j, k) element of $\mathcal{I}$ is

$$-E\left(\frac{\partial^2 \log f(\mathbf{y}, \boldsymbol{\theta})}{\partial\theta_j \partial\theta_k}\right) = E\left[\frac{\partial \log f(\mathbf{y}, \boldsymbol{\theta})}{\partial\theta_j} \cdot \frac{\partial \log f(\mathbf{y}, \boldsymbol{\theta})}{\partial\theta_k}\right].$$

When f is the probability of a single observation having multinomial probabilities $\{\pi_1(\boldsymbol{\theta}), \ldots, \pi_N(\boldsymbol{\theta})\}$, this element of $\mathcal{I}$ equals

$$\sum_{i=1}^{N} \frac{\partial \log(\pi_i(\boldsymbol{\theta}))}{\partial\theta_j} \frac{\partial \log(\pi_i(\boldsymbol{\theta}))}{\partial\theta_k} \pi_i(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\partial \pi_i(\boldsymbol{\theta})}{\partial\theta_j} \frac{\partial \pi_i(\boldsymbol{\theta})}{\partial\theta_k} \frac{1}{\pi_i(\boldsymbol{\theta})}.$$

This is the (j, k) element of $\mathbf{A}'\mathbf{A}$. Thus the asymptotic covariance is $\mathcal{I}^{-1} = (\mathbf{A}'\mathbf{A})^{-1}$.

For results of this section to apply, a ML estimator of $\boldsymbol{\theta}$ must exist and be a solution of the likelihood equations. This requires the following *strong identifiability* condition: For every $\epsilon > 0$, there exists a $\delta > 0$ such that if $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \epsilon$, then $\|\boldsymbol{\pi}(\boldsymbol{\theta}) - \boldsymbol{\pi}_0\| > \delta$. This condition implies a weaker one that two $\boldsymbol{\theta}$ values cannot have the same $\boldsymbol{\pi}$ value. When strong identifiability and the other regularity conditions hold, the probability an ML estimator is a root of the likelihood equations converges to 1 as $n \to \infty$. That estimator has the asymptotic properties given above of a solution of the likelihood equations. For proofs, see Birch (1964a) and Rao (1973, pp. 360–362).

14.2.2 Asymptotic Distribution of Cell Probability Estimators

The asymptotic distribution of the model-based estimator $\hat{\boldsymbol{\pi}}$ follows from the Taylor-series expansion

$$\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\pi}(\boldsymbol{\theta}_0) + \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(n^{-1/2}). \qquad (14.17)$$

The size of the remainder term follows from $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = O_p(n^{-1/2})$. Now $\boldsymbol{\pi}(\boldsymbol{\theta}_0) = \boldsymbol{\pi}_0$, and $\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normal with asymptotic co-

variance $(\mathbf{A}'\mathbf{A})^{-1}$. By the delta method,

$$\sqrt{n}\,(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0) \xrightarrow{d} N\left[\mathbf{0}, \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}_0}(\mathbf{A}'\mathbf{A})^{-1}\frac{\partial \boldsymbol{\pi}'}{\partial \boldsymbol{\theta}_0}\right]. \tag{14.18}$$

When the model holds with $\boldsymbol{\theta}$ having $q < N - 1$ elements, $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})$ is more efficient than the sample proportion $\mathbf{p}$ for estimating $\boldsymbol{\pi}$. More generally, for estimating a smooth function $g(\boldsymbol{\pi})$ of $\boldsymbol{\pi}$, $g(\hat{\boldsymbol{\pi}})$ has smaller asymptotic variance than $g(\mathbf{p})$. We next derive this result, discussed in Section 6.4.5. The derivation deletes the Nth component from $\mathbf{p}$ and $\hat{\boldsymbol{\pi}}$, so their covariance matrices are positive definite (Problem 14.16). The Nth proportion is linearly dependent on the first $N - 1$ since they sum to 1. Let $\boldsymbol{\Sigma} = \mathbf{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$ denote the $(N-1)\times(N-1)$ covariance matrix of $\sqrt{n}\,\mathbf{p}$. The inverse of $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma}^{-1} = \mathbf{diag}(\boldsymbol{\pi})^{-1} + \mathbf{1}\mathbf{1}'\pi_N, \tag{14.19}$$

which can be verified by evaluating $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}$ and showing that it equals the identity matrix.

Let $(\partial g/\partial \boldsymbol{\pi}_0) = (\partial g/\partial \pi_1, \ldots, \partial g/\partial \pi_{N-1})'$, evaluated at $\boldsymbol{\pi} = \boldsymbol{\pi}_0$. By the delta method,

$$\text{asymp. var}\left[\sqrt{n}\,g(\mathbf{p})\right] = \left(\frac{\partial g}{\partial \boldsymbol{\pi}_0}\right)'\left[\text{cov}\left(\sqrt{n}\,\mathbf{p}\right)\right]\frac{\partial g}{\partial \boldsymbol{\pi}_0} = \left(\frac{\partial g}{\partial \boldsymbol{\pi}_0}\right)'\boldsymbol{\Sigma}\frac{\partial g}{\partial \boldsymbol{\pi}_0}$$

and

$$\begin{aligned}\text{Asymp. var}\left[\sqrt{n}\,g(\hat{\boldsymbol{\pi}})\right] &= \left(\frac{\partial g}{\partial \boldsymbol{\pi}_0}\right)'\left[\text{Asymp. cov}\left(\sqrt{n}\,\hat{\boldsymbol{\pi}}\right)\right]\frac{\partial g}{\partial \boldsymbol{\pi}_0} \\ &= \left(\frac{\partial g}{\partial \boldsymbol{\pi}_0}\right)'\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}_0}\left[\text{Asymp. cov}\left(\sqrt{n}\,\hat{\boldsymbol{\theta}}\right)\right]\left(\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}_0}\right)'\frac{\partial g}{\partial \boldsymbol{\pi}_0}.\end{aligned}$$

Using (14.11) and (14.19) yields

$$\begin{aligned}\text{Asymp. cov}\left(\sqrt{n}\,\hat{\boldsymbol{\theta}}\right) = (\mathbf{A}'\mathbf{A})^{-1} &= \left[(\partial \boldsymbol{\pi}/\partial \boldsymbol{\theta}_0)'\mathbf{diag}(\boldsymbol{\pi}_0)^{-1}(\partial \boldsymbol{\pi}/\partial \boldsymbol{\theta}_0)\right]^{-1} \\ &= \left[(\partial \boldsymbol{\pi}/\partial \boldsymbol{\theta}_0)'\boldsymbol{\Sigma}^{-1}(\partial \boldsymbol{\pi}/\partial \boldsymbol{\theta}_0)\right]^{-1}.\end{aligned}$$

Since $\boldsymbol{\Sigma}$ is positive definite and $(\partial \boldsymbol{\pi}/\partial \boldsymbol{\theta}_0)$ has rank q, $\boldsymbol{\Sigma}^{-1}$ and $[(\partial \boldsymbol{\pi}/\partial \boldsymbol{\theta}_0)'\boldsymbol{\Sigma}^{-1}(\partial \boldsymbol{\pi}/\partial \boldsymbol{\theta}_0)]^{-1}$ are also positive definite.

To show that asymp. var$[\sqrt{n}\, g(\mathbf{p})] \geq$ asymp. var$[\sqrt{n}\, g(\hat{\boldsymbol{\pi}})]$, we show that

$$\left(\frac{\partial g}{\partial \boldsymbol{\pi}_0}\right)' \left\{ \boldsymbol{\Sigma} - \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}_0} \left[\left(\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}_0}\right)' \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}_0} \right]^{-1} \left(\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}_0}\right)' \right\} \frac{\partial g}{\partial \boldsymbol{\pi}_0} \geq 0.$$

But this quadratic form is identical to

$$(\mathbf{Y} - \mathbf{B}\boldsymbol{\zeta})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{B}\boldsymbol{\zeta})$$

where $\mathbf{Y} = \boldsymbol{\Sigma}(\partial g / \partial \boldsymbol{\pi}_0)$, $\mathbf{B} = (\partial \boldsymbol{\pi} / \partial \boldsymbol{\theta}_0)$, and $\boldsymbol{\zeta} = (\mathbf{B}'\boldsymbol{\Sigma}^{-1}\mathbf{B})^{-1}\mathbf{B}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$. The result then follows from the positive definiteness of $\boldsymbol{\Sigma}^{-1}$.

This proof is based on one given by Altham (1984). Her proof uses standard properties of ML estimators. It applies whenever regularity conditions hold that guarantee those properties. The proof applies not only to categorical data but to any situation in which a model describes the dependence of a set of parameters $\boldsymbol{\pi}$ on some smaller set $\boldsymbol{\theta}$.

14.3 ASYMPTOTIC DISTRIBUTIONS OF RESIDUALS AND GOODNESS-OF-FIT STATISTICS

We next study the distribution of Pearson X^2 and likelihood-ratio G^2 goodness-of-fit statistics for the multinomial model $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$. We first derive the asymptotic joint distribution of the sample proportions $\mathbf{p}$ and model-based estimator $\hat{\boldsymbol{\pi}}$. This distribution determines large-sample distributions of statistics that depend on both $\mathbf{p}$ and $\hat{\boldsymbol{\pi}}$. For instance, it determines the asymptotic joint distribution of the Pearson residuals, which compare $\mathbf{p}$ with $\hat{\boldsymbol{\pi}}$. Deriving the large-sample chi-squared distribution for X^2, which is the sum of squared Pearson residuals, is then straightforward. We also show that X^2 and G^2 are asymptotically equivalent, when the model holds. Our presentation borrows from Bishop et al. (1975, Chap 14), Cox (1984), Cramér (1946, pp. 432–433), and Rao (1973, Sect. 6b).

14.3.1 Joint Asymptotic Normality of p and $\hat{\boldsymbol{\pi}}$

We first express the joint dependence of $\mathbf{p}$ and $\hat{\boldsymbol{\pi}}$ on $\mathbf{p}$, in order to show the joint asymptotic normality of $\mathbf{p}$ and $\hat{\boldsymbol{\pi}}$. Let

$$\mathbf{D} = \mathbf{diag}(\boldsymbol{\pi}_0)^{1/2} \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1} \mathbf{A}' \mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}.$$

From (14.15) and (14.17),

$$\begin{aligned} \hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0 &= \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}_0} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) + \mathbf{o}_p(n^{-1/2}) \\ &= \mathbf{D}(\mathbf{p} - \boldsymbol{\pi}_0) + \mathbf{o}_p(n^{-1/2}). \end{aligned}$$

Therefore,

$$\sqrt{n}\begin{pmatrix}\mathbf{p}-\boldsymbol{\pi}_0\\ \hat{\boldsymbol{\pi}}-\boldsymbol{\pi}_0\end{pmatrix}=\begin{pmatrix}\mathbf{I}\\ \mathbf{D}\end{pmatrix}\sqrt{n}\,(\mathbf{p}-\boldsymbol{\pi}_0)+\mathbf{o}_p(1)\,,$$

where $\mathbf{I}$ is a $N\times N$ identity matrix. By the delta method,

$$\sqrt{n}\begin{pmatrix}\mathbf{p}-\boldsymbol{\pi}_0\\ \hat{\boldsymbol{\pi}}-\boldsymbol{\pi}_0\end{pmatrix}\xrightarrow{d} N(\mathbf{0},\boldsymbol{\Sigma}^*) \tag{14.20}$$

where

$$\boldsymbol{\Sigma}^*=\begin{pmatrix}\mathbf{diag}(\boldsymbol{\pi}_0)-\boldsymbol{\pi}_0\boldsymbol{\pi}_0' & [\mathbf{diag}(\boldsymbol{\pi}_0)-\boldsymbol{\pi}_0\boldsymbol{\pi}_0']\mathbf{D}'\\ \mathbf{D}[\mathbf{diag}(\boldsymbol{\pi}_0)-\boldsymbol{\pi}_0\boldsymbol{\pi}_0'] & \mathbf{D}[\mathbf{diag}(\boldsymbol{\pi}_0)-\boldsymbol{\pi}_0\boldsymbol{\pi}_0']\mathbf{D}'\end{pmatrix}. \tag{14.21}$$

The two matrix blocks on the main diagonal of $\boldsymbol{\Sigma}^*$ are cov($\sqrt{n}\,\mathbf{p}$) and asymp. cov($\sqrt{n}\,\hat{\boldsymbol{\pi}}$), derived previously. The new information here is that asymp. cov($\sqrt{n}\,\mathbf{p}, \sqrt{n}\,\hat{\boldsymbol{\pi}}$) = $[\mathbf{diag}(\boldsymbol{\pi}_0)-\boldsymbol{\pi}_0\boldsymbol{\pi}_0']\mathbf{D}'$.

14.3.2 Asymptotic Distribution of Pearson and Standardized Residuals

For cell counts $\{n_i\}$ the Pearson statistic is $X^2=\Sigma e_i^2$, where

$$e_i=\frac{n_i-\hat{\mu}_i}{\hat{\mu}_i^{1/2}}=\frac{\sqrt{n}\,(p_i-\hat{\pi}_i)}{\hat{\pi}_i^{1/2}}.$$

We next derive the asymptotic distribution of $\mathbf{e}=(e_1,\ldots,e_N)'$, which is a diagnostic measure of lack of fit. For Poisson models it is the Pearson residual. Dividing it by its standard error gives the standardized residual. The distribution of $\mathbf{e}$ is also helpful in deriving the distribution of X^2.

The residuals $\mathbf{e}$ are functions of $\mathbf{p}$ and $\hat{\boldsymbol{\pi}}$, which are jointly asymptotically normal from (14.20). To use the delta method, we calculate

$$\partial e_i/\partial p_i=\sqrt{n}\,\hat{\pi}_i^{-1/2}, \qquad \partial e_i/\partial\hat{\pi}_i=-\sqrt{n}\,(p_i+\hat{\pi}_i)/2\hat{\pi}_i^{-3/2}$$
$$\partial e_i/\partial p_j=\partial e_i/\partial\hat{\pi}_j=0 \qquad \text{for } i\neq j.$$

That is,

$$\frac{\partial\mathbf{e}}{\partial\mathbf{p}}=\sqrt{n}\,\mathbf{diag}(\hat{\boldsymbol{\pi}})^{-1/2} \qquad \text{and}$$

$$\frac{\partial\mathbf{e}}{\partial\hat{\boldsymbol{\pi}}}=-\left(\tfrac{1}{2}\right)\sqrt{n}\,[\mathbf{diag}(\mathbf{p})+\mathbf{diag}(\hat{\boldsymbol{\pi}})]\mathbf{diag}(\hat{\boldsymbol{\pi}})^{-3/2}. \tag{14.22}$$

Evaluated at $\mathbf{p}=\boldsymbol{\pi}_0$ and $\hat{\boldsymbol{\pi}}=\boldsymbol{\pi}_0$, these matrices equal $\sqrt{n}\,\mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}$ and $-\sqrt{n}\,\mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}$. Using (14.21), (14.22), and $\mathbf{A}'\boldsymbol{\pi}_0^{1/2}=\mathbf{0}$ [which follows

from (14.11)], the delta method implies that

$$\mathbf{e} \xrightarrow{d} N\left(\mathbf{0}, \mathbf{I} - \boldsymbol{\pi}_0^{1/2}\boldsymbol{\pi}_0^{1/2\prime} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\right). \tag{14.23}$$

The limiting distribution has form $N(\mathbf{0}, \mathbf{I} - \mathbf{Hat})$, where **Hat** is the *hat matrix* (Section 4.5.5). Although asymptotically normal, **e** behaves less variably than standard normal random variables. The standardized Pearson residual (Haberman 1973a) divides **e** by its estimated standard error. This statistic, which *is* asymptotically standard normal, equals

$$r_i = \frac{e_i}{\left[1 - \hat{\pi}_i - \Sigma_j\Sigma_k(1/\hat{\pi}_i)\left(\partial\pi_i/\partial\hat{\theta}_j\right)\left(\partial\pi_i/\partial\hat{\theta}_k\right)\hat{v}^{jk}\right]^{1/2}}, \tag{14.24}$$

where $\hat{v}^{jk}$ denotes the element in row j and column k of $(\hat{\mathbf{A}}'\hat{\mathbf{A}})^{-1}$. The denominator of r_i is $\sqrt{1 - \hat{h}_i}$, where the leverage $\hat{h}_i$ for observation i estimates the ith diagonal element of the hat matrix. This simplifies to (3.13) for testing independence in two-way tables.

14.3.3 Asymptotic Distribution of Pearson Statistic

The proof that the Pearson X^2 statistic has an asymptotic chi-squared distribution uses the following relationship between normal and chi-squared distributions (Rao 1973, p. 188):

> Let **X** be multivariate normal with mean $\boldsymbol{\nu}$ and covariance matrix **B**. A necessary and sufficient condition for $(\mathbf{X} - \boldsymbol{\nu})'\mathbf{C}(\mathbf{X} - \boldsymbol{\nu})$ to have a chi-squared distribution is **BCBCB** = **BCB**. The degrees of freedom equal the rank of **CB**.

When **B** is nonsingular, the condition simplifies to **CBC** = **C**.

The Pearson statistic relates to **e** by $X^2 = \mathbf{e}'\mathbf{e}$, so we apply this result by identifying **X** with **e**, $\boldsymbol{\nu} = \mathbf{0}$, $\mathbf{C} = \mathbf{I}$, and $\mathbf{B} = \mathbf{I} - \boldsymbol{\pi}_0^{1/2}\boldsymbol{\pi}_0^{1/2\prime} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\}$. Since $\mathbf{C} = \mathbf{I}$, the condition for $(\mathbf{X} - \boldsymbol{\nu}')\mathbf{C}(\mathbf{X} - \boldsymbol{\nu}) = \{\mathbf{e}'\mathbf{e}\} = X^2$ to have a chi-squared distribution simplifies to **BBB** = **BB**. A direct computation using $\mathbf{A}'\boldsymbol{\pi}_0^{1/2} = \mathbf{0}$ shows that **B** is idempotent, so the condition holds. Since **e** is asymptotically multivariate normal, X^2 is asymptotically chi-squared.

For symmetric idempotent matrices, the rank equals the trace. The trace of **I** is N; the trace of $\boldsymbol{\pi}_0^{1/2}\boldsymbol{\pi}_0^{1/2\prime}$ equals the trace of $\boldsymbol{\pi}_0^{1/2\prime}\boldsymbol{\pi}_0^{1/2} = \Sigma\pi_{i0} = 1$, which is 1; the trace of $\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ equals the trace of $(\mathbf{A}'\mathbf{A})^{-1}(\mathbf{A}'\mathbf{A})$ = identity matrix of size $q \times q$, which is q. Thus, the rank of $\mathbf{B} = \mathbf{CB}$ is $N - q - 1$, and the asymptotic chi-squared distribution has df $= N - q - 1$.

This result, due to Fisher (1922), is remarkably simple. When the sample size is large, the distribution of X^2 does not depend on $\boldsymbol{\pi}_0$ or the model form. It depends only on the difference between the dimension of $\boldsymbol{\pi}$, which is $N - 1$, and the dimension of $\boldsymbol{\theta}$. With $q = 0$ parameters, X^2 is Pearson's

(1900) statistic (1.15) for testing that multinomial probabilities equal certain specified values, and df $= N - 1$ as Pearson claimed. Watson (1959) showed that the same result holds for the asymptotic conditional distribution, given a sufficient statistic for nuisance parameters.

14.3.4 Asymptotic Distribution of Likelihood-Ratio Statistic

When the model holds, the likelihood-ratio statistic G^2 is asymptotically equivalent to X^2 as $n \to \infty$. To show this, we express

$$G^2 = 2\sum_i n_i \log\frac{n_i}{\hat{\mu}_i} = 2n\sum_i p_i \log\left(1 + \frac{p_i - \hat{\pi}_i}{\hat{\pi}_i}\right)$$

and apply the expansion

$$\log(1 + x) = x - x^2/2 + x^3/3 - \cdots \quad \text{for } |x| < 1.$$

We identify x with $(p_i - \hat{\pi}_i)/\hat{\pi}_i$, which converges in probability to 0 when the model holds. For large n,

$$\begin{aligned} G^2 &= 2n\sum_i \left[\hat{\pi}_i + (p_i - \hat{\pi}_i)\right]\left[\frac{p_i - \hat{\pi}_i}{\hat{\pi}_i} - \left(\frac{1}{2}\right)\frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i^2} + \cdots\right] \\ &= 2n\sum_i \left[(p_i - \hat{\pi}_i) - \left(\frac{1}{2}\right)\frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i} + \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i} + O_p(p_i - \hat{\pi}_i)^3\right] \\ &= n\sum_i \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i} + 2nO_p(n^{-3/2}) = X^2 + O_p(n^{-1/2}) = X^2 + o_p(1), \end{aligned}$$

since $\Sigma(p_i - \hat{\pi}_i) = 0$ and $(p_i - \hat{\pi}_i) = (p_i - \pi_i) - (\hat{\pi}_i - \pi_i)$, both of which are $O_p(n^{-1/2})$. Thus, when the model holds, the difference between X^2 and G^2 converges in probability to 0. As a consequence, G^2, like X^2, has an asymptotic chi-squared distribution with df $= N - q - 1$.

The parameter value that maximizes the likelihood is the one that minimizes G^2. To show this, we let

$$G^2(\boldsymbol{\pi}; \mathbf{p}) = 2n\sum p_i \log(p_i/\pi_i).$$

The kernel of the multinomial log likelihood is

$$\begin{aligned} L(\boldsymbol{\theta}) &= n\sum p_i \log \pi_i(\boldsymbol{\theta}) \\ &= -n\sum p_i \log\frac{p_i}{\pi_i(\boldsymbol{\theta})} + n\sum p_i \log p_i \\ &= -\left(\frac{1}{2}\right)G^2(\boldsymbol{\pi}(\boldsymbol{\theta}); \mathbf{p}) + n\sum p_i \log p_i. \end{aligned}$$

The second term in the last expression does not depend on $\boldsymbol{\theta}$, so maximizing $L(\boldsymbol{\theta})$ is equivalent to minimizing G^2 with respect to $\boldsymbol{\theta}$.

A fundamental result for G^2 concerns comparisons of nested models. Suppose that model M_0 is a special case of model M_1. Let q_0 and q_1 denote the numbers of parameters in the two models. Let $\{\hat{\pi}_{0i}\}$ and $\{\hat{\pi}_{1i}\}$ denote ML estimators of cell probabilities for the two models. Then

$$G^2(M_0) - G^2(M_1) = 2n \sum p_i \log(\hat{\pi}_{1i}/\hat{\pi}_{0i})$$

has the form of -2(log likelihood ratio) for testing that M_0 holds against the alternative that M_1 holds. Theory for likelihood-ratio tests suggests that when the simpler model holds, the asymptotic distribution of $G^2(M_0) - G^2(M_1)$ is chi-squared with $q_1 - q_2$ degrees of freedom. For details, see Bishop et al. (1975, pp. 525–526), Haberman (1974a, p. 108), and Rao (1973, pp. 418–419). The statistic $X^2(M_0|M_1)$ defined in (9.4) is a quadratic approximation for the G^2 difference. Haberman (1977a) noted that these tests can perform well even for large, sparse tables, as long as $q_1 - q_0$ is small compared to the sample size and no expected frequency has larger order of magnitude than the others.

14.3.5 Asymptotic Noncentral Distributions

Results in this chapter assume that a certain parametric model holds. In practice, any unsaturated model almost surely does not hold perfectly, so one might question the scope of these results. This is not problematic if we regard models merely as convenient approximations for reality. For instance, the ML estimator $\hat{\boldsymbol{\theta}}$ converges to a value $\boldsymbol{\theta}_0$ that describes the best fit of the chosen model to reality. In this sense, inferences for $\boldsymbol{\theta}$ give us information about a useful approximation for reality. Similarly, model-based inferences about cell probabilities are inconsistent for the true probabilities when the model does not hold; nevertheless, those inferences are consistent for describing a useful smoothing of reality.

For goodness-of-fit statistics, a relevant distinction exists between limiting behavior when the model holds and when it does not hold. When the model holds, we've seen X^2 and G^2 have a limiting chi-squared distribution, and the difference between them disappears as n increases. When the model does not hold, X^2 and G^2 tend to grow unboundedly as n increases, and $|X^2 - G^2|$ need not go to zero. One method for obtaining proper limiting distributions considers a sequence of situations $\boldsymbol{\pi}_n$ for which the lack of fit diminishes as n increases. Specifically, the model is $\boldsymbol{\pi} = \mathbf{f}(\boldsymbol{\theta})$, but in reality

$$\boldsymbol{\pi}_n = \mathbf{f}(\boldsymbol{\theta}) + \boldsymbol{\delta}/\sqrt{n}\,. \tag{14.25}$$

The best fit of the model to the population has ith probability equal to $f_i(\boldsymbol{\theta})$, but the true value differs from that by $\delta_i/\sqrt{n}$.

For this representation, Mitra (1958) showed that the Pearson X^2 has a limiting noncentral chi-squared distribution, with df $= N - q - 1$ and non-

centrality parameter

$$\lambda = n \sum_{i=1}^{n} \frac{\left[\pi_{ni} - f_i(\boldsymbol{\theta})\right]^2}{f_i(\boldsymbol{\theta})}.$$

This has the form of X^2, with the sample values p_i and $\hat{\pi}_i$ replaced by population values π_{ni} and $f_i(\boldsymbol{\theta})$. Similarly, the noncentrality of the likelihood-ratio statistic has the form of G^2, with the same substitution. Haberman (1974a, pp. 109–112) showed that under certain conditions G^2 and X^2 have the same limiting distribution; that is, their noncentrality values converge to a common value as $n \to \infty$.

Representation (14.25) means that for large n, the noncentral chi-squared approximation is valid when the model is just barely incorrect. In practice, it is often reasonable to adopt (14.25) for fixed, finite n to approximate the distribution of X^2, even though (14.25) would not be plausible as we obtain more data. The alternative representation

$$\boldsymbol{\pi} = \mathbf{f}(\boldsymbol{\theta}) + \boldsymbol{\delta} \tag{14.26}$$

in which $\boldsymbol{\pi}$ differs from $\mathbf{f}(\boldsymbol{\theta})$ by a *fixed* amount as $n \to \infty$ may seem more natural. In fact, this is more appropriate than (14.25) for proving the test to be consistent (i.e., for convergence to 1 of the probability of rejecting the hypothesis that the model holds). For (14.26), however, the noncentrality parameter λ grows unboundedly as $n \to \infty$, and a proper limiting distribution does not result for X^2 and G^2.

When the model holds, $\boldsymbol{\delta} = \mathbf{0}$ in either representation (14.25) or (14.26). That is, $\mathbf{f}(\boldsymbol{\theta}) = \boldsymbol{\pi}(\boldsymbol{\theta})$, $\lambda = 0$, and the results in Sections 14.3.3 and 14.3.4 apply.

14.4 ASYMPTOTIC DISTRIBUTIONS FOR LOGIT / LOGLINEAR MODELS

For loglinear models, formulas in Section 8.6 for the asymptotic covariance matrices of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\pi}}$ are special cases of ones derived in Section 14.2. We present these for the multinomial form of the models, which relates directly to that section. Then we discuss the connection to Poisson loglinear models.

To constrain probabilities to sum to 1, we express loglinear models for multinomial sampling as

$$\boldsymbol{\pi} = \exp(\mathbf{X}\boldsymbol{\theta}) / \left[\mathbf{1}'\exp(\mathbf{X}\boldsymbol{\theta})\right] \tag{14.27}$$

where $\mathbf{X}$ is a model matrix and $\mathbf{1}' = (1, \ldots, 1)$. Letting $\mathbf{x}_i$ denote row i of $\mathbf{X}$,

$$\pi_i = \pi_i(\boldsymbol{\theta}) = \frac{\exp(\mathbf{x}_i\boldsymbol{\theta})}{\sum_k \exp(\mathbf{x}_k\boldsymbol{\theta})}.$$

14.4.1 Asymptotic Covariance Matrices

A model affects covariance matrices through the Jacobian. Since

$$\frac{\partial \pi_i}{\partial \theta_j} = \frac{\left[\sum_k \exp(\mathbf{x}_k\boldsymbol{\theta})\right]\left[\exp(\mathbf{x}_i\boldsymbol{\theta})\right]x_{ij} - \left[\exp(\mathbf{x}_i\boldsymbol{\theta})\right]\left[\sum_k x_{kj}\exp(\mathbf{x}_k\boldsymbol{\theta})\right]}{\left[\sum_k \exp(\mathbf{x}_k\boldsymbol{\theta})\right]^2}$$

$$= \pi_i x_{ij} - \pi_i \sum_k x_{kj}\pi_k ,$$

the matrix of these elements has the form

$$\partial \boldsymbol{\pi}/\partial \boldsymbol{\theta} = \left[\mathbf{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'\right]\mathbf{X}.$$

Using this with (14.14) and (14.16), the information matrix at $\boldsymbol{\theta}_0$ is

$$\begin{aligned}\mathbf{A}'\mathbf{A} &= (\partial \boldsymbol{\pi}/\partial \boldsymbol{\theta}_0)'\mathbf{diag}(\boldsymbol{\pi}_0)^{-1}(\partial \boldsymbol{\pi}/\partial \boldsymbol{\theta}_0)\\ &= \mathbf{X}'\left[\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0'\right]'\mathbf{diag}(\boldsymbol{\pi}_0)^{-1}\left[\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0'\right]\mathbf{X}\\ &= \mathbf{X}'\left[\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0'\right]\mathbf{X}.\end{aligned}$$

Thus, for multinomial loglinear models, $\hat{\boldsymbol{\theta}}$ is asymptotically normally distributed with estimated covariance matrix

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\theta}}) = \left\{\mathbf{X}'\left[\mathbf{diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}'\right]\mathbf{X}\right\}^{-1}/n. \tag{14.28}$$

Similarly, from (14.23) the estimated asymptotic covariance matrix of $\hat{\boldsymbol{\pi}}$ is

$$\begin{aligned}\widehat{\mathrm{cov}}(\hat{\boldsymbol{\pi}}) = \left[\mathbf{diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}'\right]&\mathbf{X}\left\{\mathbf{X}'\left[\mathbf{diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}'\right]\mathbf{X}\right\}^{-1}\\ &\times\mathbf{X}'\left[\mathbf{diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}}\hat{\boldsymbol{\pi}}'\right]/n.\end{aligned}$$

From (14.23), the Pearson residuals $\mathbf{e}$ are asymptotically normal with

$$\begin{aligned}\text{asymp. cov}(\mathbf{e}) &= \mathbf{I} - \boldsymbol{\pi}_0^{1/2}\left(\boldsymbol{\pi}_0^{1/2}\right)' - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\\ &= \mathbf{I} - \boldsymbol{\pi}_0^{1/2}\left(\boldsymbol{\pi}_0^{1/2}\right)' - \mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}\left[\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0'\right]\mathbf{X}\\ &\quad\times\left\{\mathbf{X}'\left[\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0'\right]\mathbf{X}\right\}^{-1}\mathbf{X}'\\ &\quad\times\left[\mathbf{diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0'\right]\mathbf{diag}(\boldsymbol{\pi}_0)^{-1/2}.\end{aligned}$$

14.4.2 Connection with Poisson Loglinear Models

This book expressed loglinear models in terms of Poisson expected cell frequencies $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)'$, using formulas of the form

$$\log \boldsymbol{\mu} = \mathbf{X}_a\boldsymbol{\theta}_a. \tag{14.29}$$

The model matrix $\mathbf{X}_a$ and parameter vector $\boldsymbol{\theta}_a$ in this formula are slightly different from $\mathbf{X}$ and $\boldsymbol{\theta}$ in multinomial model (14.27). The Poisson expression (14.29) does not have constraints on $\boldsymbol{\mu}$. For multinomial model (14.27), $\sum_i \mu_i = n$ is fixed, and $\boldsymbol{\pi} = \boldsymbol{\mu}/n$ satisfies

$$\begin{aligned}\log \boldsymbol{\mu} &= \log n\boldsymbol{\pi} = \mathbf{X}\boldsymbol{\theta} + \left[\log n - \log(\mathbf{1}'\exp(\mathbf{X}\boldsymbol{\theta}))\right]\mathbf{1} \\ &= \mathbf{X}\boldsymbol{\theta} + \mathbf{1}\mu\end{aligned}$$

where $\mu = \log n - \log(\mathbf{1}'\exp(\mathbf{X}\boldsymbol{\theta}))]$. In other words, multinomial model (14.27) implies Poisson model (14.29) with

$$\mathbf{X}_a = [\mathbf{1}\colon \mathbf{X}] \quad \text{and} \quad \boldsymbol{\theta}_a = (\mu, \boldsymbol{\theta}')'.$$

The columns of $\mathbf{X}$ in the multinomial representation must be linearly independent of $\mathbf{1}$; that is, the parameter μ, which relates to the total sample size, does not appear in $\boldsymbol{\theta}$. The dimension of $\boldsymbol{\theta}$ is 1 less than the number of parameters reported in this text for Poisson loglinear models. For instance, for the saturated model, $\boldsymbol{\theta}$ has $N - 1$ elements for the multinomial representation, reflecting the sole constraint on $\boldsymbol{\pi}$ of $\sum \pi_i = 1$.

NOTES

Section 14.1: Delta Method

14.1. For detailed discussion of large-sample theory including the delta method, see Bishop et al. (1975, Chap. 14) and Sen and Singer (1993).

14.2. In applying the delta method to a function g of an asymptotically normal random vector $\mathbf{T}_n$, suppose that the first-order, ..., $(a-1)$st-order differentials of the function are zero at $\boldsymbol{\theta}$, but the ath-order differential is nonzero. A generalization of the delta method implies that $n^{a/2}[g(\mathbf{T}_n) - g(\boldsymbol{\theta})]$ has limiting distribution involving products of order a of components of a normal random vector. When $a = 2$, the limiting distribution is a quadratic form in a multivariate normal vector, which often relates to a chi-squared distribution; in the univariate case, it is $\sigma^2(g''(\theta))/2$ times a χ_1^2 variable (Casella and Berger 2001, p. 244).

Resampling methods such as the jackknife and the bootstrap are alternative tools for estimating standard errors and obtaining confidence intervals. They can be helpful when use of the delta method is questionable—for instance, for small samples, highly sparse data, or complex sampling designs. For details, see Davison and Hinkley (1997), Fay (1985), Parr and Tolley (1982), and Simonoff (1986).

Section 14.3: Asymptotic Distributions of Residuals and Goodness-of-Fit Statistics

14.3. If Y is Poisson with $E(Y) = \mu$, then for large μ the delta method implies $Y^{1/2}$ is approximately normal with standard deviation $\frac{1}{2}$. This motivates an alternative goodness-of-fit statistic, the *Freeman–Tukey statistic*, $\text{FT} = 4\Sigma(\sqrt{y_i} - \sqrt{\hat{\mu}_i})^2$. When the model holds, $\text{FT} - X^2$ is also $o_p(1)$ as $n \to \infty$. See Bishop et al. (1975, p. 514) for details.

Results of this chapter do not apply when the number of cells N grows as $n \to \infty$, or when different expected frequencies grow at different rates. Haberman (1988) showed the consistency of X^2 breaks down with non-standard asymptotics.

14.4. Drost et al. (1989) showed noncentral approximations using other sequences of alternatives than the local and fixed ones (14.25) and (14.26).

PROBLEMS

14.1 Explain why:

a. If $c > 0$, $n^{-c} = o(1)$ as $n \to \infty$.

b. If $c \neq 0$, cz_n has the same order as z_n; that is, $o(cz_n)$ is equivalent to $o(z_n)$ and $O(cz_n)$ is equivalent to $O(z_n)$.

c. $o(y_n)o(z_n) = o(y_n z_n)$, $O(y_n)O(z_n) = O(y_n z_n)$, $o(y_n)O(z_n) = o(y_n z_n)$.

14.2 If X^2 has an asymptotic chi-squared distribution with fixed df as $n \to \infty$, then explain why $X^2/n = o_p(1)$.

14.3 **a.** Use Tchebychev's inequality to show that if $E(X_n) = \mu_n$ and $\text{var}(X_n) = \sigma_n^2 < \infty$, then $(X_n - \mu_n) = O_p(\sigma_n)$.

b. Suppose that $Y_1, \ldots, Y_n$ are independent with $E(Y_i) = \mu$ and $\text{var}(Y_i) = \sigma^2$ for $i = 1, \ldots, n$. Let $\bar{Y}_n = (\Sigma_i Y_i)/n$. Apply part (a) to show that $\bar{Y}_n - \mu = O_p(n^{-1/2})$.

14.4 Let Y be a Poisson random variable with mean μ.

a. For a constant $c > 0$, show that

$$E[\log(Y + c)] = \log \mu + \left(c - \tfrac{1}{2}\right)/\mu + O(\mu^{-2})$$

(*Hint:* Note that $\log(Y + c) = \log \mu + \log[1 + (Y + c - \mu)/\mu]$.)

b. Cell counts in a 2×2 table are independent Poisson random variables. Use part (a) to argue that to reduce bias in estimating the log odds ratio, a sensible estimator is the sample log odds ratio after adding $\frac{1}{2}$ to each cell.

14.5 Let p denote the sample proportion for n independent Bernoulli trials. Find the asymptotic distribution of the estimator $[p(1 - p)]^{1/2}$ of the standard deviation. What happens when $\pi = 0.5$?

14.6 Suppose that T_n has a Poisson distribution with mean $\lambda = n\mu$, for fixed $\mu > 0$. For large n, show that the distribution of $\log T_n$ is approximately normal with mean $\log(\lambda)$ and variance λ^{-1}. [*Hint:* By

the central limit theorem, T_n/n is approximately $N(\mu, \mu/n)$ for large n.]

14.7 **a.** Refer to Problem 14.6. If T_n is Poisson, show $\sqrt{T_n}$ has asymptotic variance $\frac{1}{4}$.

b. For a binomial sample with n trials and sample proportion p, show the asymptotic variance of $\sin^{-1}(\sqrt{p})$ is $1/4n$. [This transformation and the one in part (a) are *variance stabilizing*, producing variates with asymptotic variances that are the same for all values of the parameter. Traditionally, these transformations were employed to make ordinary least squares applicable to count data. See Cochran 1940 for discussion and ML analyses.]

14.8 For a multinomial $(n, \{\pi_i\})$ distribution, show the correlation between p_i and p_j is $-[\pi_i\pi_j/(1-\pi_i)(1-\pi_j)]^{1/2}$. What does this equal when $\pi_i = 1 - \pi_j$ and $\pi_k = 0$ for $k \neq i, j$?

14.9 An animal population has N species, with population proportion π_i of species i. *Simpson's index of ecological diversity* (Simpson 1949) is $I(\boldsymbol{\pi}) = 1 - \Sigma\pi_i^2$. [Rao (1982) surveyed diversity measures.]

a. Two animals are randomly chosen from the population, with replacement. Show $I(\boldsymbol{\pi})$ is the probability they are different species.

b. For proportions $\mathbf{p}$ for a random sample, show that the estimated asymptotic standard error of $I(\mathbf{p})$ is

$$2\left\{\left[\sum_i p_i^3 - \left(\sum_i p_i^2\right)^2\right]\Big/ n\right\}^{1/2}.$$

14.10 Let $\{Y_i\}$ be independent Poisson random variables. Show by the delta method that the estimated asymptotic variance of $\Sigma a_i \log(Y_i)$ is $\Sigma a_i^2/y_i$. [This formula applies to ML estimators of parameters for the saturated loglinear model, which are contrasts of $\{\log(y_i)\}$. Formula (14.9) yields the asymptotic covariance structure of such estimators; see Lee (1977).]

14.11 Assuming two independent binomial samples, derive the asymptotic standard error of the log relative risk (Section 3.1.4).

14.12 Refer to Problem 3.27. The sample size may need to be quite large for the sampling distribution of $\hat{\gamma}$ to be approximately normal, especially if $|\gamma|$ is large. The Fisher-type transform $\hat{\xi} = \frac{1}{2}\log[(1+\hat{\gamma})/(1-\hat{\gamma})]$ (Agresti 1984, pp. 166–167, 177; O'Gorman and Woolson 1988) converges more quickly to normality.

a. Show that the asymptotic variance of $\hat{\xi}$ equals the asymptotic variance of $\hat{\gamma}$ multiplied by $(1 - \gamma^2)^{-2}$.

b. Explain how to construct a confidence interval for ξ and use it to obtain one for γ.

c. Show that $\hat{\xi} = \frac{1}{2}\log(C/D)$. For 2×2 tables, show that this is half the log odds ratio.

14.13 Let $\phi^2(\mathbf{T}) = \sum_i (T_i - \pi_{i0})^2/\pi_{i0}$. Then $\phi^2(\mathbf{p}) = X^2/n$, where X^2 is the Pearson statistic (1.15) for testing H_0: $\pi_i = \pi_{i0}$, $i = 1, \ldots, N$, and $n\phi^2(\boldsymbol{\pi})$ is the noncentrality for that test when $\boldsymbol{\pi}$ is the true value. Under H_0, why does the delta method not yield an asymptotic normal distribution for $\phi^2(\mathbf{p})$? (See Note 14.2.)

14.14 In an $I \times J$ contingency table, let θ_{ij} denote local odds ratio (2.10), and let $\hat{\theta}_{ij}$ denote its sample value.

a. Show that asymp. $\text{cov}(\sqrt{n}\log\hat{\theta}_{ij}, \sqrt{n}\log\hat{\theta}_{i+1,j}) = -[\pi^{-1}_{i+1,j} + \pi^{-1}_{i+1,j+1}]$.

b. Show that asymp. $\text{cov}(\sqrt{n}\log\hat{\theta}_{ij}, \sqrt{n}\log\hat{\theta}_{i+1,j+1}) = \pi^{-1}_{i+1,j+1}$.

c. When $\hat{\theta}_{ij}$ and $\hat{\theta}_{hk}$ use mutually exclusive sets of cells, show that asymp. $\text{cov}(\sqrt{n}\log\hat{\theta}_{ij}, \sqrt{n}\log\hat{\theta}_{hk}) = 0$.

d. State the asymptotic distribution of $\log\hat{\theta}_{ij}$.

14.15 For loglinear model (XY, XZ, YZ), ML estimates of $\{\mu_{ijk}\}$ and hence the X^2 and G^2 statistics are not direct. Alternative approaches may yield direct analyses. For $2 \times 2 \times 2$ tables, find a statistic for testing the hypothesis of no three-factor interaction, using the delta method with the asymptotic normality of $\log\hat{\theta}_{111}$, where

$$\hat{\theta}_{111} = \frac{p_{111}p_{221}/p_{121}p_{211}}{p_{112}p_{222}/p_{122}\ p_{212}}.$$

14.16 Refer to Section 14.2.2, with $\boldsymbol{\Sigma} = \mathbf{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$ the covariance matrix of $\sqrt{n}\,(p_1, \ldots, p_{N-1})'$. Let

$$Z = \begin{cases} c_i & \text{with probability } \pi_i, \quad i = 1, \ldots, N-1 \\ 0 & \text{with probability } \pi_N \end{cases}$$

and let $\mathbf{c} = (c_1, \ldots, c_{N-1})'$.

a. Show that $E(Z) = \mathbf{c}'\boldsymbol{\pi}$, $E(Z^2) = \mathbf{c}'\mathbf{diag}(\boldsymbol{\pi})\mathbf{c}$, and $\text{var}(Z) = \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}$.

b. Suppose that at least one $c_i \neq 0$, and all $\pi_i > 0$. Show $\text{var}(Z) > 0$, and deduce that $\boldsymbol{\Sigma}$ is positive definite.

c. If $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)'$, so $\boldsymbol{\Sigma}$ is $N \times N$, prove that $\boldsymbol{\Sigma}$ is not positive definite.

14.17 Consider the model for a 2×2 table, $\pi_{11} = \theta^2$, $\pi_{12} = \pi_{21} = \theta(1 - \theta)$, $\pi_{22} = (1 - \theta)^2$, where θ is unknown (Problems 3.31 and 10.34).

a. Find the matrix $\mathbf{A}$ in (14.14) for this model.

b. Use $\mathbf{A}$ to obtain the asymptotic variance of $\hat{\theta}$. (As a check, it is simple to find it directly using the inverse of $-E\partial^2 L/\partial\theta^2$, where L is the log likelihood.) For which θ value is the variance maximized? What is the distribution of $\hat{\theta}$ if $\theta = 0$ or $\theta = 1$?

c. Find the asymptotic covariance matrix of $\sqrt{n}\,\hat{\boldsymbol{\pi}}$.

d. Find df for testing fit using X^2.

14.18 Refer to the model for the calf data in Section 1.5.6. Obtain the asymptotic variance of $\hat{\pi}$.

14.19 Justify the use of *estimated* asymptotic covariance matrices. For instance, for large samples, why is $\hat{\mathbf{A}}'\hat{\mathbf{A}}$ close to $\mathbf{A}'\mathbf{A}$?

14.20 Cell counts $\{Y_i\}$ are independent Poisson random variables, with $\mu_i = E(Y_i)$. Consider the Poisson loglinear model

$$\log \boldsymbol{\mu} = \mathbf{X}_a \boldsymbol{\theta}_a, \quad \text{where } \boldsymbol{\mu} = (\mu_1, \ldots, \mu_N).$$

Using arguments similar to those in Section 14.2, show that the large-sample covariance matrix of $\hat{\boldsymbol{\theta}}_a$ can be estimated by $[\mathbf{X}_a'\mathbf{diag}(\hat{\boldsymbol{\mu}})\mathbf{X}_a]^{-1}$, where $\hat{\boldsymbol{\mu}}$ is the ML estimator of $\boldsymbol{\mu}$.

14.21 For a given set of parameter constraints, show that weak identifiability conditions hold for the independence loglinear model for a two-way table; that is, when two values for $\boldsymbol{\theta}$ give the same $\boldsymbol{\pi}$, those parameter vectors must be identical.

14.22 Use the delta method, with derivatives (14.22), to derive the asymptotic covariance matrix in (14.23) for residuals. Show that this matrix is idempotent.

14.23 In some situations, X^2 and G^2 take very similar values. Explain the joint influence on this event of (**a**) whether the model holds, (**b**) whether the sample size n is large, and (**c**) whether the number of cells N is large.

14.24 Show $\mathbf{X}$ and $\boldsymbol{\theta}$ in multinomial representation (14.27) for the independence model for an $I \times J$ table. By contrast, show $\mathbf{X}_a$ for the corresponding Poisson loglinear model (14.29).

14.25 Using (14.18) and (14.28), derive the asymptotic $\widehat{\text{cov}}(\hat{\boldsymbol{\pi}})$ for a multinomial loglinear model.

14.26 Consider the ML estimator $\hat{\pi}_{ij} = p_{i+}p_{+j}$ of π_{ij} for the independence model, when that model does not hold. Show that $E(p_{i+}p_{+j}) = \pi_{i+}\pi_{+j}(n-1)/n + \pi_{ij}/n$. To what does $\hat{\pi}_{ij}$ converge as n increases?

14.27 Let ζ denote a generic measure of association. For K independent multinomial samples of sizes $\{n_k\}$, suppose that $\sqrt{n_k}(\hat{\zeta}_k - \zeta_k) \xrightarrow{d} N(0, \sigma_k^2)$ as $n_k \to \infty$. A summary measure is

$$\bar{\zeta} = \frac{\Sigma_k\left(n_k/\hat{\sigma}_k^2\right)\hat{\zeta}_k}{\Sigma_k\left(n_k/\hat{\sigma}_k^2\right)}.$$

a. Show that $\Sigma_k z_k^2 = V + [\,\bar{\zeta}^{\,2}/\hat{\sigma}^{\,2}(\bar{\zeta})]$, where

$$V = \sum_k \frac{n_k\left(\hat{\zeta}_k - \bar{\zeta}\right)^2}{\hat{\sigma}_k^2}, \quad z_k = \frac{n_k^{1/2}\hat{\zeta}_k}{\hat{\sigma}_k}, \quad \hat{\sigma}^{\,2}\left(\bar{\zeta}\right) = \left(\sum_k \frac{n_k}{\hat{\sigma}_k^2}\right)^{-1}.$$

b. Suppose that $n \to \infty$ with $n_k/n \to \rho_k > 0$, $k = 1, \ldots, K$. State the asymptotic chi-squared distribution for each component in the partitioning in part (a). Indicate the hypothesis that each tests.

Categorical Data Analysis, Second Edition. Alan Agresti

ISBN: 0-471-36093-7

CHAPTER 15

Alternative Estimation Theory for Parametric Models

In this book we have used the maximum likelihood (ML) approach to inference. This is by far the most common approach for categorical data analysis. Other paradigms have been used, however. In this chapter we discuss some of them. These methods have similar asymptotic properties as maximum likelihood, so the large-sample theory of Chapter 14 applies also to them.

In Section 15.1 we discuss weighted least squares for fitting models for categorical data. This and related quasi-likelihood methods introduced in Sections 4.7 and 11.4 are sometimes simpler to apply than ML.

The Bayesian paradigm is increasingly popular as computations become easier to implement. A full discussion of modern developments with this approach is beyond our scope, but in Section 15.2 we present Bayesian methods of estimating cell probabilities in a contingency table. Four other methods of estimation for categorical data are described in the final section.

15.1 WEIGHTED LEAST SQUARES FOR CATEGORICAL DATA

Weighted least squares (WLS) is an extension of ordinary least squares that permits responses to be correlated and to have nonconstant variance. Familiarity with the WLS method is useful because:

1. WLS computations have a standard form that is simple to apply for a wide variety of models.
2. Algorithms for calculating ML estimates often consist of iterative use of WLS. An example is the Fisher scoring method for generalized linear models (Section 4.6.3).
3. When the model holds, WLS and ML estimators are asymptotically equivalent, both falling in the class of best asymptotically normal (BAN)

estimators. For large samples, the estimators are approximately normally distributed around the parameter value, and the ratio of their variances converges to 1.

Grizzle, Starmer, and Koch (1969) popularized WLS for categorical data analyses. In honor of them, WLS for such analyses is often called the *GSK method*. This section summarizes the ingredients of this approach.

15.1.1 Notation and Preliminaries for WLS Approach

For a response variable Y with J categories, consider multinomial samples of sizes $n_1, \ldots, n_I$ at I levels of an explanatory variable or combinations of levels of several explanatory variables. Let $\boldsymbol{\pi} = (\boldsymbol{\pi}_1', \ldots, \boldsymbol{\pi}_I')'$, where

$$\boldsymbol{\pi}_i = (\pi_{1|i}, \pi_{2|i}, \ldots, \pi_{J|i})' \quad \text{with} \quad \sum_j \pi_{j|i} = 1$$

denotes the conditional distribution of Y at level i. Let $\mathbf{p}$ denote corresponding sample proportions, with $\mathbf{V}$ their $IJ \times IJ$ covariance matrix. When the I samples are independent,

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & & & \mathbf{0} \\ & \mathbf{V}_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \mathbf{V}_I \end{bmatrix}$$

From Section 14.1.4, the covariance matrix of $\sqrt{n_i}\,\mathbf{p}_i$ is

$$n_i\mathbf{V}_i = \begin{bmatrix} \pi_{1|i}(1-\pi_{1|i}) & -\pi_{1|i}\pi_{2|i} & \cdots & -\pi_{1|i}\pi_{J|i} \\ -\pi_{2|i}\pi_{1|i} & \pi_{2|i}(1-\pi_{2|i}) & \cdots & -\pi_{2|i}\pi_{J|i} \\ \vdots & \vdots & & \vdots \\ -\pi_{J|i}\pi_{1|i} & -\pi_{J|i}\pi_{2|i} & \cdots & \pi_{J|i}(1-\pi_{J|i}) \end{bmatrix}.$$

Each set of proportions has $(J-1)$ linearly independent elements.

Let $\mathbf{F}$ be a vector of $u \le I(J-1)$ response functions

$$\mathbf{F}(\boldsymbol{\pi}) = [F_1(\boldsymbol{\pi}), \ldots, F_u(\boldsymbol{\pi})]'.$$

The WLS approach applies to linear models for $\mathbf{F}$ of form

$$\mathbf{F}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}, \tag{15.1}$$

where $\boldsymbol{\beta}$ is a $q \times 1$ vector of parameters and $\mathbf{X}$ is a $u \times q$ model matrix of known constants having rank q. From Section 8.5.4, loglinear and logit response functions are special cases of $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{C}\log(\mathbf{A}\boldsymbol{\pi})$ for certain matrices $\mathbf{C}$ and $\mathbf{A}$.

Let $\mathbf{F}(\mathbf{p})$ denote the sample response functions. We assume that $\mathbf{F}$ has continuous second-order partial derivatives in an open region containing $\boldsymbol{\pi}$. This assumption enables the delta method to determine the large-sample normal distribution for $\mathbf{F}(\mathbf{p})$. The asymptotic covariance matrix of $\mathbf{F}(\mathbf{p})$ depends on the $u \times IJ$ matrix

$$\mathbf{Q} = \frac{\partial F_k(\boldsymbol{\pi})}{\partial \pi_{j|i}}$$

for $k = 1, \ldots, u$ and all IJ combinations (i, j). Linear response models have response functions of form $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{A}\boldsymbol{\pi}$ for a matrix of known constants $\mathbf{A}$, in which case $\mathbf{Q} = \mathbf{A}$. For the generalized loglinear model $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{C}\log(\mathbf{A}\boldsymbol{\pi})$ (recall Sections 8.5.4 and 11.2.5), $\mathbf{Q} = \mathbf{C}[\text{diag}(\mathbf{A}\boldsymbol{\pi})]^{-1}\mathbf{A}$. [See Magnus and Neudecker 1988 for matrix differential calculus.] By the multivariate delta method (Section 14.1.5), the asymptotic covariance matrix of $\mathbf{F}(\mathbf{p})$ is

$$\mathbf{V}_F = \mathbf{QVQ}'.$$

Let $\hat{\mathbf{V}}_F$ denote the sample version of $\mathbf{V}_F$, substituting sample proportions in $\mathbf{Q}$ and $\mathbf{V}$. For subsequent formulas, this matrix must be nonsingular.

15.1.2 Inference Using the WLS Approach to Model Fitting

For the general model (15.1), the WLS estimate of $\boldsymbol{\beta}$ is

$$\mathbf{b} = \left(\mathbf{X}'\hat{\mathbf{V}}_{\mathbf{F}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{F}(\mathbf{p}).$$

This is the $\boldsymbol{\beta}$ value that minimizes the quadratic form

$$[\mathbf{F}(\mathbf{p}) - \mathbf{X}\boldsymbol{\beta}]'\hat{\mathbf{V}}_F^{-1}[\mathbf{F}(\mathbf{p}) - \mathbf{X}\boldsymbol{\beta}].$$

The ordinary least squares estimate, for uncorrelated responses with constant variance, results when $\hat{\mathbf{V}}_F$ is a constant multiple of the identity matrix.

The WLS estimator has an asymptotic multivariate normal distribution, with estimated covariance matrix

$$\widehat{\text{cov}}(\mathbf{b}) = \left(\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X}\right)^{-1}.$$

The normal distribution improves as the sample size increases and $\mathbf{F}(\mathbf{p})$ is more nearly normally distributed.

The estimate $\mathbf{b}$ yields predicted values $\hat{\mathbf{F}} = \mathbf{Xb}$ for the response functions. Since they satisfy the model, these predicted values are smoother than the sample response functions $\mathbf{F(p)}$. When the model holds, $\hat{\mathbf{F}}$ is asymptotically better than $\mathbf{F(p)}$ as an estimator of $\mathbf{F}(\boldsymbol{\pi})$ (Section 14.2.2). The estimated covariance matrix of the predicted values is

$$\hat{\mathbf{V}}_{\hat{F}} = \mathbf{X}\left(\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'.$$

The test of model goodness of fit uses the residual term

$$W = [\mathbf{F(p)} - \mathbf{Xb}]'\hat{\mathbf{V}}_F^{-1}[\mathbf{F(p)} - \mathbf{Xb}] = \mathbf{F(p)}'\hat{\mathbf{V}}_F^{-1}\mathbf{F(p)} - \mathbf{b}'\left(\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X}\right)\mathbf{b},$$

which compares the sample response functions with their model predicted values. Under H_0: $\mathbf{F}(\boldsymbol{\pi}) - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}$ that the model holds, W is asymptotically chi-squared with df $= u - q$, the difference between the number of response functions and the number of model parameters.

One can more closely check the model fit by studying the residuals, $\mathbf{F(p)} - \hat{\mathbf{F}}$. They are orthogonal to the fit $\hat{\mathbf{F}}$, so

$$\text{cov}[\mathbf{F(p)}] = \text{cov}\left\{\left[\mathbf{F(p)} - \hat{\mathbf{F}}\right] + \hat{\mathbf{F}}\right\} = \text{cov}\left[\mathbf{F(p)} - \hat{\mathbf{F}}\right] + \text{cov}(\hat{\mathbf{F}}).$$

Thus, the estimated covariance matrix of the residuals equals

$$\text{cov}[\mathbf{F(p)}] - \text{cov}(\hat{\mathbf{F}}) = \hat{\mathbf{V}}_F - \hat{\mathbf{V}}_{\hat{F}} = \hat{\mathbf{V}}_F - \mathbf{X}\left(\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'.$$

Dividing the residuals by their standard errors yields standardized residuals having large-sample standard normal distributions.

Hypotheses about contrasts and other effects of explanatory variables have form H_0: $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, where $\mathbf{C}$ is a known $c \times q$ matrix with $c \leq q$, having rank c. The estimator $\mathbf{Cb}$ of $\mathbf{C}\boldsymbol{\beta}$ is asymptotically normal with mean $\mathbf{0}$ under H_0 and with covariance matrix estimated by $\mathbf{C}(\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X})^{-1}\mathbf{C}'$. The Wald statistic

$$W_C = \mathbf{b}'\mathbf{C}'\left[\mathbf{C}\left(\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X}\right)\mathbf{C}'\right]^{-1}\mathbf{Cb} \tag{15.2}$$

has an approximate chi-squared null distribution with df $= c$. This statistic also equals the difference between residual chi-squared statistics for the reduced model implied by H_0 and the full model. For the special case H_0: $\beta_i = 0$, $W_C = b_i^2/\text{var}(b_i)$ has df $= 1$.

15.1.3 Scope of WLS versus ML Estimation

The WLS approach requires estimating the multinomial covariance matrix of sample responses at each setting of the explanatory variables. It is inapplicable when explanatory variables are continuous, since there may be only one

observation at each such setting. WLS also becomes less appropriate as the number of explanatory variables increases, since few observations may occur at each of the many combinations of settings. By contrast, in principle, continuous explanatory variables or many explanatory settings are not problematic to ML.

When a certain model holds, with large cell expected frequencies ML and WLS give similar results. Both estimators are in the class of best asymptotically normal estimators. However, practical considerations often favor ML estimation. For example, zero cell counts often adversely affect the WLS approach. The sample response functions may then be ill-defined or have a singular estimated covariance matrix.

WLS shares with quasi-likelihood the feature that inferential results depend only on specifying a model for the mean responses and specifying a variance function and covariance structure (here, based on the multinomial). It does not use the likelihood function for the complete distribution. Thus, inference uses Wald methods.

Historically, an advantage of the WLS approach was computational simplicity. This is not relevant now that software is available for ML analyses and for extensions of WLS (e.g., quasi-likelihood methods such as GEE) that do not have some of its disadvantages. Thus, WLS is now used much less frequently than it was about 25 years ago. Nonetheless, it has close connections with more sophisticated methods. Some algorithms for calculating ML estimates iteratively use WLS. Also, Miller et al. (1993) showed that under certain conditions the solution of the first iteration in the GEE fitting process gives the WLS estimate. This equivalence uses initial estimates based directly on sample values and assumes a saturated association structure that allows a separate correlation parameter for each pair of response categories and each pair of observations in a cluster. In this sense, GEE is an iterated form of WLS. Moreover, in this case, the covariance matrix for the estimates is the same in both approaches.

15.2 BAYESIAN INFERENCE FOR CATEGORICAL DATA

Methodology using the Bayesian paradigm has advanced tremendously in the past decade. New computational methods make it easier to evaluate posterior distributions for model parameters. Nonetheless, Bayesian inference is not as fully developed or commonly used for categorical data analysis as in many other areas of statistics. For multiway contingency table analysis, partly this is because of the plethora of parameters for multinomial models, often necessitating substantial prior specification. Bayesian theory and methods are beyond the scope of this book. We present only relatively elementary problems in which the Bayesian approach applies quite naturally and is sometimes more appealing than ML. We then briefly summarize more complex developments.

The first applications of Bayesian methods to contingency tables involved smoothing cell counts to improve estimation of cell probabilities (e.g., Good 1965). The sample proportions are ordinary ML estimators for the saturated model. When data are sparse, these can have undesirable features. Large sparse tables often contain many sampling zeros, for which 0.0 is unappealing as a probability estimate. In addition, Stein's results for estimating multivariate normal means suggest that lower total mean-squared error occurs with Bayes estimators that shrink the sample proportions toward some average value (Efron and Morris 1975).

In considering Bayesian estimators, we cannot hope to find one that is uniformly better than ML. For instance, suppose that a true cell probability $\pi_i = 0$. Then the sample proportion $p_i = 0$ with probability 1, and the sample proportion is better than any other estimator. Because parameter values exist for which the sample proportion is optimal, no other estimator is uniformly better over the entire parameter space. Here the criterion of comparison is the expected value of a *loss function* that measures distance between the estimator and the parameter, such as squared error. In decision-theoretic terms the sample proportion is an *admissible* estimator, for standard loss functions (Johnson 1971). In this sense, the sample mean for the multinomial or multivariate binomial differs from the sample mean for the multivariate normal, which is inadmissible (dominated by Bayes estimators) when the dimension of the mean vector is at least three (Ferguson 1967, p. 170). Meeden et al. (1998) gave related results for decomposable loglinear models.

Another approach for estimating cell probabilities fits an unsaturated model. Often, though, there is no particular model expected to describe the table well. For $I \times J$ cross-classifications of nominal variables, for instance, the independence model rarely fits well. When unsaturated models approximate the true relationship poorly, model-based estimators also have undesirable properties. Although they smooth the data, the smoothing is too severe for large samples. The model-based estimators are inconsistent, converging to values that may be far from the true cell probabilities as n increases.

A Bayesian approach to estimating cell probabilities compromises between sample proportions and model-based estimators. A model still provides part of the smoothing mechanism, with the Bayes estimators shrinking the sample proportions toward a set of proportions satisfying the model.

15.2.1 Bayesian Estimation of Binomial Parameter

We illustrate basic ideas with Bayesian inference for a binomial parameter. Let y denote a bin(n, π) variate. Since π falls between 0 and 1, a natural prior density for π is the beta [(13.8) in Section 13.3.1] for some choice of $\alpha > 0$ and $\beta > 0$. This satisfies $E(\pi) = \alpha/(\alpha + \beta)$.

In Bayesian inference the posterior density of a parameter, given the data, is proportional to the product of the prior density with the likelihood

function. Here, the beta prior depends on π through $\pi^{\alpha-1}(1-\pi)^{\beta-1}$, and the binomial likelihood has kernel depending on π through $\pi^{y}(1-\pi)^{n-y}$. Thus, the posterior density $h(\pi|y)$ of π is proportional to

$$h(\pi|y) \propto \left[\pi^{y}(1-\pi)^{n-y}\right]\left[\pi^{\alpha-1}(1-\pi)^{\beta-1}\right] = \pi^{y+\alpha-1}(1-\pi)^{n-y+\beta-1},$$

for $0 \leq \pi \leq 1$. The beta is the conjugate prior distribution. The posterior density is also beta, with parameters $\alpha^* = y + \alpha$ and $\beta^* = n - y + \beta$.

The mean of the posterior distribution is a Bayesian estimator of a parameter. This is optimal when a squared-error loss function $(T-\pi)^2$ describes the consequence of estimating π by an estimator T (Ferguson 1967, p. 46). The mean of the beta posterior distribution for π is

$$\begin{aligned} E(\pi|y) &= \alpha^*/(\alpha^* + \beta^*) = (y+\alpha)/(n+\alpha+\beta) \\ &= w(y/n) + (1-w)[\alpha/(\alpha+\beta)], \end{aligned}$$

where $w = n/(n+\alpha+\beta)$. This is a weighted average of the sample proportion $p = y/n$ and the mean of the prior distribution. For fixed (α, β), the weight given the sample increases as n increases. The standard deviation of the posterior distribution describes the accuracy of this estimator. This equals the square root of

$$\text{var}(\pi|y) = \alpha^*\beta^*/(\alpha^* + \beta^*)^2(\alpha^* + \beta^* + 1).$$

For large n the standard deviation is roughly $\sqrt{p(1-p)/n}$, the ordinary standard error for the ML estimator $\hat{\pi} = p$.

The Bayes estimator requires selecting parameters (α, β) for the prior distribution. Complete ignorance about π might suggest a uniform prior distribution. This is the beta distribution with $\alpha = \beta = 1$. The posterior distribution then has the same shape as the binomial likelihood function. The Bayes estimator is then

$$E(\pi|y) = (y+1)/(n+2).$$

This shrinks the sample proportion slightly toward $\frac{1}{2}$.

Alternatively, a popular prior with Bayesians is the *Jeffreys prior*. This is proportional to the square root of the determinant of the Fisher information matrix for the parameters of interest, for a single observation. With a single parameter θ, this is $[E(\partial^2 \log f(y|\theta)/\partial\theta^2)]^{1/2}$. In the binomial case with $\theta = \pi$ and $n = 1$, this equals $[\pi(1-\pi)]^{-1/2}$ and the prior is beta with $\alpha = \beta = .5$. Brown et al. (2001) showed that the posterior generated by this prior yields a confidence interval for π with good performance. It approximates the Clopper–Pearson interval with the mid-P adjustment (Sections

1.4.4 and 1.4.5). For a test of H_0: $\pi \geq \frac{1}{2}$ against H_a: $\pi < \frac{1}{2}$, a Bayesian P-value is the posterior probability that $\pi \geq \frac{1}{2}$. Routledge (1994) showed that with the Jeffreys prior, this posterior probability approximately equals the one-sided mid-P-value for the ordinary binomial test.

15.2.2 Dirichlet Prior and Posterior for Multinomial Parameters

These ideas generalize from the binomial to the multinomial (Good 1965). Suppose that cell counts $(n_1, \ldots, n_N)$ have a multinomial distribution with $n = \sum n_i$ and parameters $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)'$. The multinomial likelihood is proportional to

$$\prod_{i=1}^{N} \pi_i^{n_i}.$$

For a prior distribution over potential $\boldsymbol{\pi}$ values, the multivariate generalization of the beta is the *Dirichlet density*

$$g(\boldsymbol{\pi}) = \frac{\Gamma(\sum \beta_i)}{\prod_i \Gamma(\beta_i)} \prod_{i=1}^{N} \pi_i^{\beta_i - 1} \quad \text{for } 0 \leq \pi_i \leq 1 \text{ all } i, \qquad \sum_i \pi_i = 1,$$

where $\{\beta_i > 0\}$. For it, $E(\pi_i) = \beta_i/(\sum_j \beta_j)$.

The posterior density is also Dirichlet, with parameters $\{n_i + \beta_i\}$. The Bayes estimator of π_i is

$$E(\pi_i | n_1, \ldots, n_N) = (n_i + \beta_i) \Big/ \Big(n + \sum_j \beta_j\Big). \tag{15.3}$$

Let $K = \sum \beta_j$ and $\gamma_i = E(\pi_i) = \beta_i/K$. The $\{\gamma_i\}$ are prior guesses for the cell probabilities. Bayes estimator (15.3) equals the weighted average

$$[n/(n + K)]\, p_i + [K/(n + K)]\, \gamma_i. \tag{15.4}$$

From (15.3) the Bayes estimator is a sample proportion when the prior information corresponds to $\sum_j \beta_j$ trials with β_i outcomes of type i, $i = 1, \ldots, N$. This interpretation may provide guidance for choosing $\{\beta_i\}$. The Jeffreys prior sets all $\beta_i = 0.5$. Good referred to K as a *flattening constant*, since with identical $\{\beta_i\}$ (15.4) shrinks each sample proportion toward the uniform value $\gamma_i = 1/N$. Greater flattening occurs as K increases, for fixed n. Hierarchical models treat $\{\beta_i\}$ as unknown and specify a second-stage prior for them (e.g., Albert and Gupta 1982).

Bayes estimators combine good characteristics of sample proportions and model-based estimators. Like sample proportions and unlike model-based

estimators, they are consistent even when the model does not hold. Unless the model holds, the weight given the sample proportion increases to 1.0 as the sample size increases. Like model-based estimators and unlike sample proportions, the Bayes estimators smooth the data. The resulting estimates, although slightly biased, usually have smaller total mean-squared error than the sample proportions.

15.2.3 Development of Bayesian Methods for Categorical Data

We now summarize the development of Bayesian methods for categorical data since Good's (1965) work on smoothing multinomial proportions. Leonard and Hsu (1994) provided a more detailed review. We begin with methods for two-way contingency tables.

For 2×2 tables, Altham (1969) gave a Bayesian analysis comparing parameters for two independent binomial samples. She tested H_0:$\pi_1 \leq \pi_2$ against $\pi_1 > \pi_2$ using independent beta(α_i, β_i) priors for π_1 and π_2. Altham showed that the P-value that is the posterior probability that $\pi_1 \leq \pi_2$ can equal the one-sided P-value for Fisher's exact test. This happens when one uses improper prior distributions $(\alpha_1, \beta_1) = (1, 0)$ and $(\alpha_2, \beta_2) = (0, 1)$. These represent prior belief favoring the null hypothesis, in effect penalizing against concluding that $\pi_1 > \pi_2$. That is, Fisher's exact test corresponds to a conservative prior distribution.

If $\alpha_i = \beta_i = \gamma$, $i = 1, 2$, with $0 \leq \gamma \leq 1$, Altham showed that the Bayesian P-value is smaller than the Fisher P-value. The difference between the two is no greater than the null probability of the observed data. Use of Jeffreys priors with $\alpha_i = \beta_i = 0.5$ provides a type of continuity correction to Fisher's exact test in much the way the mid-P-value does for the frequentist approach. Howard (1998) showed that with these priors the posterior probability that $\pi_1 \leq \pi_2$ approximates the one-sided P-value for the large-sample z test using pooled variance (i.e., the signed square root of the Pearson statistic; see Problem 3.30) for testing H_0: $\pi_1 = \pi_2$ against H_a: $\pi_1 > \pi_2$. Howard also discussed other priors for 2×2 tables, including ones that treat π_1 and π_2 as dependent.

Altham (1971) showed Bayesian analyses for binomial proportions from matched-pairs data. For a simple model in which the probability of success is the same for each subject at a given occasion, she again showed that the classical exact P-value (Section 10.1.4, using the binomial distribution) is a Bayesian P-value for a prior distribution favoring H_0. For a model similar to (10.8) in which the probability varies by subject but the occasion effect is constant, she showed that the Bayesian evidence against the null is weaker as the number of pairs giving the same response at both occasions increases, for fixed values of the numbers of pairs giving different responses at the two occasions. This differs from the conditional ML result, which does not

depend on such pairs (Section 10.2.3). Ghosh et al. (2000) showed related results.

The Bayesian approaches presented so far focused directly on cell probabilities by using a prior distribution for them. Lindley (1964) did this with $I \times J$ contingency tables. He considered the posterior distribution of contrasts of log probabilities, such as the log odds ratio. An alternative approach (Laird 1978; Leonard 1975) focused on parameters of the saturated loglinear model, using normal priors. This is not a conjugate prior, but normal distributions can approximate the posterior. Using independent normal $N(0, \sigma^2)$ distributions for the association parameters is a way of inducing shrinkage toward the independence model (Laird 1978). A hierarchical approach puts second-stage priors on the parameters of the prior distribution (Leonard 1975).

Historically, a barrier for the Bayesian approach has been the difficulty of calculating the posterior distribution when the prior is not conjugate. This is less problematic with modern ways of approximating posterior distributions by simulating samples from them. These include the importance sampling generalization of Monte Carlo simulation (Zellner and Rossi 1984) and Markov chain Monte Carlo methods such as Gibbs sampling (Gelfand and Smith 1990). Zellner and Rossi used Bayesian methods for logistic regression and Gelfand and Smith considered a class of multinomial models with Dirichlet prior. Zeger and Karim (1991) fitted generalized linear mixed models (GLMMs) essentially using a Bayesian framework with priors for fixed and random effects.

The focus on distributions for random effects in GLMMs in articles such as Zeger and Karim (1991) led to the treatment of parameters in GLMs as random variables with a fully Bayesian approach. Dey et al. (2000) edited a collection of articles that provided Bayesian analyses for GLMs. For instance, in that volume Gelfand and Ghosh surveyed the subject, Albert and Ghosh reviewed item response modeling, Chib modeled correlated binary data, and Chen and Dey modeled correlated ordinal data.

Bayesian methods are used increasingly in applications. For instance, Skene and Wakefield (1990) modeled multicenter binary response studies with a logit model that allows the treatment–response log odds ratio to vary among centers. This gives a Bayesian alternative to the GLMM analysis presented in Section 12.3.4. Daniels and Gatsonis (1999) used multi-level GLMs to analyze geographic and temporal trends with clustered longitudinal binary data. This built on hierarchical modeling ideas introduced by Wong and Mason (1985). An article by Landrum and Normand in Dey et al. (2000) gave a case study using Bayesian ordinal probit and logit models. Chaloner and Larntz (1989) used a Bayesian approach to determining optimal design for experiments using logistic regression. J. Albert has suggested Bayesian models for a variety of categorical data analyses. For instance, Albert (1997) modeled associations in two-way tables and Albert and Chib (1993) studied

binary regression modeling, focusing on the probit case with extensions to ordered multinomial responses.

15.2.4 Data-Dependent Choice of Prior Distribution

With Bayesian analyses, careful prior specification is necessary. The use of an improper prior, such as the uniform prior over the entire or positive real line, sometimes results in improper posteriors. One may not realize this from the output of software for Bayesian fitting. In addition, with simulation methods it may not be obvious when convergence has occurred. Be suspicious if results are dramatically different from ordinary ML frequentist results.

Some dislike the subjectivity of the Bayesian approach inherent in selecting a prior distribution. Instead of choosing particular parameters for a prior distribution, it is increasingly popular to use a hierarchical approach in which those parameters themselves have a second-stage prior distribution. Alternatively, the empirical Bayes approach lets the data suggest parameter values for use in the prior distribution (e.g., Efron and Morris 1975). This approach uses the prior that maximizes the marginal probability of the observed data, integrating out with respect to the prior. Laird (1978) did this for the loglinear model, estimating σ^2 in normal priors for association parameters by finding the value that maximizes an approximation for the marginal distribution of the cell counts, evaluated at the observed data. A disadvantage of empirical Bayes compared to the hierarchical approach is that it does not take into account the source of variability due to substituting estimates for prior parameters.

Fienberg and Holland (1973) proposed analyses for contingency tables with data-dependent priors. For a particular choice of Dirichlet means $\{\gamma_i\}$ for the Bayes estimator (15.4), they showed that the minimum total mean-squared error occurs when

$$K = \left(1 - \sum \pi_i^2\right) \Big/ \left[\sum (\gamma_i - \pi_i)^2\right]. \tag{15.5}$$

The optimal $K = K(\boldsymbol{\gamma}, \boldsymbol{\pi})$ depends on $\boldsymbol{\pi}$, so they used the estimate $K(\boldsymbol{\gamma}, \mathbf{p})$ of K in which the sample proportion $\mathbf{p}$ replaces $\boldsymbol{\pi}$. As $\mathbf{p}$ falls closer to the prior guess $\boldsymbol{\gamma}$, $K(\boldsymbol{\gamma}, \mathbf{p})$ increases and the prior guess receives more weight in the posterior estimate. They selected the prior pattern $\{\gamma_i\}$ for the cell probabilities based on the fit of a simple model. For two-way tables, they used the independence fit $\{\gamma_{ij} = p_{i+}p_{+j}\}$. The Bayes estimator then shrinks sample proportions toward that fit.

As in other inference, Bayesian modeling should normally account for any ordering in the response categories. For instance, in the method just mentioned for smoothing contingency tables, one could shrink toward an ordinal model.

15.3 OTHER METHODS OF ESTIMATION

In this final section we describe some alternative estimation methods for categorical data. Consider estimation of $\boldsymbol{\pi}$ or $\boldsymbol{\theta}$, assuming a model $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$. Let $\tilde{\boldsymbol{\theta}}$ denote a generic estimator of $\boldsymbol{\theta}$, for which $\tilde{\boldsymbol{\pi}} = \boldsymbol{\pi}(\tilde{\boldsymbol{\theta}})$ estimates $\boldsymbol{\pi}$. The ML estimator $\hat{\boldsymbol{\theta}}$ maximizes the likelihood. It also minimizes the deviance statistic G^2 comparing observed and fitted proportions (Section 14.3.4).

15.3.2 Minimum Chi-Squared Estimators

Other estimators minimize other measures of distance between $\boldsymbol{\pi}(\boldsymbol{\theta})$ and $\mathbf{p}$. The value $\tilde{\boldsymbol{\theta}}$ that minimizes the Pearson statistic

$$X^2[\boldsymbol{\pi}(\boldsymbol{\theta}), \mathbf{p}] = n \sum \frac{[p_i - \pi_i(\boldsymbol{\theta})]^2}{\pi_i(\boldsymbol{\theta})}$$

is called the *minimum chi-squared* estimate. It is simpler to calculate the estimate that minimizes the modified statistic

$$X^2_{\text{mod}}[\boldsymbol{\pi}(\boldsymbol{\theta}), \mathbf{p}] = n \sum \frac{[p_i - \pi_i(\boldsymbol{\theta})]^2}{p_i} \tag{15.6}$$

that replaces the denominator by the sample proportion. This is called the *minimum modified chi-squared* estimate. It is the solution for $\boldsymbol{\theta}$ to the equations

$$\sum_i \frac{\pi_i(\boldsymbol{\theta})}{p_i} \left(\frac{\partial \pi_i(\boldsymbol{\theta})}{\partial \theta_j} \right) = 0, \qquad j = 1, \ldots, q.$$

Neyman (1949) introduced minimum modified chi-squared estimators. He showed that they and minimum chi-squared estimators are best asymptotically normal (BAN) estimators. When the model holds, they are asymptotically (as $n \to \infty$) equivalent to ML estimators. Under the model, different estimation methods (ML, WLS, minimum chi-squared, etc.) yield nearly identical estimates of parameters when n is large. This happens partly because the estimators are consistent, converging in probability to $\boldsymbol{\theta}$ as n increases. When the model does not hold, estimates for different methods can be quite different, even when n is large. The estimators converge to values for which the model gives the best approximation to reality, and this approximation is different when best is defined in terms of minimizing G^2 rather than minimizing X^2 or some other measure.

For any n, minimum modified chi-squared estimates are sometimes identical to WLS estimates. The connection refers to an alternative way of

specifying a model, using a set of *constraint equations* for π,

$$\{g_j(\pi_1, \ldots, \pi_N) = 0\}.$$

For instance, for an $I \times J$ table, the $(I-1)(J-1)$ constraint equations

$$\log \pi_{ij} - \log \pi_{i,j+1} - \log \pi_{i+1,j} + \log \pi_{i+1,j+1} = 0$$

specify the model of independence. The number of constraint equations equals the residual df for the model.

Neyman (1949) noted that minimum modified chi-squared estimates result from minimizing

$$\sum_{i=1}^{N} \frac{(p_i - \pi_i)^2}{p_i} + \sum_{j=1}^{N-q} \lambda_j g_j(\pi_1, \ldots, \pi_N)$$

with respect to $\boldsymbol{\pi}$, where the $\{\lambda_j\}$ are Lagrange multipliers. When the constraint equations are linear in $\boldsymbol{\pi}$, the resulting estimating equations are linear. Then Bhapkar (1966) showed that these estimators are identical to WLS estimators. The statistic (15.6) then equals the WLS residual statistic (Section 15.1.2) for testing model fit.

Usually, however, constraint equations are nonlinear in $\boldsymbol{\pi}$, such as for the independence model. The WLS estimator is then the minimum modified chi-squared estimator based on a linearized version of the constraints,

$$g_j(\mathbf{p}) + (\boldsymbol{\pi} - \mathbf{p})' \partial g_j(\boldsymbol{\pi}) / \partial \boldsymbol{\pi} = 0,$$

with differential vector evaluated at $\mathbf{p}$.

Berkson (1944, 1955, 1980) was a strong advocate of minimum chi-squared methods. For logistic regression, his *minimum logit chi-squared* estimators minimized a weighted sum of squares between sample logits and linear predictions. Mantel (1985) criticized such methods, noting that their consistency requires group sizes to grow large, whereas ML (or conditional ML, when there are many nuisance parameters) is consistent however information goes to the limit (see also Problem 15.14).

15.3.2 Minimum Discrimination Information

Kullback (1959) formulated estimation by *minimum discrimination information* (MDI). The discrimination information for two probability vectors $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$ is

$$I(\boldsymbol{\pi}; \boldsymbol{\gamma}) = \sum_{i=1}^{N} \pi_i \log(\pi_i / \gamma_i). \tag{15.7}$$

This directed measure of distance between $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$ is nonnegative, equaling 0 only when $\boldsymbol{\pi} = \boldsymbol{\gamma}$. Gokhale and Kullback (1978) studied MDI estimates that minimize $I(\boldsymbol{\pi}; \boldsymbol{\gamma})$, subject to model constraints, using $\boldsymbol{\gamma} = \mathbf{p}$ for some problems and $\boldsymbol{\gamma}$ with $\gamma_1 = \gamma_2 = \cdots = \gamma_N = 1/N$ for others. Good (1963) conducted related work in the area of *maximum entropy*.

In some cases with $\{\gamma_i = 1/N\}$, the MDI estimator is identical to the ML estimator (Simon 1973). With $\boldsymbol{\gamma} = \mathbf{p}$ it is not ML, but it has similar asymptotic properties, being best asymptotically normal (BAN). Then Gokhale and Kullback recommended testing goodness of fit using twice the minimized value of $I(\boldsymbol{\pi}; \mathbf{p})$. This statistic reverses the roles of $\mathbf{p}$ and $\boldsymbol{\pi}$ relative to G^2, much as X^2_{mod} in (15.6) reverses their roles relative to X^2. Both statistics fall in the class of power divergence statistics (Cressie and Read 1984; see also Problem 3.34) and have similar asymptotic properties. More generally, one could choose any member of the power divergence statistics and define estimates to be the values minimizing it. Under regularity conditions, they are all BAN.

15.3.3 Kernel Smoothing

Kernel estimation is a smoothing method that estimates a probability density or mass function without assuming a parametric distribution. Let $\mathbf{K}$ denote a matrix containing nonnegative elements and having column sums equal to 1. Kernel estimates of cell probabilities in a contingency table have form

$$\tilde{\boldsymbol{\pi}} = \mathbf{K}\mathbf{p}. \tag{15.8}$$

For unordered multinomials with N categories, Aitchison and Aitken (1976) used

$$\begin{aligned} k_{ij} &= \lambda, \quad i = j \\ &= (1 - \lambda)/(N - 1), \quad i \neq j \end{aligned}$$

for $(1/N) \leq \lambda \leq 1$. The resulting kernel estimator of $\boldsymbol{\pi}$ has form

$$(1 - \alpha)\mathbf{p} + \alpha(\mathbf{1}/N), \tag{15.9}$$

where $\alpha = N(1 - \lambda)/(N - 1)$. This estimator shrinks the sample proportion toward $(1/N, \ldots, 1/N)$. As λ decreases from 1 to $1/N$, the smoothing parameter α increases from 0 to 1. Brown and Rundell (1985) proved that when no $\pi_i = 1$, $\lambda < 1$ exists such that the total mean squared error is smaller for this kernel estimator than for the sample proportions. Results for other shrinkage estimators applied to multivariate means suggest that the improvement for the kernel estimator can be large when n is small and the true cell probabilities are roughly equal.

Brown and Rundell generalized kernel smoothing for multiway contingency tables that may contain both nominal and ordinal variables. For a

T-way table, let $\mathbf{L}_k$ be a stochastic matrix (i.e., row and column sums equal to 1) with elements

$$\ell_{k,ij} = \begin{cases} \lambda_k, & i = j \\ d_k(i,j)(1-\lambda_k), & i \neq j, \end{cases}$$

$k = 1, \ldots, T$. They let $\mathbf{K}$ in (15.8) be the Kronecker product

$$\mathbf{K} = \mathbf{L}_1 \otimes \cdots \otimes \mathbf{L}_T.$$

When variable k is ordinal, shrinkage alone is not enough, and it helps to borrow information from nearby cells. Then $d_k(i,j)$ is chosen to be smaller for greater distances between categories i and j. If variable k is nominal, the natural choice is $d_k(i,j) = 1/(I_k - 1)$, where I_k is the number of categories for variable k. For fixed $\{\lambda_k\}$, collapsing the smoothed table gives the same result as smoothing the corresponding collapsing of the original table. With $\{\lambda_k = \lambda,\ k = 1, \ldots, T\}$, Brown and Rundell described ways of finding λ to minimize an unbiased estimate of the total mean squared error.

Dong and Simonoff (1995) and Simonoff (1986) described other approaches for ordered categories. Most such kernels yield probability estimates of the form

$$\tilde{\pi}_i = (1-\alpha)p_i + \alpha \times \text{smoother}_i,$$

where the smoothing is designed to work well when true probabilities in nearby cells are similar.

15.3.4 Penalized Likelihood

Good and Gaskins (1971) introduced the *penalized likelihood* method for density estimation. For log likelihood $L(\boldsymbol{\pi})$, the estimator maximizes

$$L^*(\boldsymbol{\pi}) = L(\boldsymbol{\pi}) - \alpha(\boldsymbol{\pi})$$

where $\alpha(\cdot)$ is a function that provides a roughness penalty. That is, $\alpha(\boldsymbol{\pi})$ decreases as elements of $\boldsymbol{\pi}$ are smoother, in some sense. The penalized likelihood estimator has a Bayesian interpretation. With prior density proportional to $\exp[-\alpha(\boldsymbol{\pi})]$, the posterior density is proportional to the penalized likelihood function. Hence, the mode of the posterior distribution equals the penalized likelihood estimator.

Simonoff (1983) applied penalized likelihood to estimating cell probabilities $\boldsymbol{\pi}$. Like Bayesian and kernel methods, it provides estimates that are smoother than the sample proportions. For a single multinomial with ordered categories, Simonoff (1983) used penalty function $\alpha(\boldsymbol{\pi}) = \lambda \sum_{i=1}^{N-1} (\log \pi_i - \log \pi_{i+1})^2$, which encourages adjacent category estimates to be similar. For

two-way contingency tables, Simonoff suggested using $\alpha(\boldsymbol{\pi}) = \lambda\sum_i\sum_j(\log\theta_{ij})^2$ with the local odds ratios. This provides shrinkage toward the independence estimator. One chooses the smoothing parameter λ to minimize an approximation for the mean-squared error of the estimator.

In evaluating smoothing methods such as kernel smoothing and penalized likelihood, it is useful to distinguish between large-sample asymptotics with a fixed number of cells N and sparse-data asymptotics for which N grows with n (recall Section 6.3.4). For the former, these smoothing methods and Bayesian inference behave asymptotically like ordinary ML (i.e., the sample proportions). They have the same rate of convergence to true probabilities. These methods then improve over ML primarily for small samples, where the benefit of "borrowing from the whole" occurs. For sparse-data asymptotics, however, smoothing is particularly beneficial. As the dimensions of a table increase, the number of cells grows exponentially and the "curse of dimensionality" occurs. Accurate estimation becomes more difficult, with estimators converging more slowly to true values. The table then has an increasing proportion of empty cells. Smoothing can be better than ML even asymptotically. For such results, see Fienberg and Holland (1973) for the Dirichlet-based Bayes multinomial estimator and Simonoff (1983) for penalized likelihood with the multinomial. Simonoff showed that consistency can occur with the latter estimator in the sense that $\sup_i|\hat{\pi}_i/\pi_i - 1| \xrightarrow{p} 0$ as n and N grow and the probabilities themselves approach 0.

For surveys of smoothing methods, see Fahrmeir and Tutz (2001, Chap. 5), Lloyd (1999, Chap. 5), and Simonoff (1996, Chap. 6; 1998). As Simonoff noted, all smoothing methods attempt to balance the low bias of undersmoothing with the low variability of oversmoothing. The methods require input from the user about the degree of smoothness, whether it be determined by a prior distribution or some type of smoothing parameter.

In summary, many methods exist for smoothing categorical data. Besides those discussed in this section, there are traditional model-building methods. Some of these, such as generalized additive models (Section 4.8), are also specifically directed toward smoothing. A particular type of smoothing method may seem most natural for a given application. An advantage of the Bayesian approach is that its entire formulation seems less ad hoc than some others.

NOTES

Section 15.1: Weighted Least Squares for Categorical Data

15.1. Applications of WLS include fitting mean response models (Grizzle et al. 1969) and models for marginal distributions (Koch et al. 1977). For general discussion, see Bhapkar and Koch (1968), Imrey et al. (1981), and Koch et al. (1985).

Section 15.2: Bayesian Inference for Categorical Data

15.2. Other literature on Bayesian analyses of categorical responses includes Fienberg et al. (1999), Forster and Smith (1998), Good (1976), Knuiman and Speed (1988), Spiegelhalter and Smith (1982), and Walley (1996).

Section 15.3: Other Methods of Estimation

15.3. For further discussion of minimum chi-squared methods, see Bhapkar (1966), Koch et al. (1985), Neyman (1949), and Rao (1963).

15.4. For the use of minimum discrimination information, see Gokhale and Kullback (1978), Ireland and Kullback (1968a, b), Ireland et al. (1969), and Ku et al. (1971).

15.5. Hall and Titterington (1987) studied rates of convergence for multinomial kernel estimators. They defined one that achieves the optimal rate. Ordinary kernel estimators tend to be biased toward zero at the boundary of a table. Dong and Simonoff (1994) dealt with improving kernel estimates on the boundary of large sparse tables. Kernel methods are also useful for discrete regression modeling. For binary response data, Copas (1983) used one to display in a nonparametric manner the dependence of $P(Y = 1)$ on x.

PROBLEMS

Applications

15.1 Consider the mean response model fitted in Section 7.4.6. Show how to use WLS for this analysis. Identify the number of multinomial samples I, the number of response categories J, the response functions $\mathbf{F}$, the model matrix $\mathbf{X}$, the parameter vector $\boldsymbol{\beta}$, and the estimated covariance matrix $\hat{\mathbf{V}}_F$.

15.2 Use WLS to conduct the longitudinal analysis of depression in Section 11.2.1. Using software (e.g., SAS: PROC CATMOD), obtain WLS estimates and standard errors and compare to the ML results.

15.3 Refer to Problem 15.2. Using these data, describe the differences between (**a**) WLS and ML, and (**b**) WLS and GEE methods for marginal models with multivariate categorical response data.

15.4 Using data from Section 1.4.3, obtain a Bayesian estimate of the proportion of vegetarians. Explain how you chose the prior distribution. Compare results to those with ML.

15.5 Refer to Table 9.8. Consider the model that simultaneously assumes (9.12) as well as linear logit relationships for the marginal effects of age on breathlessness and on wheeze.

a. Specify $\mathbf{C}$, $\mathbf{A}$, and $\mathbf{X}$ for which this model has form $\mathbf{C}\log\mathbf{A}\boldsymbol{\pi} = \mathbf{X}\boldsymbol{\beta}$.

b. Using software, fit the model and interpret estimates.

Theory and Methods

15.6 Consider marginal homogeneity for an $I \times I$ table.

a. Letting $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{A}\boldsymbol{\pi}$, explain how **(i)** $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{0}$, where $\mathbf{A}$ has $I - 1$ rows, and **(ii)** $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{A}$ has $2(I - 1)$ rows and $\boldsymbol{\beta}$ has $I - 1$ elements. In part (ii), show $\mathbf{A}$, $\boldsymbol{\pi}$, $\mathbf{X}$, $\boldsymbol{\beta}$ when $I = 3$.

b. Explain how to use WLS to test marginal homogeneity. [This is Bhapkar's test (10.16).]

15.7 For WLS with $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{C}[\log(\mathbf{A}\boldsymbol{\pi})]$, show that $\mathbf{Q} = \mathbf{C}[\mathbf{diag}(\mathbf{A}\boldsymbol{\pi})]^{-1}\mathbf{A}$.

15.8 With WLS, show that $[\mathbf{F}(\mathbf{p}) - \mathbf{X}\boldsymbol{\beta}]'\hat{\mathbf{V}}_F^{-1}[\mathbf{F}(\mathbf{p}) - \mathbf{X}\boldsymbol{\beta}]$ is minimized by $\boldsymbol{\beta} = (\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}_F^{-1}\mathbf{F}(\mathbf{p})$.

15.9 The response functions $\mathbf{F}(\mathbf{p})$ have asymptotic covariance matrix $\mathbf{V}_F$. Derive the asymptotic covariance matrix of the WLS model parameter estimator $\mathbf{b}$ and the predicted values $\hat{\mathbf{F}} = \mathbf{X}\mathbf{b}$.

15.10 Consider the Bayes estimator of a binomial parameter π using a beta prior distribution.

a. Does any beta prior distribution produce a Bayes estimator that coincides with the ML estimator?

b. Show that the ML estimator is a limit of Bayes estimators, for a certain sequence of beta prior parameter values.

c. Find an improper prior density (one for which its integral is not finite) such that the Bayes estimator coincides with the ML estimator. (In this sense, the ML estimator is a *generalized Bayes estimator*.)

d. For Bayesian inference using loss function $w(\theta)(T - \theta)^2$, the Bayes estimator of θ is the posterior expected value of $\theta w(\theta)$ divided by the posterior expected value of $w(\theta)$ (Ferguson 1967, p. 47). With loss function $(T - \pi)^2/[\pi(1 - \pi)]$, show the ML estimator of π is a Bayes estimator for the uniform prior distribution.

e. The risk function is the expected loss, treated as a function of π. For the loss function in part (d), show the risk function is constant. (Bayes' estimators with constant risk are *minimax*; their maximum risk is no greater than the maximum risk for any other estimator.)

f. Show that the Jeffreys prior for π equals the beta density with $\alpha = \beta = .5$.

15.11 For the Dirichlet prior for multinomial probabilities, show the posterior expected value of π_i is (15.3). Derive the expression for this Bayes estimator as a weighted average of p_i and $E(\pi_i)$.

15.12 For Bayes estimator (15.4), show that the total mean squared error is

$$[K/(n+K)]^2\Big[\sum(\pi_i-\gamma_i)^2\Big]+[n/(n+K)]^2\Big(1-\sum\pi_i^2\Big).$$

Show that (15.5) is the value of K that minimizes this.

15.13 Refer to Problem 15.6. For marginal homogeneity, explain why the minimum modified chi-squared estimates are identical to WLS estimates.

15.14 Let y_i be a bin(n_i, π_i) variate for group i, $i = 1, \ldots, N$, with $\{y_i\}$ independent. Consider the model that $\pi_1 = \cdots = \pi_N$. Denote that common value by π.

a. Show that the ML estimator of π is $p = (\Sigma_i y_i)/(\Sigma_i n_i)$.

b. The minimum chi-squared estimator $\tilde{\pi}$ is the value of π minimizing

$$\sum_{i=1}^{N}\frac{[(y_i/n_i)-\pi]^2}{\pi}+\sum_{i=1}^{N}\frac{[(y_i/n_i)-\pi]^2}{1-\pi}.$$

The second term results from comparing $(1 - y_i/n_i)$ to $(1 - \pi)$, the proportions in the second category. If $n_1 = \cdots = n_N = 1$, show that $\tilde{\pi}$ minimizes $Np(1-\pi)/\pi + N(1-p)\pi/(1-\pi)$. Hence show

$$\tilde{\pi}=p^{1/2}/\Big[p^{1/2}+(1-p)^{1/2}\Big].$$

Note the bias toward $\frac{1}{2}$ in this estimator.

c. Argue that as $N \to \infty$ with all $n_i = 1$, the ML estimator is consistent but the minimum chi-squared estimator is not (Mantel 1985).

15.15 Refer to Problem 15.14. For $N = 2$ groups with n_1 and n_2 independent observations, find the minimum modified chi-squared estimator of π. Compare it to the ML estimator.

15.16 Show that the kernel estimator (15.9) is the same as the Bayes estimator (15.3) for the Dirichlet prior with $\{\beta_i = \alpha n/(1-\alpha)N\}$. Using this result, suggest a way of letting the data determine the value of α in the kernel estimator.

Categorical Data Analysis, Second Edition. Alan Agresti

ISBN: 0-471-36093-7

CHAPTER 16

Historical Tour of Categorical Data Analysis*

This book concludes with an informal historical overview of the evolution of methods for categorical data analysis (CDA). We have seen that categorical scales are pervasive in the social sciences and the biomedical sciences. Not surprisingly, the development of GLMs for categorical responses was fostered by statisticians having ties to the social sciences or to the biomedical sciences.

Only in the last quarter of the twentieth century did these models receive the attention given early in the century to models for continuous data. Regression models for continuous variables evolved out of Francis Galton's breakthroughs in the 1880s. The strong influence of R. A. Fisher, G. Udny Yule, and other statisticians on experimentation in agriculture and biological sciences ensured widespread adoption of regression and ANOVA modeling by the mid-twentieth century. On the other hand, despite influential articles around 1900 by Karl Pearson and Yule on association between categorical variables, models for categorical responses received scant attention until the 1960s.

The beginnings of CDA were often shrouded in controversy. Key figures in the development of statistical science made groundbreaking contributions, but these statisticians were often in heated disagreement with one another.

16.1 PEARSON–YULE ASSOCIATION CONTROVERSY

Much of the early development of methods for CDA took place in England, and it is fitting that we begin our historical tour in London at the beginning of the twentieth century. The year 1900 is an apt starting point, since in that year Karl Pearson introduced his chi-squared statistic (X^2) and G. Udny Yule presented the odds ratio and related measures of association. Before

then most work focused on descriptive aspects for relatively simple measures. For instance, Goodman and Kruskal (1959) noted that the Belgian social statistician Adolphe Quetelet used the relative risk in 1849.

By 1900, Karl Pearson (1857–1936) was already well known in the statistics community. He was head of a statistical laboratory at University College in London. His work the previous decade included developing a family of skewed probability distributions (called *Pearson curves*), obtaining the product-moment estimate of the correlation coefficient and finding its standard error, and extending Galton's work on linear regression. In fact, Pearson was a true renaissance man, writing on a wide variety of topics that included art, religion, philosophy, law, socialism, women's rights, physics, genetics, eugenics, and evolution. Pearson's motivation for developing the chi-squared test included testing whether outcomes on a roulette wheel in Monte Carlo varied randomly, checking the fit to various data sets of normal distributions and Pearson curves, and testing statistical independence in two-way contingency tables.

Much of the literature on CDA early in the twentieth century consisted of vocal debates about appropriate ways to summarize association. Pearson's approach assumed that continuous bivariate distributions underlie two-way contingency tables (Pearson 1904, 1913). He argued in favor of approximating a measure, such as the correlation, for the underlying continuum. In 1904, Pearson introduced the term *contingency* as a "measure of the total deviation of the classification from independent probability," and he introduced measures to describe its extent. The *tetrachoric correlation* is a ML estimate of the correlation for a bivariate normal distribution assumed to underlie counts in 2×2 tables. It is the correlation value ρ in the bivariate normal density that would produce cell probabilities equal to the sample cell proportions when that density is collapsed to a 2×2 table having the same marginal proportions as the observed table. The *mean-square contingency* and the *contingency coefficient* are normalizations of X^2 to the $(0, 1)$ scale. Pearson's contingency coefficient (Problem 3.33) for $I \times J$ tables standardized X^2 to approximate an underlying correlation.

George Udny Yule (1871–1951), a British contemporary of Pearson's, took a different approach. Having completed pioneering work developing multiple regression models and multiple and partial correlation coefficients, Yule turned his attention between 1900 and 1912 to association in contingency tables. He believed that many categorical variables, such as (vaccinated, unvaccinated) and (died, survived), are inherently discrete. Yule defined indices directly using cell counts without assuming an underlying continuum. He popularized the odds ratio θ [which Goodman (2000) noted may first have been proposed by a Hungarian statistician, J. Kőrösy] and a transformation of it to the $[-1, +1]$ scale, $Q = (\theta - 1)/(\theta + 1)$, now called *Yule's Q* (Problem 2.36). Discussing one of Pearson's measures that assumes underlying normality, Yule argued (1912, p. 612) that "at best the normal coefficient can only be said to give us in cases like these a hypothetical correlation between

supposititious variables. The introduction of needless and unverifiable hypotheses does not appear to me a desirable proceeding in scientific work." Yule (1903) also showed the potential discrepancy between marginal and conditional associations in contingency tables, later studied by E. H. Simpson (1951) and now called *Simpson's paradox*.

In the first quarter of the twentieth century, Karl Pearson was the rarely challenged leader of statistical science in Britain. Pearson's strong personality did not take kindly to criticism, and he reacted negatively to Yule's ideas. He argued that Yule's own coefficients were unsuitable. For instance, Pearson claimed that their values were unstable, since different collapsings of $I \times J$ tables to 2×2 tables could produce quite different values of the measures. Pearson and D. Heron (1913) filled more than 150 pages of *Biometrika*, a journal he co-founded and edited, with a scathing reply to Yule's criticism. In a passage critical also of Yule's well-received book *An Introduction to the Theory of Statistics*, they stated "If Mr. Yule's views are accepted, irreparable damage will be done to the growth of modern statistical theory. ... [Yule's Q] has never been and never will be used in any work done under his [Pearson's] supervision. ... We regret having to draw attention to the manner in which Mr. Yule has gone astray at every stage in his treatment of association, but criticism of his methods has been thrust on us not only by Mr. Yule's recent attack, but also by the unthinking praise which has been bestowed on a text-book which at many points can only lead statistical students hopelessly astray." Pearson and Heron attacked Yule's "half-baked notions" and "specious reasoning" and argued that Yule would have to withdraw his ideas "if he wishes to maintain any reputation as a statistician."

In retrospect, Pearson and Yule both had valid points. Some classifications, such as most nominal variables, have no apparent underlying continuous distribution. On the other hand, many applications relate naturally to an underlying continuum, and that fact can motivate models and inference (e.g., Section 7.2.3). Goodman (1981a, b) noted that the ordinal models presented in Sections 9.4.1 and 9.6.1 provide a sort of reconciliation between Yule and Pearson, since Yule's odds ratio characterizes models that fit well when underlying distributions are approximately normal.

Half a century after the Pearson–Yule controversy, Leo Goodman and William Kruskal surveyed the development of association measures for contingency tables and made many contributions of their own. Their 1979 book reprinted four influential articles of theirs from the *Journal of the American Statistical Association* on this topic. Initial development of many measures occurred in the nineteenth century. Their 1959 article contains the following quote from M. H. Doolittle in 1887, which illustrates the lack of precision in early attempts to quantify the meaning of *association* even in 2×2 tables: "Having given the number of instances respectively in which things are both thus and so, in which they are thus but not so, in which they are so but not thus, and in which they are neither thus nor so, it is required to eliminate the general quantitative relativity inhering in the mere thingness

of the things, and to determine the special quantitative relativity subsisting between the thusness and the soness of the things." Goodman (2000) added to the historical survey and proposed a new measure.

16.2 R. A. FISHER'S CONTRIBUTIONS

Pearson's disagreements with Yule were minor compared to his later ones with Ronald A. Fisher (1890–1962). Using a geometric representation, Fisher (1922) introduced *degrees of freedom* to characterize the family of chi-squared distributions. Fisher claimed that for tests of independence in $I \times J$ tables, X^2 has df $= (I-1)(J-1)$. By contrast, Pearson (1900, 1904) had argued that for any application of X^2, the index that Fisher later identified as df equals the number of cells minus 1, or $IJ-1$ for two-way tables. Fisher pointed out, however, that estimating hypothesized cell probabilities using estimated row and column probabilities resulted in an additional $(I-1)+(J-1)$ constraints on the fitted values, thus affecting the distribution of X^2.

Not surprisingly, Pearson (1922) reacted critically to Fisher's suggestion that his df formula was incorrect. He stated: "I hold that such a view [Fisher's] is entirely erroneous, and that the writer has done no service to the science of statistics by giving it broad-cast circulation in the pages of the *Journal of the Royal Statistical Society*. ... I trust my critic will pardon me for comparing him with Don Quixote tilting at the windmill; he must either destroy himself, or the whole theory of probable errors, for they are invariably based on using sample values for those of the sampled population unknown to us." Pearson claimed that using row and column sample proportions to estimate unknown probabilities had negligible effect on large-sample distributions, although he had realized (Pearson 1917) that df must be adjusted when the cell counts have linear constraints. Fisher was unable to get his rebuttal published by the Royal Statistical Society, and he ultimately resigned his membership.

Statisticians soon realized that Fisher was correct, but he maintained much bitterness over this and other dealings with Pearson. In the preface to a later volume of his collected works, he remarked that his 1922 article "had to find its way to publication past critics who, in the first place, could not believe that Pearson's work stood in need of correction, and who, if this had to be admitted, were sure that they themselves had corrected it." Writing about Pearson: he stated: "If peevish intolerance of free opinion in others is a sign of senility, it is one which he had developed at an early age." In Fisher (1926), he was able to dig the knife a bit deeper into the Pearson family using 11,688 2×2 tables randomly generated assuming independence by Karl Pearson's son, E. S. Pearson. Fisher showed that the sample mean of X^2 for these tables was 1.00001, much closer to the 1.0 predicted by his formula for $E(X^2)$ of df $= (I-1)(J-1) = 1$ than Pearson's $IJ-1=3$. His daughter,

Joan Fisher Box (1978), discussed this and other conflicts between Fisher and Pearson. Hald (1998, pp. 652–663), Plackett (1983), and Stigler (1999, Chap. 19) summarized the chi-squared controversy.

Fisher's preeminent reputation among statisticians today accrues mainly from his theoretical work (introducing concepts such as sufficiency, information, and optimal properties of ML estimators) and his methodological contributions to the design of experiments and the analysis of variance. Although not so well known for work in CDA, he made other interesting contributions. Moreover, he made good use of the methods in his applied work. For instance, Fisher was also a famed geneticist. In one article, he used Pearson's goodness-of-fit test to check Mendel's theories of natural inheritance and showed that the fit was *too* good (Section 1.5.3).

Fisher realized the limitations of large-sample methods for laboratory work, and he was at the forefront of advocating specialized small-sample methods. Writing about large-sample methods in the preface to the first edition of his classic text *Statistical Methods for Research Workers*, he stated: "[T]he traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a cannon to shoot a sparrow, but it misses the sparrow! The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data." Fisher was among the first to promote the work by W. S. Gosset (pseudonym "Student") on the t distribution. The fifth edition of *Statistical Methods for Research Workers* (1934) introduced Fisher's exact test for 2×2 contingency tables. In his 1935 book *The Design of Experiments*, Fisher described the tea-tasting experiment (Section 3.5.2) motivated by his experience at an afternoon tea break while employed at Rothamsted Experiment Station.

The mid-1930s finally saw some model building for categorical responses. Chester Bliss (1934, 1935), following up a 1933 report on quantal response methods by J. H. Gaddum, popularized the probit model for applications in toxicology with a binary response. Bliss introduced the term *probit* but used the inverse normal cdf with mean 5 (rather than 0, in order to avoid negative values) and standard deviation 1. In the appendix of Bliss (1935), Fisher (1935b) outlined an algorithm for finding ML estimates of model parameters. That algorithm was a Newton–Raphson type of method using expected information, today commonly called *Fisher scoring* (Section 4.6.2). Stigler (1986, p. 246) and Finney (1971) attributed the first use of inverse normal cdf transformations of proportions to the German physicist Gustav Fechner in his 1860 book *Elemente der Psychophysik*. See Finney (1971) and McCulloch (2000) for other history of the probit method.

The definition for homogeneous association (no interaction) in contingency tables originated in an article by the British statistician Maurice Bartlett (1935) about $2 \times 2 \times 2$ tables. Bartlett showed how to find ML

estimates of cell probabilities satisfying the property of equality of odds ratios between two variables at each level of the third. He attributed the idea to Fisher.

In 1940, Fisher developed canonical correlation methods for contingency tables. He showed how to assign scores to rows and columns of a contingency table to maximize the correlation. His work relates to the later development, particularly in France, of *correspondence analysis* methods (e.g., Benzécri 1973).

R. A. Fisher has had the greatest influence on the practice of modern statistical science. The biography by his daughter (Box 1978) gives a fascinating account of his impressive contributions to statistics and genetics. Fienberg (1980) summarized his contributions to CDA.

16.3 LOGISTIC REGRESSION

Bartlett (1937) used $\log[y/(1-y)]$ in regression and ANOVA to transform observations y that are continuous proportions (Problem 6.33). In a book of statistical tables published in 1938, R. A. Fisher and Frank Yates suggested it as a possible transformation of a binomial parameter for analyzing binary data. In 1944, the physician and statistician Joseph Berkson introduced the term *logit* for this transformation. Berkson showed that the model using the logit fitted similarly to the probit model, and his subsequent work did much to popularize logistic regression. In 1951, Jerome Cornfield, another statistician with strong medical ties, used the odds ratio to approximate relative risks in case–control studies. Dyke and Patterson (1952) apparently first used the logit in models with qualitative predictors.

Sir David R. Cox introduced many statisticians to logistic regression, through his 1958 article and 1970 book, *The Analysis of Binary Data*. About the same time, an article by the Danish statistician and mathematician Georg Rasch sparked an enormous literature on item response models. The most important of these is the logit model with subject and item parameters, now called the *Rasch model* (Section 12.1.4). This work was highly influential in the psychometric community of northern Europe (especially in Denmark, the Netherlands, and Germany) and spurred many generalizations in the educational testing community in the United States.

The extension of logistic regression to multicategory responses received occasional attention before 1970 (e.g., Mantel 1966) but substantial work after about that date. For nominal responses, early work was mainly in the econometrics literature. See Bock (1970), McFadden (1974), Nerlove and Press (1973), and Theil (1969, 1970). In 2000, Daniel McFadden won the Nobel Prize in Economics for his work in the 1970s and 1980s on the discrete-choice model (Section 7.6). For cumulative logit models for ordinal responses, see Bock and Jones (1968), Simon (1974), Snell (1964), Walker and Duncan (1967), and Williams and Grizzle (1972). The cumulative probit case,

based on an underlying normal response, has a longer history; see, for instance, Aitchison and Silvey (1957) and Bock and Jones (1968, Chap. 8). Cumulative logit and probit models received much more attention following publication of McCullagh (1980), which provided a Fisher scoring algorithm for ML fitting of all cumulative link models.

The next major advances with logistic regression dealt with its application to case–control studies (e.g., Breslow 1996; Mantel 1973; Prentice 1976a; Prentice and Pyke 1979; see also Section 5.1.4) and the conditional ML approach to model fitting for those studies and others with numerous nuisance parameters (Breslow et al. 1978, with related work in Breslow 1976, 1982; Breslow and Day 1980; Breslow and Powers 1978; Cox 1970; Farewell 1979; Prentice 1976a; Prentice and Breslow 1978; Zelen 1971; see also Sections 6.7 and 10.2). The conditional approach was later exploited in small-sample exact inference (Hirji et al. 1987; Mehta and Patel 1995; see also Section 6.7).

Nathan Mantel, whose name appears in the preceding two paragraphs, made a variety of interesting contributions to CDA. Although best known for the 1959 Mantel–Haenszel test and related odds ratio estimator, he also discussed trend tests (1963), multinomial logit and loglinear modeling (1966), logistic regression for case–control data (1973), the number of contingency tables having fixed margins (Gail and Mantel 1977), the analysis of square contingency tables (Mantel and Byar 1978), and problems with minimum chi-squared and Wald tests (1985, 1987a).

More recently, attention has focused on fitting logistic models to correlated responses for clustered data. One strand of this is marginal modeling of longitudinal data (Diggle et al. 2002; Liang and Zeger 1986; Liang et al. 1992). Much of this literature focuses on quasi-likelihood methods such as generalized estimating equations (GEE). Another strand is generalized linear mixed models (e.g., Breslow and Clayton 1993).

Perhaps the most far-reaching contribution of the past half century has been the introduction by British statisticians John Nelder and R. W. M. Wedderburn in 1972 of the concept of *generalized linear models*. This unifies the logistic and probit regression models for binomial data with loglinear models for Poisson data and with long-established regression and ANOVA models for normal-response data. Interestingly, the algorithm they used to fit GLMs is Fisher scoring, which R. A. Fisher introduced in 1935 for ML fitting of probit models. McCulloch (2000) reviewed the journey from probit models to GLMs and their further generalizations such as quasi-likelihood.

16.4 MULTIWAY CONTINGENCY TABLES AND LOGLINEAR MODELS

The quarter century following the end of World War II saw the development of a theoretical underpinning for models for contingency tables. H. Cramér

(1946) derived general expressions for large-sample distributions of parameter estimators. C. R. Rao (1957, 1963) conducted related work.

In 1949, the Berkeley-based statistician Jerzy Neyman, who had already performed fundamental work on hypothesis testing and interval estimation methods with E. S. Pearson, introduced the family of *best asymptotically normal* (BAN) estimators. These have the same optimal large-sample properties as ML estimators. The BAN family includes estimators obtained by minimizing chi-squared-type measures comparing observed proportions to proportions predicted by the model (Section 15.3.1). This type of estimator itself includes some *weighted least squares* (WLS) estimators. The simplicity of their computation, compared to ML estimators, was an important consideration before the advent of modern computing. Neyman's (1949) only mention of Fisher was the suggestion that Fisher did not realize that estimators other than ML could be BAN, stating that "the results ... contradict the assertion of R. A. Fisher, not a very clear one, that 'the maximum likelihood equation may indeed be derived from the conditions that it shall be linear in frequencies, and efficient for all values of θ'." Fisher, of course, returned the compliment: for instance, writing (1956) about proposals for an unconditional test for 2×2 tables, "the Principles of Neyman and Pearson's 'Theory of Testing Hypotheses' are liable to mislead those who follow them into much wasted effort."

In the early 1950s, William Cochran published work dealing with a variety of important topics in CDA. Scottish-born, Cochran spent most of his career at American universities: Iowa State, North Carolina State, Johns Hopkins, and Harvard. He (1940) modeled Poisson and binomial responses with variance-stabilizing transformations. He (1943) recognized and discussed ways of dealing with overdispersion. He (1950) introduced a generalization (Cochran's Q) of McNemar's test for comparing proportions in several matched samples. His classic 1954 article is a mixture of new methodology and advice for applied statisticians. It gave sample-size guidelines for chi-squared approximations to work well for the X^2 statistic. It also stressed the importance of directing inferences toward narrow (e.g., single-degree-of-freedom) alternatives and partitioning chi-squared statistics into components. One instance of this was Cochran's proposed test of conditional independence in several 2×2 tables, which was closely related to the Mantel and Haenszel (1959) test (Section 6.3.2). Another was a test for a linear trend in proportions across quantitatively defined rows of an $I \times 2$ table (Section 5.3.5). See also Cochran (1955). Fienberg (1984) reviewed Cochran's contributions to CDA.

Bartlett's work on interaction structure in $2 \times 2 \times 2$ contingency tables had relatively little impact for 20 years. Indeed, in presenting methods for partitioning X^2 in $2 \times 2 \times 2$ tables, Lancaster (1951) noted that "Doubtless little use will ever be made of more than a three-dimensional classification." However, in the mid-1950s and early 1960s, Bartlett's work was extended in many ways to multiway tables. See, for instance, Darroch (1962), Good

(1963), Goodman (1964b), Plackett (1962), Roy and Kastenbaum (1956), and Roy and Mitra (1956). These articles as well as influential articles by Martin W. Birch (1963, 1964a, b, 1965) were the genesis of research work on loglinear models between about 1965 and 1975. Birch's work was part of a never-submitted Ph.D. thesis at the University of Glasgow. He showed how to obtain ML estimates of cell probabilities in three-way tables, under various conditions. He showed the equivalence of those ML estimates for Poisson and multinomial sampling. He (and Watson 1959) extended theoretical results of Cramér and Rao on large-sample distributions for contingency table models. Mantel (1966) discussed early results and made the loglinear model formula explicit. A survey article by the French statistician Henri Caussinus (1966), based partly on his Ph.D. thesis, provides a good glimpse of the state-of-the-art of CDA just before this decade of advances. There, Caussinus introduced the quasi-symmetry model for square tables.

Much of the work in the next decades on loglinear and related logit modeling took place at three American universities: the University of Chicago, Harvard University, and the University of North Carolina. At Chicago, Leo Goodman wrote a series of groundbreaking articles, dealing with such topics as partitionings of chi-squared, models for square tables (e.g., quasi-independence), stepwise logit and loglinear model-building procedures, deriving asymptotic variances of ML estimates of loglinear parameters, latent class models, association models, correlation models, and correspondence analysis. For surveys of his early work, see Goodman (1968, an R. A. Fisher memorial lecture, 1970). For later work, see Goodman (1985, 1996, 2000). Goodman also wrote a stream of articles for social science journals that had a substantial impact on popularizing loglinear and logit methods for applications (e.g., Goodman 1969b).

Over the past 50 years, Goodman has been the most prolific contributor to the advancement of CDA methodology. The field owes tremendous gratitude to his steady and impressive body of work. In addition, some of Goodman's students at Chicago also made fundamental contributions. In 1970, Shelby Haberman completed a Ph.D. dissertation (the basis of his 1974a monograph) making substantial theoretical contributions to loglinear modeling. Among topics he considered were residual analyses, existence of ML estimates, loglinear models for ordinal variables, and theoretical results for models (such as the Rasch model) for which the number of parameters grows with the sample size. Clifford Clogg followed in Goodman's steps by having influence in the social sciences and in statistics with his work on association models, demography, models for rates, the census, and various other topics.

Simultaneously with Goodman's work, related research on ML methods for loglinear-logit models occurred at Harvard by students of Frederick Mosteller (such as Stephen Fienberg) and William Cochran. Much of this research was inspired by problems arising in analyzing large, multivariate data sets in the National Halothane Study (Bishop and Mosteller 1969; see also p. 345 of an interview with Lincoln Moses in *Statist. Sci.* **14**, 1999). That

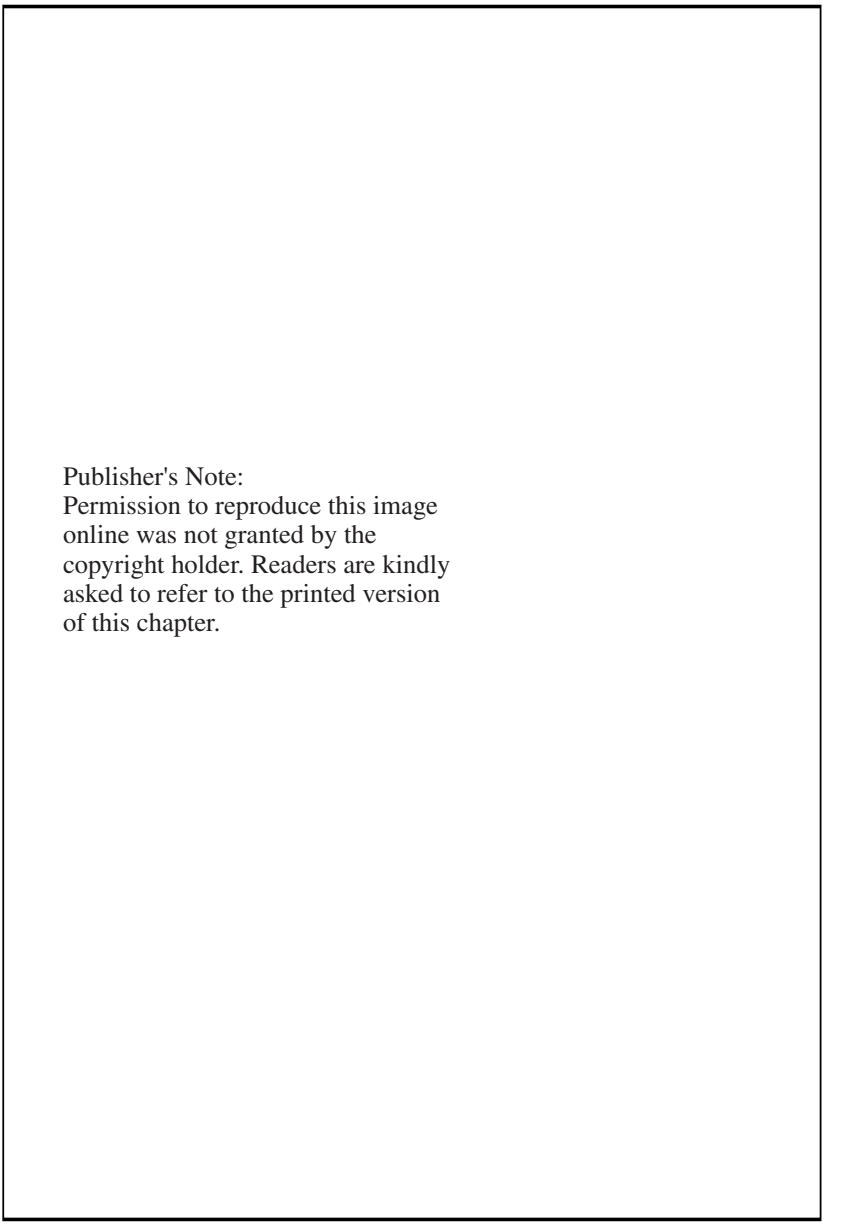


FIGURE 16.1 Four leading figures in the development of categorical data analysis.

study investigated whether halothane was more likely than other anesthetics to cause death due to liver damage. A presidential address by Mosteller (1968) to the American Statistical Association described early uses of loglinear models for smoothing multidimensional discrete data sets. Fienberg and his own students advanced this work further. A landmark book in 1975 by him with Yvonne Bishop and Paul Holland, *Discrete Multivariate Analysis*, was largely responsible for introducing loglinear models to the general statistical community and remains an excellent reference.

Research at North Carolina by Gary Koch and several students and co-workers has been highly influential in the biomedical sciences. Their research developed WLS methods for categorical data models (Section 15.1). The 1969 article by Koch with J. Grizzle and F. Starmer popularized this approach. Koch and colleagues extended it in later articles to an impressive variety of problems, including problems for which ML methods are awkward to use, such as the analysis of repeated categorical measurement data (Koch et al. 1977). In 1966, Vasant Bhapkar showed that the WLS estimator is often identical to Neyman's minimum modified chi-squared estimator.

The early literature on loglinear models treated all classifications as nominal. Haberman (1974b) and Simon (1974) showed how to exploit ordinality of classifications in loglinear models. This work was extended in several articles by Leo Goodman (1979a, 1981a, b, 1983, 1985, 1986). The extensions included association models, which replace ordered scores in loglinear models by parameters (Section 9.5). Goodman (1985, 1986, 1996) also discussed related correlation models and provided a model-based perspective for the closely related correspondence analysis methods.

Certain loglinear models with conditional independence structure provide graphical models for contingency tables. These relate to the association graphs used in Section 9.1. Darroch et al. (1980) was the genesis of much of this work.

16.5 RECENT (AND FUTURE?) DEVELOPMENTS

The most active area of new research in CDA in the past decade has been the modeling of clustered data, such as occur in longitudinal studies and other forms of repeated measurement. A variety of ways now exist of modeling while accounting for the correlation among responses in the same cluster.

As discussed in Chapters 11 and 12, ML estimation is difficult for such models. For complex forms of generalized linear mixed models, for instance, it is a challenge to estimate well regression parameters and variance components. Integrating out the random effect to obtain the likelihood function requires an approximation such as numerical integration. Not surprisingly, various Monte Carlo approaches are applied increasingly here. A promising

approach is a Monte Carlo EM algorithm that uses a Monte Carlo approximation for the E step (Booth and Hobert 1999). The Monte Carlo error can be assessed at each iteration, and one can accurately reproduce the ML estimates with sufficiently many iterations.

The modeling of clustered correlated data is likely to be an active area of research in coming years. The class of generalized linear mixed models is certain to see substantial work and further generalization. One extension is generalized *additive* mixed models. Time-series models for categorical responses have so far received relatively little attention. For all such models with correlated responses, model diagnostics are of vital importance and need development. For longitudinal data, missing data are a common problem. This area currently has much activity.

Another important recent advance is the development of efficient algorithms for exact small-sample methods. With such methods, one can guarantee that the size of a test is no greater than some prespecified level and that the coverage probability for a confidence interval is at least the nominal level. The "exactness" refers only to inference being based on probability distributions that do not depend on unknown parameters. There is no unique way to do this, and certain methods can be highly conservative because of discreteness. Most literature deals with the conditional approach, which eliminates nuisance parameters by conditioning on their sufficient statistics. Hence, the basic idea builds on Fisher's exact test. Conditional methods are versatile, applying to exponential family linear models that use the canonical link function, such as loglinear models for Poisson responses and logit models for binomial responses. Many of the computational advances with the exact conditional approach occurred in a series of articles by Cyrus Mehta, Nitin Patel, and colleagues at Harvard (e.g., Mehta and Patel 1983), using the network algorithm. See surveys by Agresti (1992), Mehta (1994), Mehta and Patel (1995), and the *StatXact* and *LogXact* manuals (Cytel Software, Cambridge, MA, founded by Mehta and Patel).

Although the development of "exact" methods has seen considerable progress, certain analyses are still infeasible and likely to be so for some time because of the exponential increase in computing time as the table size or sample size increases. There are an ever-increasing variety of methods for accurate approximation of exact methods. These include simple Monte Carlo (e.g., Agresti et al. 1979), Monte Carlo with importance sampling (e.g., Booth and Butler 1999; Mehta et al. 1988), Markov chain Monte Carlo (MCMC; Forster et al. 1996), saddlepoint approximations (Pierce and Peters 1992, Strawderman and Wells 1998), and related work on an *approximate* conditioning approach (Pierce and Peters 1999) in which discreteness is not so problematic.

Finally, the development of Bayesian approaches to CDA is an increasingly active area. The multiplicity of parameters complicates Bayesian modeling. For early use of Bayesian estimation of probabilities, see Good (1965) and Lindley (1964). Good's (1965) article apparently evolved from his work

during World War II with Alan Turing at Bletchley Park, England, on breaking Nazi codes. The development of the Bayesian approach for CDA is discussed in Section 15.2.3.

Predicting the future is always dangerous. However, it is likely that much future research will focus on computationally intensive methods such as generalized linear mixed models. Another hot topic, largely outside the realm of traditional modeling, is the development of algorithmic methods for huge data sets with large numbers of variables. Such methods, often referred to as *data mining*, deal with the handling of complex data structures, with a premium on predictive power at the sacrifice of simplicity and interpretability of structure. Important areas of application include genetics, such as the analysis of discrete DNA sequences in the form of very high-dimensional contingency tables, and business applications such as credit scoring and tree-structured methods for predicting future behavior of customers.

Sources for the historical tour in this chapter include Stigler (1986), *Studies in the History of Probability and Statistics*, edited by E. S. Pearson and M. G. Kendall (London: Griffin, 1970), and personal conversations over the years with several statisticians, including Erling Andersen, R. L. Anderson, Henri Caussinus, William Cochran, Sir David Cox, John Darroch, Leo Goodman, Gary Koch, Frederick Mosteller, John Nelder, C. R. Rao, Stephen Stigler, Geoffrey Watson, and Marvin Zelen. To readers who have made it this far, I congratulate your perseverance! To develop a more complete understanding of the historical development of CDA, you may want to study the following chronological list of 25 sources. These convey a sense of how methodology has evolved. Alternatively, look at some early books on this topic, such as A. E. Maxwell's *Analysing Qualitative Data* (New York: Methuen, 1961), R. L. Plackett's *The Analysis of Categorical Data* (London: Griffin, 1974), and the Bishop, Fienberg, and Holland *Discrete Multivariate Analysis* (Cambridge, MA: MIT Press 1975).

Pearson (1900)
Yule (1912)
Fisher (1922)
Bartlett (1935)
Berkson (1944)
Neyman (1949)
Cochran (1954)
Goodman and Kruskal (1954)
Roy and Mitra (1956)
Cox (1958a)
Mantel and Haenszel (1959)
Birch (1963)
Birch (1964b)
Caussinus (1966)
Goodman (1968)
Mosteller (1968)
Grizzle et al. (1969)
Goodman (1970)
Haberman (1974a)
Nelder and Wedderburn (1972)
McFadden (1974)
Goodman (1979a)
McCullagh (1980)
Liang and Zeger (1986)
Breslow and Clayton (1993)

Categorical Data Analysis, Second Edition. Alan Agresti

ISBN: 0-471-36093-7

APPENDIX A

Using Computer Software to Analyze Categorical Data

In this appendix we discuss statistical software for categorical data analysis, with emphasis on SAS. We begin by mentioning major software that can perform the analyses discussed in this book. Then we illustrate, by chapter, SAS code for the analyses. Information about other packages (such as S-Plus, R, SPSS, and Stata), as well as updated information about SAS, is at the Web site (*www.stat.ufl.edu/~aa/cda/cda.html*.) Section A.2 on SAS also lists other software for analyses not currently available in SAS.

A.1 SOFTWARE FOR CATEGORICAL DATA ANALYSIS

A.1.1 SAS

SAS is general-purpose software for a wide variety of statistical analyses. The main procedures (PROCs) for categorical data analyses are FREQ, GENMOD, LOGISTIC, NLMIXED, and CATMOD.

PROC FREQ computes measures of association and their estimated standard errors. It also performs generalized Cochran–Mantel–Haenszel tests of conditional independence, and exact tests of independence in $I \times J$ tables.

PROC GENMOD fits generalized linear models. It fits cumulative link models for ordinal responses. It can perform GEE analyses for marginal models. One can form one's own variance function and allow scale parameters, making it suitable for quasi-likelihood analyses.

PROC LOGISTIC gives ML fitting of binary response models, cumulative link models for ordinal responses, and baseline-category logit models for nominal responses. It incorporates model selection procedures, regression diagnostic options, and exact conditional inference. PROC PROBIT also conducts ML fitting of binary and cumulative link models as well as quantal

response models that permit a strictly positive probability as the linear predictor decreases to $-\infty$.

PROC CATMOD fits baseline-category logit models. It is also useful for WLS fitting of a wide variety of models for categorical data.

PROC NLMIXED fits generalized linear mixed models (GLMMs). It approximates the likelihood using adaptive Gauss–Hermite quadrature.

Other programs run on SAS that are not specifically supported by the SAS Institute. For further details about SAS for categorical data analyses, see the very helpful guide by Stokes et al. (2000). Also useful are SAS publications on logistic regression (Allison 1999) and graphics (Friendly 2000).

A.1.2 Other Software Packages

Most major statistical software has procedures for categorical data analyses. For instance, see SPSS (*SPSS Regression Models 10.0* by M. J. Norusis, SPSS Inc., 1999), Stata (*A Handbook of Statistical Analyses Using Stata*, 2nd ed., by S. Rabe-Hesketh and B. Everitt, CRC Press, Boca Raton, FL, 2000), S-Plus (*Modern Applied Statistics with S-Plus*, 3rd ed., by W. N. Venables and B. D. Ripley, Springer-Verlag, New York, 1999), and the related free package, R, and GLIM (Aitkin et al. 1989). Most major software now follows the lead of GLIM and includes a generalized linear models routine. Examples are PROC GENMOD in SAS and the glm function in R and S-Plus.

For certain analyses, specialized software is better than the major packages. A good example is StatXact (Cytel Software, Cambridge, Massachusetts), which provides exact analysis for categorical data methods and some nonparametric methods. Among its procedures are small-sample confidence intervals for differences and ratios of proportions and for odds ratios, and Fisher's exact test and its generalizations for $I \times J$ tables. It can also conduct exact tests of conditional independence and of equality of odds ratios in $2 \times 2 \times K$ tables, and exact confidence intervals for the common odds ratio in several 2×2 tables. StatXact uses Monte Carlo methods to approximate exact P-values and confidence intervals when a data set is too large for exact inference to be computationally feasible. Its companion, LogXact, performs exact conditional logistic regression.

Other examples of specialized software are SUDAAN for GEE-type analyses that handle clustering in survey data (Research Triangle Institute, Research Triangle Park, North Carolina), Latent GOLD for latent class modeling (Statistical Innovations, Belmont, Massachusetts), MLn (Institute of Education, London) and HLM (Scientific Software, Chicago) for multilevel models, and PASS for power analyses (NCSS Statistical Software, Kaysville, Utah). S-Plus and R functions are also available from individuals or from published work for particular analyses. For instance, *Statistical Models in S* by J. M. Chambers and T. J. Hastie (Wadsworth, Belmont, California, 1993, p. 227) showed the use of S-Plus in quasi-likelihood analyses using the quasi and make.family functions.

TABLE A.1 SAS Code for Chi-Squared, Measures of Association, and Residuals for Education–Religion Data in Table 3.2

```
data table;
   input degree religion $ count @@;
datalines;
1 fund 178     1 mod 138     1 lib 108
2 fund 570     2 mod 648     2 lib 442
3 fund 138     3 mod 252     3 lib 252
   ;
proc freq order=data; weight count;
  tables degree*religion/chisq expected measures cmh1;
proc genmod order=data; class degree religion;
  model count=degree religion/dist=poi link=log residuals;
```

A.2 EXAMPLES OF SAS CODE BY CHAPTER

The examples below show SAS code (Version 8.1). We focus on basic model fitting rather than the great variety of options. The material is organized by chapter of presentation. For convenience, data for examples are entered in the form of the contingency table displayed in the text. In practice, one would usually enter data at the subject level. These tables and the full data sets are available at *www.stat.ufl.edu/~aa/cda/cda.html*.

Chapters 1–3: Introduction, Two-Way Contingency Tables

Table A.1 uses SAS to analyze Table 3.2. The @@ symbol indicates that each line of data contains more than one observation. Input of a variable as characters rather than numbers requires an accompanying $ label in the INPUT statement. PROC FREQ forms the table with the TABLES statement, ordering row and column categories alphanumerically. To use instead the order in which the categories appear in the data set (e.g., to treat the variable properly in an ordinal analysis), use the ORDER = DATA option in the PROC statement. The WEIGHT statement is needed when one enters the cell counts instead of subject-level data. PROC FREQ can conduct chi-squared tests of independence (CHISQ option), show its estimated expected frequencies (EXPECTED), provide a wide assortment of measures of association and their standard errors (MEASURES), and provide ordinal statistic (3.15) with a "nonzero correlation" test (CMH1). One can also perform chi-squared tests using PROC GENMOD (using loglinear models discussed in the Chapters 8–9 section of this appendix), as shown. Its RESIDUALS option provides cell residuals. The output labeled "StReschi" is the standardized Pearson residual (3.13).

Table A.2 analyzes Table 3.8. With PROC FREQ, for 2×2 tables the MEASURES option in the TABLES statement provides confidence intervals

TABLE A.2 SAS Code for Fisher's Exact Test and Confidence Intervals for Odds Ratio for Tea-Tasting Data in Table 3.8

```
data fisher;
input poured guess count @@;
datalines;
1 1 3   1 2 1   2 1 1   2 2 3
;
proc freq;   weight count;
  tables poured*guess / measures riskdiff;
  exact fisher or / alpha = .05;
proc logistic descending; freq count;
  model guess = poured / clodds = pl;
```

for the odds ratio (labeled "case-control" on output) and the relative risk, and the RISKDIFF option provides intervals for the proportions and their difference. For tables having small cell counts, the EXACT statement can provide various exact analyses. These include Fisher's exact test and its generalization for $I \times J$ tables, treating variables as nominal, with keyword FISHER. The OR keyword gives the odds ratio and its large-sample confidence interval (3.2) and the small-sample interval based on (3.20). Other EXACT statement keywords include binomial tests for 1×2 tables (keyword BINOMIAL), exact trend tests for $I \times 2$ tables (TREND), and exact chi-squared tests (CHISQ) and exact correlation tests for $I \times J$ tables (MHCHI). One can use Monte Carlo simulation (option MC) to estimate exact P-values when the exact calculation is too time consuming. Table A.2 also uses PROC LOGISTIC to get a profile-likelihood confidence interval for the odds ratio (CLODDS = PL). LOGISTIC uses FREQ to serve the same purpose as PROC FREQ uses WEIGHT.

Other

StatXact provides small-sample confidence intervals for a binomial parameter, the difference of proportions, relative risk, and odds ratio. Blaker (2000) gave S-Plus functions that provide his confidence interval for a binomial parameter.

Chapter 4: Models for Binary Response Variables

PROC GENMOD fits GLMs. It specifies the response distribution in the DIST option ("poi" for Poisson, "bin" for binomial, "mult" for multinomial, "negbin" for negative binomial) and specifies the link in the LINK option. Table A.3 illustrates for Table 4.2. For binomial models with grouped data, the response in the model statements takes the form of the number of "successes" divided by the number of cases.

TABLE A.3 SAS Code for Binary GLMs for Snoring Data in Table 4.2

```
data glm;
input snoring disease total @@;
datalines;
0 24 1379    2 35 638    4 21 213    5 30 254
;
proc genmod; model disease/total=snoring/dist=bin link=identity;
proc genmod; model disease/total=snoring/dist=bin link=logit;
proc genmod; model disease/total=snoring/dist=bin link=probit;
```

TABLE A.4 SAS Code for Poisson and Negative Binomial GLMs for Horseshoe Crab Data in Table 4.3

```
data crab;
input color spine width satell weight;
datalines;
3  3  28.3  8  3.05
4  3  22.5  0  1.55
...
3  2  24.5  0  2.00
;
proc genmod;
  model satell=width/dist=poi link=log;
proc genmod;
  model satell=width/dist=poi link=identity;
proc genmod;
  model satell=width/dist=negbin link=identity;
```

Table A.4 uses GENMOD for count modeling of Table 4.3. Each observation refers to a single crab. Using width as the predictor, the first two models use Poisson regression. The third model uses the identity link assuming a negative binomial distribution.

Table A.5 uses GENMOD for the overdispersed data of Table 4.5. A CLASS statement requests dummy variables for the groups. With no intercept in the model (option NOINT) for the identity link, the estimated parameters are the four group probabilities. The ESTIMATE statement provides an estimate, confidence interval, and test for a contrast of model parameters, in this case the difference in probabilities for the first and second groups. The second analysis uses the Pearson statistic to scale standard errors to adjust for overdispersion. PROC LOGISTIC can also provide overdispersion modeling of binary responses; see Table A.27 in the Chapter 13 part of this appendix.

PROC GAM (starting in Version 8.2) fits generalized additive models.

TABLE A.5 SAS Code for Overdispersion Modeling of Teratology Data in Table 4.5

```
data moore;
  input litter group n y @@;
datalines;
 1 1 10 1     2 1 11 4     3 1 12 9     4 1 4 4     5 1 10 10
...
55 4 14 1    56 4  8 0    58 4 17 0
;
proc genmod;  class group;
  model y/n = group / dist = bin link = identity noint;
estimate  'pi1- pi2 ' group 1 -1 0 0;
proc genmod;  class group;
  model y/n = group / dist = bin link = identity noint scale = pearson;
```

Chapters 5 and 6: Logistic Regression

One can fit logistic regression models using either software for GLMs or specialized software for logistic regression. PROC GENMOD uses Newton-Raphson, whereas PROC LOGISTIC uses Fisher scoring. Both yield ML estimates, but SE values use observed information in GENMOD and expected information in LOGISTIC. These are the same for the logit link.

Table A.6 applies GENMOD and LOGISTIC to Table 5.2, when "y" out of "n" crabs had satellites at a given width level. In GENMOD, the LRCI option provides profile likelihood confidence intervals. The ALPHA = option can specify an error probability other than the default of 0.05. The TYPE3 option provides likelihood-ratio tests for each parameter. (In the Chapter 8–9 section we discuss the second GENMOD analysis.)

TABLE A.6 SAS Code for Modeling Grouped Crab Data in Table 5.2

```
data crab;
input width y n satell;  logcases = log(n);
datalines;
22.69  5 14 14
...
30.41 14 14 72
;
proc genmod;
  model y/n = width / dist = bin link = logit 1rci alpha = .01 type3;
proc logistic;
  model y/n = width / influence stb;
  output out = predict p = pi_hat lower = LCL upper = UCL;
proc print data = predict;
proc genmod;
  model satell = width / dist = poi link = log offset = logcases residuals;
```

TABLE A.7 SAS Code for Logit Modeling of AIDS Data in Table 5.5

```
data aids;
input race $ azt $ y n @@;
datalines;
  White Yes 14 107    White No 32 113    Black Yes 11 63    Black No 12 55
;
proc genmod;  class race azt;
  model y/n = azt race / dist = bin type3 lrci residuals obstats;
proc logistic;  class race azt / param = reference;
  model y/n = azt race / aggregate scale = none clparm = both clodds = both;
  output out = predict p = pi_hat lower = lower upper = upper;
proc print data = predict;
proc logistic;  class race azt (ref = first) / param = ref;
  model y/n = azt / aggregate = (azt race) scale = none;
```

With PROC LOGISTIC, logistic regression is the default for binary data. LOGISTIC has a built-in check of whether logistic regression ML estimates exist. It can detect a complete separation of data points with 0 and 1 outcomes. LOGISTIC can also apply other links, such as the probit. Its INFLUENCE option provides Pearson and deviance residuals and diagnostic measures (Pregibon 1981). The STB option provides standardized estimates by multiplying by $s_{x_j}\sqrt{3}/\pi$ (Section 5.4.7 and Note 5.9). Following the model statement, Table A.6 requests predicted probabilities and lower and upper 95% confidence limits for the probabilities.

Table A.7 uses GENMOD and LOGISTIC to fit a logit model with qualitative predictors to Table 5.5. In GENMOD, the OBSTATS option provides various "observation statistics," including predicted values and their confidence limits. The RESIDUALS option requests residuals such as the Pearson and standardized Pearson residuals (labeled "Reschi" and "StReschi"). A CLASS statement requests dummy variables for the factor. By default, in GENMOD the parameter estimate for the last level of each factor equals 0. In LOGISTIC, estimates sum to zero. That is, dummies take the effect coding $(1, -1)$ of 1 when in the category and -1 when not, for which parameters sum to 0. In the CLASS statement in LOGISTIC, the option PARAM = REF requests $(1, 0)$ dummy variables with the last category as the reference level. Also putting REF = FIRST next to a variable name requests its first category as the reference level. The CLPARM = BOTH and CLODDS = BOTH options provide Wald and profile likelihood confidence intervals for parameters and odds ratio effects of explanatory variables. With AGGREGATE SCALE = NONE in the model statement, LOGISTIC reports Pearson and deviance tests of fit; it forms groups by aggregating data into the possible combinations of explanatory variable values, without overdispersion adjustments. Adding variables in parentheses after AGGREGATE (as in the second use of LOGISTIC in Table A.7) specifies the predictors used for forming the table on which to test fit, even when some predictors may have no effect in the model.

TABLE A.8 SAS Code for Logistic Regression Models with Horseshoe Crab Data in Table 4.3

```
data crab;
input color spine width satell weight;
if satell>0 then y = 1; if satell = 0 then y = 0;
if color = 4 then light = 0; if color<4 then light = 1;
datalines;
2 3 28.3 8 3.05
...
2 2 24.5 0 2.00
;
proc genmod descending; class color;
  model y = width color / dist = bin link = logit lrci type3 obstats;
  contrast 'a-d' color 1 0 0 -1;
proc genmod descending;
  model y = width color / dist = bin link = logit;
proc genmod descending;
  model y = width light / dist = bin link = logit;
proc genmod descending; class color spine;
  model y = width weight color spine / dist = bin link = logit type3;
proc logistic descending; class color spine / param = ref;
  model y = width weight color spine / selection = backward lackfit
     outroc = classif1;
proc plot data = classif1; plot _sensit_*_lmspec_ ;
```

Table A.8 shows logistic regression analyses for Table 4.3. The models refer to a constructed binary variable Y that equals 1 when a horseshoe crab has satellites and 0 otherwise. With binary data entry, GENMOD and LOGISTIC order the levels alphanumerically, forming the logit with $(1, 0)$ responses as $\log[P(Y = 0)/P(Y = 1)]$. Invoking the procedure with DESCENDING following the PROC name reverses the order. The first two GENMOD statements use both color and width as predictors; color is qualitative in the first model (by the CLASS statement) and quantitative in the second. A CONTRAST statement tests contrasts of parameters, such as whether parameters for two levels of a factor are identical. The statement shown contrasts the first and fourth color levels. The third GENMOD statement uses a dummy variable for color, indicating whether a crab is light or dark (color = 4). The fourth GENMOD statement fits the main effects model using all the predictors from Table 4.3. LOGISTIC has options for stepwise selection of variables, as the final model statement shows. The LACKFIT option yields the Hosmer–Lemeshow statistic. Using the OUTROC option, LOGISTIC can output a data set for plotting a ROC curve.

Table A.9 analyzes Table 6.9. The CMH option in PROC FREQ specifies the CMH statistic, the Mantel–Haenszel estimate of a common odds ratio and its confidence interval, and the Breslow–Day statistic. FREQ uses the

TABLE A.9 SAS Code for CMH Analysis of Clinical Trial Data in Table 6.9

```
data crab;
input center $ treat response count @@ ;
datalines;
a  1  1 11    a  1  2 25    a  2  1 10    a  2  2 27
...
h  1  1  4    h  1  2  2    h  2  1  6    h  2  2  1
;
proc freq; weight count;
  tables center*treat*response/ cmh chisq;
```

two rightmost variables in the TABLES statement as the rows and columns for each partial table; the CHISQ option yields chi-square tests of independence for each partial table. For $I \times 2$ tables the TREND keyword in the TABLES statement provides the Cochran–Armitage trend test.

Exact conditional logistic regression is available in PROC LOGISTIC with the EXACT statement. It provides ordinary and mid-*P*-values as well as confidence limits for each model parameter and the corresponding odds ratio with the ESTIMATE = BOTH option. One can also conduct the exact conditional version of the Cochran–Armitage test using the TREND option in the EXACT statement with PROC FREQ. Version 9 of SAS will include asymptotic conditional logistic regression, using a STRATA statement to indicate the stratification parameters to be conditioned out. One can also use PROC PHREG to do this (Stokes et al. 2000).

Models with probit and complementary log-log (CLOGLOG) links are available with PROC GENMOD, PROC LOGISTIC, or PROC PROBIT. O'Brien (1986) gave a SAS macro for computing powers using the noncentral chi-squared distribution.

Other

LogXact provides exact conditional logistic regression and StatXact provides exact inference about the odds ratio in $2 \times 2 \times K$ tables. PASS (NCSS Statistical Software) provides power analyses.

Chapter 7: Multinomial Response Models

PROC LOGISTIC fits baseline-category logit models (as of Version 8.2) using the LINK = GLOGIT option. The final response category is the default baseline for the logits. Exact inference is also available using the conditional distribution to eliminate nuisance parameters. PROC CATMOD also fits baseline-category logit models, as Table A.10 shows. CATMOD codes estimates for a factor so that they sum to zero. The PRED = PROB and PRED = FREQ options provide predicted probabilities and fitted values and their standard errors. The POPULATION statement provides the

TABLE A.10 SAS Code for Baseline-Category Logit Models with Alligator Data in Table 7.1

```
data gator;
input lake gender size food count @@;
datalines;
1 1 1 1 7  1 1 1 2 1  1 1 1 3 0  1 1 1 4 0  1 1 1 5 5
...
4 2 2 1 8  4 2 2 2 1  4 2 2 3 0  4 2 2 4 0  4 2 2 5 1
;
proc logistic; freq count; class lake size / param = ref;
  model food(ref = '1') = lake size / link = glogit
      aggregate scale = none;
proc catmod; weight count;
  population lake size gender;
  model food = lake size / pred = freq pred = prob;
```

variables that define the predictor settings. For instance, with "gender" in that statement, the model with lake and size effects is fitted to the full table also classified by gender.

PROC GENMOD can fit the proportional odds version of cumulative logit models using the DIST = MULTINOMIAL and LINK = CLOGIT options. Table A.11 fits it to Table 7.5. When the number of response categories exceeds 2, by default PROC LOGISTIC fits this model. It also gives a score test of the proportional odds assumption of identical effect parameters for each cutpoint. Both procedures use the $\alpha_j + \beta x$ form of the model. Cox (1995) used PROC NLIN for the more general model (7.8) having a scale parameter.

Both GENMOD and LOGISTIC can use other links in cumulative link models. GENMOD uses LINK = CPROBIT for the cumulative probit model and LINK = CCLL for the cumulative complementary log-log model. Table A.11 uses LINK = PROBIT in LOGISTIC to fit a cumulative probit model.

TABLE A.11 SAS Code for Cumulative Logit and Probit Models with Mental Impairment Data in Table 7.5

```
data impair;
input mental ses life;
datalines;
1 1 1
...
4 0 9
;
proc genmod ;
  model mental = life ses / dist = multinomial link = clogit lrci type3;
proc logistic;
  model mental = life ses / link = probit;
```

TABLE A.12 SAS Code for Adjacent-Categories Logit and Mean Response Models and CMH Analysis of Job Satisfaction Data in Table 7.8

```
data jobsat;
input gender income satisf count @@;
count2 = count + .01;
datalines;
1 1 1 1  1 1 2 3  1 1 3 11  1 1 4 2
...
0 4 1 0  0 4 2 1  0 4 3  9  0 4 4 6
;
proc catmod order = data; * ML analysis of adj-cat logit (ACL) model;
    weight count;
    population gender income;
    model satisf =
        (1 0 0  3 3,  0 1 0 2 2,  0 0 1 1 1,
         1 0 0  6 3,  0 1 0 4 2,  0 0 1 2 1,
         1 0 0  9 3,  0 1 0 6 2,  0 0 1 3 1,
         1 0 0 12 3,  0 1 0 8 2,  0 0 1 4 1,
         1 0 0  3 0,  0 1 0 2 0,  0 0 1 1 0,
         1 0 0  6 0,  0 1 0 4 0,  0 0 1 2 0,
         1 0 0  9 0,  0 1 0 6 0,  0 0 1 3 0,
         1 0 0 12 0,  0 1 0 8 0,  0 0 1 4 0)
          /ml pred = freq;
proc catmod order = data; weight count2; * WLS analysis of ACL model;
  response alogits; population gender income; direct gender income;
  model satisf = _response_ gender income;
proc catmod; weight count; * mean response model;
  population gender income; response mean; direct gender income;
  model satisf = gender income / covb;
proc freq; weight count;
  tables gender*income*satisf / cmh scores = table;
```

One can fit adjacent-categories logit models in CATMOD by fitting equivalent baseline-category logit models. Table A.12 uses it for Table 7.8, where each line of code in the model statement specifies the predictor values (for the three intercepts, income, and gender) for the three logits. The income and gender predictor values are multiplied by 3 for the first logit, 2 for the second, and 1 for the third, to make effects comparable in the two models. PROC CATMOD has options (CLOGITS and ALOGITS) for fitting cumulative logit and adjacent-categories logit models to ordinal responses; however, those options provide weighted least squares (WLS) rather than ML fits. A constant must be added to empty cells for WLS to run. CATMOD treats zero counts as structural zeros, so they must be replaced by small constants when they are actually sampling zeros. The DIRECT statements identify predictors treated as quantitative. The second analysis in Table A.12 uses the ALOGITS option. CATMOD can also fit mean response models using WLS, as the third analysis in Table A.12 shows.

With the CMH option, PROC FREQ provides the generalized CMH tests of conditional independence. The statistic for the "general association"

alternative treats X and Y as nominal [statistic (7.20)], the statistic for the "row mean scores differ" alternative treats X as nominal and Y as ordinal, and the statistic for the "nonzero correlation" alternative treats X and Y as ordinal [statistic (7.21)]. Table A.12 analyzes Table 7.8, using scores $(1, 2, 3, 4)$ for each variable.

PROC MDC fits multinomial discrete choice models, with logit and probit links. One can also use PROC PHREG, which is designed for the Cox proportional hazards model for survival analysis, because the partial likelihood for that analysis has the same form as the likelihood for the multinomial model (Allison 1999, Chap. 7; Chen and Kuo 2001).

Other

LogXact provides exact conditional analyses for baseline-category logit models. Joseph Lang (*jblang@stat.uiowa.edu*) has an R function that can fit mean response models by ML.

Chapters 8 and 9: Loglinear Models

Table A.13 uses GENMOD to fit model (AC, AM, CM) to Table 8.3. Table A.14 uses GENMOD for table raking of Table 8.15. Table A.15 uses GENMOD to fit the linear-by-linear association model (9.6) and the row effects model (9.8) to Table 9.3 (with column scores $1, 2, 4, 5$). The defined

TABLE A.13 SAS Code for Fitting Loglinear Models to Drug Survey Data in Table 8.3

```
data drugs;
input a c m count @@;
datalines;
1 1 1 911    1 1 2 538    1 2 1 44    1 2 2 456
2 1 1   3    2 1 2  43    2 2 1  2    2 2 2 279
;
proc genmod;  class a c m;
  model count = a c m a*m a*c c*m / dist = poi link = log lrci type3 obstats;
```

TABLE A.14 SAS Code for Raking Table 8.15

```
data rake;
input school atti count @@;
log_c = log(count); pseudo = 100 / 3;
data lines;
1 1 209    1 2 101    1 3 237
...
;
proc genmod; class school atti;
  model pseudo = school atti / dist = poi link = log offset = log_c obstats;
```

TABLE A.15 SAS Code for Fitting Association Models to GSS Data in Table 9.3

```
data sex;
input premar birth u v count @@; assoc = u*v ;
datalines;
1 1 1 1 38    1 2 1 2 60    1 3 1 4 68    1 4 1 5 81
...
;
proc genmod;  class premar birth;
  model count = premar birth assoc / dist = poi link = log;
proc genmod;  class premar birth;
  model  count = premar birth premar*v / dist = poi link = log;
```

variable "assoc" represents the cross-product of row and column scores, which has β parameter as coefficient in model (9.6). Table A.6 uses GENMOD to fit the Poisson regression model with log link for the grouped data of Table 5.2. It models the total number of satellites at each width level (variable "satell"), using the log of the number of cases as offset.

Correspondence analysis is available with PROC CORRESP.

Other

Prof. Joseph Lang (*jblang@stat.uiowa.edu*) has R and S-Plus functions for ML fitting of the generalized loglinear model (8.18). Becker (1990) gave a FORTRAN program that fits the $RC(M)$ model.

Chapter 10: Models for Matched Pairs

Table A.16 analyzes Table 10.1. For square tables, the AGREE option in PROC FREQ provides the McNemar chi-squared statistic for binary matched pairs, the X^2 test of fit of the symmetry model (also called *Bowker's test*),

TABLE A.16 SAS Code for McNemar's Test and Comparing Proportions for Matched Samples in Table 10.1

```
data matched;
input first second count @@;
datalines;
1 1 794    1 2 150    2 1 86    2 2 570
;
proc freq;  weight count;
  tables first*second/ agree;  exact mcnem;
proc catmod;  weight count;
  response marginals;
  model first*second= (1  0 ,
                       1  1 ;
```

TABLE A.17 SAS Code for Testing Marginal Homogeneity with Migration Data in Table 10.6

```
data migrate;
input then $ now $ count  m11 m12 m13 m21 m22 m23 m31 m32 m33 m44 m1 m2 m3;
datalines;
  ne  ne 11607   1   0   0   0   0   0   0   0   0   0   0   0
  ne  mw   100   0   1   0   0   0   0   0   0   0   0   0   0
  ne   s   366   0   0   1   0   0   0   0   0   0   0   0   0
  ne   w   124  -1  -1  -1   0   0   0   0   0   0   1   0   0
  mw  ne    87   0   0   0   1   0   0   0   0   0   0   0   0
  mw  mw 13677   0   0   0   0   1   0   0   0   0   0   0   0
  mw   s   515   0   0   0   0   0   1   0   0   0   0   0   0
  mw   w   302   0   0   0  -1  -1  -1   0   0   0   0   1   0
   s  ne   172   0   0   0   0   0   0   1   0   0   0   0   0
   s  mw   225   0   0   0   0   0   0   0   1   0   0   0   0
   s   s 17819   0   0   0   0   0   0   0   0   1   0   0   0
   s   w   270   0   0   0   0   0   0  -1  -1  -1   0   0   0   1
   w  ne    63  -1   0   0  -1   0   0  -1   0   0   0   1   0   0
   w  mw   176   0  -1   0   0  -1   0   0  -1   0   0   0   1   0
   w   s   286   0   0  -1   0   0  -1   0   0  -1   0   0   0   1
   w   w 10192  0   0   0   0   0   0   0   0   0   1   0   0   0
       ;
proc genmod;
  model count = m11 m12 m13 m21 m22 m23 m31 m32 m33 m44 m1 m2 m3
     / dist = poi  link = identity;
proc catmod;  weight count;  response marginals;
  model then*now = _response_ /freq;
  repeated time 2;
```

and Cohen's kappa and weighted kappa with SE values. The MCNEM keyword in the EXACT statement provides a small-sample binomial version of McNemar's test. PROC CATMOD can provide the confidence interval for the difference of proportions. The code forms a model for the marginal proportions in the first row and the first column, specifying a model matrix in the model statement that has an intercept parameter (the first column) that applies to both proportions and a slope parameter that applies only to the second; hence the second parameter is the difference between the second and first marginal proportions.

PROC LOGISTIC can conduct conditional logistic regression.

Table A.17 shows ways of testing marginal homogeneity for Table 10.6. The GENMOD code shows the Lipsitz et al. (1990) approach, expressing the I^2 expected frequencies in terms of parameters for the $(I-1)^2$ cells in the first $I-1$ rows and $I-1$ columns, the cell in the last row and last column, and $I-1$ marginal totals (which are the same for rows and columns). Here, m11 denotes expected frequency μ_{11}, m1 denotes $\mu_{1+} = \mu_{+1}$, and so on. This parameterization uses formulas such as $\mu_{14} = \mu_{1+} - \mu_{11} - \mu_{12} - \mu_{13}$ for terms in the last column or last row. CATMOD provides the Bhapkar test (10.16) of marginal homogeneity, as shown.

TABLE A.18 SAS Code Showing Square-Table Analysis of Table 10.5

```
data sex;
input premar extramar symm qi count @@;
unif = premar*extramar;
datalines;
1 1 1 1 144   1 2 2 5  2   1 3 3 5  0   1 4  4 5 0
2 1 2 5  33   2 2 5 2  4   2 3 6 5  2   2 4  7 5 0
3 1 3 5  84   3 2 6 5 14   3 3 8 3  6   3 4  9 5 1
4 1 4 5 126   4 2 7 5 29   4 3 9 5 25   4 4 10 4 5
;
proc genmod;  class symm;
  model  count = symm / dist = poi  link = log; * symmetry;
proc genmod;  class extramar premar symm;
  model  count = symm extramar premar / dist = poi  link = log; *QS;
proc genmod;  class symm;
  model  count = symm extramar premar / dist = poi  link = log; * ordinal QS;
proc genmod;  class extramar premar qi;
  model  count = extramar premar qi / dist = poi  link = log; * quasi indep;
proc genmod;  class extramar premar;
  model  count = extramar premar unif / dist = poi  link = log;
data sex2;
input score below above @@;  trials = below + above;
datalines;
1 33 2   1 14 2   1 25 1   2 84 0   2 29 0   3 126 0
;
proc genmod data = sex2;
  model above / trials = score / dist = bin link = logit noint;
 proc genmod data = sex2;
  model  above / trials = /dist = bin  link = logit noint;
proc genmod data = sex2;
  model  above / trials = /dist = bin  link = logit;
```

Table A.18 shows various square-table analyses of Table 10.5. The "symm" factor indexes the pairs of cells that have the same association terms in the symmetry and quasi-symmetry models. For instance, "symm" takes the same value for cells $(1, 2)$ and $(2, 1)$. Including this term as a factor in a model invokes a parameter λ_{ij} satisfying $\lambda_{ij} = \lambda_{ji}$. The first model fits this factor alone, providing the symmetry model. The second model looks like the third except that it identifies "premar" and "extramar" as class variables (for quasi-symmetry), whereas the third model statement does not (for ordinal quasi-symmetry). The fourth model fits quasi-independence. The "qi" factor invokes the δ_i parameters. It takes a separate level for each cell on the main diagonal and a common value for all other cells. The fifth model fits the quasi-uniform association model (10.29).

The bottom of Table A.18 fits square-table models as logit models. The pairs of cell counts (n_{ij}, n_{ji}), labeled as "above" and "below" with reference to the main diagonal, are six sets of binomial counts. The variable defined as "score" is the distance $(u_j - u_i) = j - i$. The first two cases are symmetry

TABLE A.19 SAS Code for Fitting Bradley–Terry Model to Table 10.10

```
data baseball;
input  wins  games  milw  detr  toro  newy  bost  clev  balt;
datalines;
7  13  1 -1  0  0  0  0  0
...
6  13  0  0  0  0  0  1 -1
;
proc genmod;
  model  wins / games = milw  detr  toro  newy  bost  clev  balt /
  dist = bin  link = logit noint covb;
```

and ordinal quasi-symmetry. Neither model contains an intercept (NOINT), and the ordinal model uses "score" as the predictor. The third model allows an intercept and is the conditional symmetry model (10.28).

Table A.19 uses GENMOD for logit fitting of the Bradley–Terry model to Table 10.10 by forming an artificial explanatory variable for each team. For a given observation, the variable for team i is 1 if it wins, -1 if it loses, and 0 if it is not one of the teams for that match. Each observation lists the number of wins ("wins") for the team with variate-level equal to 1 out of the number of games ("games") against the team with variate-level equal to -1. The model has these artificial variates, one of which is redundant, as explanatory variables with no intercept term. The COVB option provides the estimated covariance matrix of parameter estimators.

Chapter 11: Analyzing Repeated Categorical Response Data

Table A.20 uses GENMOD for the likelihood-ratio test of marginal homogeneity for Table 11.1, where for instance *m11p* denotes μ_{11+}. The marginal homogeneity model expresses the eight cell expected frequencies in terms of

TABLE A.20 SAS Code for Testing Marginal Homogeneity with Crossover Study of Table 11.1

```
data crossover;
input  a  b  c  count  m111  m11p  m1p1  mp11  m1pp  m222  @@;
datalines;
1  1  1  6   1  0  0  0  0  0    1  1  2  16 -1  1  0  0  0  0
1  2  1  2  -1  0  1  0  0  0    1  2  2   4  1 -1 -1  0  1  0
2  1  1  2  -1  0  0  1  0  0    2  1  2   4  1 -1  0 -1  1  0
2  2  1  6   1  0 -1 -1  1  0    2  2  2   6  0  0  0  0  0  1
;
proc genmod;
  model  count = m111  m11p  m1p1  mp11  m1pp  m222 / dist = poi  link = identity;
proc catmod;  weight count;  response marginals;
  model  a*b*c = _response_ /freq;
  repeated drug 3;
```

TABLE A.21 SAS Code for Marginal Modeling of Depression Data in Table 11.2

```
data depress;
input case diagnose drug time outcome @@; * outcome = 1 is normal;
datalines;
  1  0  0  0  1    1  0  0  1  1    1  0  0  2  1
...
340  1  1  0  0  340  1  1  1  0  340  1  1  2  0
;
proc genmod descending; class case;
  model; outcome = diagnose drug time drug*time / dist = bin link = logit type3;
  repeated subject = case / type = exch corrw;
proc n1mixed qpoints = 200;
  parms alpha = -.03 beta1 = -1.3 beta2 = -.06 beta3 = .48 beta4 = 1.02 sigma = .066;
  eta = alpha + beta1*diagnose + beta2*drug + beta3*time + beta4*drug*time + u;
  p = exp(eta) / (1 + exp(eta));
  model outcome ~ binary(p);
  random u ~ normal(0, sigma*sigma) subject = case;
```

TABLE A.22 SAS Code for GEE and Random Intercept Cumulative Logit Analysis of Insomnia Data in Table 11.4

```
data francom;
  input case treat time outcome @@;
datalines;
  1  1  0  1    1  1  1  1
...
239  0  0  4  239  0  1  4
;
proc genmod; class case;
  model outcome = treat time treat*time / dist = multinomial
      link = clogit;
  repeated subject = case / type = indep corrw;
proc n1mixed qpoints = 40;
  bounds i2>0; bounds i3>0;
  eta1 = i1 + treat*beta1 + time*beta2 + treat*time*beta3 + u;
  eta2 = i1 + i2 + treat*beta1 + time*beta2 + treat*time*beta3 + u;
  eta3 = i1 + i2 + i3 + treat*beta1 + time*beta2 + treat*time*beta3 + u;
  p1 = exp(eta) / (1 + exp(eta1));
  p2 = exp(eta2) / (1 + exp(eta2)) - exp(eta1) / (1 + exp(eta1));
  p3 = exp(eta3) / (1 + exp(eta3)) - exp(eta2) / (1 + exp(eta2));
  p4 = 1 - exp(eta3) / (1 + exp(eta3));
  ll = y1*log(p1) + y2*log(p2) + y3*log(p3) + y4*log(p4);
  model y1 ~ general(ll);
  estimate 'interc2' i1 + i2; * this is alpha_2 in model, and
      i1 is alpha_1;
  estimate 'interc3' i1 + i2 + i3; * this is alpha_3 in model;
  random u ~ normal(0, sigma*sigma) subject = case;
```

μ_{111}, μ_{11+}, μ_{1+1}, μ_{+11}, μ_{1++}, and μ_{222} (since $\mu_{+1+} = \mu_{++1} = \mu_{1++}$). Note, for instance, that $\mu_{112} = \mu_{11+} - \mu_{111}$ and $\mu_{122} = \mu_{111} + \mu_{1++} - \mu_{11+} - \mu_{1+1}$. CATMOD provides the generalized Bhapkar test (11.5) of marginal homogeneity.

Table A.21 uses GENMOD to analyze Table 11.2 using GEE. Possible working correlation structures are TYPE = EXCH for exchangeable, TYPE = AR for autoregressive, TYPE = INDEP for independence, and TYPE = UNSTR for unstructured. Output shows estimates and standard errors under the naive working correlation and based on the sandwich matrix incorporating the empirical dependence. Alternatively, the working association structure in the binary case can use the log odds ratio (e.g., using LOGOR = EXCH for exchangeability). The type 3 option in GEE provides score tests about effects. See Stokes et al. (2000, Sec. 15.11) for the use of GEE with missing data.

Table A.22 uses GENMOD to implement GEE for a cumulative logit model for Table 11.4. For multinomial responses, independence is currently the only working correlation structure.

Other

Joseph Lang (*jblang@stat.uiowa.edu*) has R and S-Plus functions for ML fitting of marginal models through the generalized loglinear model (11.8), using the constraint approach with Lagrange multipliers. The program MAREG (Kastner et al. 1997) provides GEE fitting and ML fitting of marginal models with the Fitzmaurice and Laird (1993) approach, allowing multicategory responses. See *www.stat.uni-muenchen.de/~andreas/mareg/winmareg.html*.

Chapter 12: Random Effects: Generalized Linear Mixed Models

PROC NLMIXED extends GLMs to GLMMs by including random effects. Table A.23 analyzes the matched pairs model (12.3). Table A.24 analyzes the election data in Table 12.2.

TABLE A.23 SAS Code for Fitting Model (12.3) for Matched Pairs to Table 12.1

```
data matched;
input case occasion response count @@;
datalines;
2  0  1 794    1  1  1 794    2  0  1 150    2  1  0 150
3  0  0  86    3  1  1  86    4  0  0 570    4  1  0 570
;
proc nlmixed;
  eta = alpha + beta*occasion + u;  p = exp(eta) / (1 + exp(eta));
  model  response ~ binary(p);
  random u ~ normal(0, sigma*sigma) subject = case;
  replicate count;
```

TABLE A.24 SAS Code for GLMM Analysis of Election Data in Table 12.2

```
data vote;
input y n;
case = _n_;
datalines;
 1   5
16  32
...
 1   4
;
proc n1mixed;
  eta = alpha + u;  p = exp(eta) / (1 + exp(eta));
  model y ~ binomial(n,p);
  random u ~ normal (0, sigma*sigma) subject = case;
  predict p out = new;
proc print data = new;
```

TABLE A.25 SAS Code for GLMM Modeling of Opinions in Table 10.13

```
data new;
input sex poor single any count;
datalines;
1  1  1  1 342
...
2  0  0  0 457
;
data new;  set new;
  sex = sex- 1;  case = _n_;
  q1 = 1; q2 = 0; resp = poor; output;
  q1 = 0, q2 = 1; resp = single;  output;
  q1 = 0; q2 = 0; resp = any; output;
drop poor single any;
proc n1mixed  qpoints = 50;
  parms alpha = 0 beta1 = .8 beta2 = .3 gamma = 0 sigma = 8.6;
  eta = alpha + beta1*q1 + beta2*q2 + gamma*sex + u;
  p = exp(eta) / (1 + exp(eta));
  model resp ~ binary(p);
  random u ~ normal(0, sigma*sigma) subject = case;
  replicate count;
```

TABLE A.26 SAS Code for GLMM for Leading Crowd Data in Table 12.8

```
data crowd;
input  mem1  att1  mem2  att2  count;
datalines;
  1  1  1  1 458
...
  0  0  0  0 554
;
data new;  set crowd;
  case = _n_;
  x1m = 1;  x1a = 0;  x2m = 0;  x2a = 0;  var = 1;  resp = mem1;  output;
  x1m = 0;  x1a = 1;  x2m = 0;  x2a = 0;  var = 0;  resp = att1;  output;
  x1m = 0;  x1a = 0;  x2m = 1;  x2a = 0;  var = 1;  resp = mem2;  output;
  x1m = 0;  x1a = 0;  x2m = 0;  x2a = 1;  var = 0;  resp = att2; output;
  drop  mem1  att1  mem2  att2;
proc  n1mixed  data = new;
  eta = beta1m*x1m + beta1a*x1a + beta2m*x2m + beta2a*x2a + um*var +
     ua*(1 - var);
  p = exp(eta) / (1 + exp(eta));
  model resp ~ binary(p);
  random  um  ua ~ normal([0,0],[s1*s1, cov12, s2*s2]) subject = case;
  replicate count;
  estimate 'mem change' beta2m - beta1m; estimate 'att change'
     beta2a - beta1a;
```

Table A.25 fits model (12.10) to Table 10.13. This shows how to set initial values and set the number of quadrature points for Gauss–Hermite quadrature (e.g., QPOINTS =). One could let SAS fit without initial values but then take that fit as initial values in further runs, increasing QPOINTS until estimates and standard errors converge to the necessary precision.

Table A.21 uses NLMIXED for Table 11.2. Table A.22 uses NLMIXED for ordinal modeling of Table 11.4, defining a general multinomial log likelihood. Table A.26 shows a correlated bivariate random effect analysis of Table 12.8. Agresti et al. (2000) showed NLMIXED examples for clustered data, Agresti and Hartzel (2000) showed code for multicenter trials such as Table 12.5, and Hartzel et al. (2001a) showed code for multicenter trials with an ordinal response. The Web site for the journal *Statistical Modelling* shows NLMIXED code for an adjacent-categories logit model and a nominal model at the data archive for Hartzel et al. (2001b). Chen and Kuo (2001) discussed fitting multinomial logit models, including discrete-choice models, with random effects.

Other

MLn (Institute of Education, London) and HLM (Scientific Software, Chicago) fit multilevel models. MIXOR is a FORTRAN program for ML

TABLE A.27 SAS Code for Overdispersion Analysis of Table 4.5

```
data moore;
input  litter  group  n  y  @@;
  z2 = 0;  z3 = 0;  z4 = 0;
  if group = 2 then z2 = 1;  if group = 3 then z3 = 1;  if group = 4
     then z4 = 1;
datalines;
 1  1 10  1    2  1 11  4    3  1 12  9    4  1  4  4
...
55  4 14  1   56  4  8  0   57  4  6  0   58  4 17  0
;
proc logistic;
  model y / n = z2 z3 z4 / scale = williams;
proc logistic;
  model y / n = z2 z3 z4 / scale = pearson;
proc n1mixed  qpoints = 200;
  eta = alpha + beta2*z2 + beta3*z3 + beta4*z4 + u;
  p = exp(eta) / (1 + exp(eta));
  model y ~ binomial(n,p);
  random u ~ normal(0, sigma*sigma) subject = litter;
```

TABLE A.28 SAS Code for Fitting Models to Murder Data in Table 13.6

```
data new;
input white black other response;
datalines;
1070  119  55   0
  60   16   5   1
...
   1    0   0   6
;
data new; set new; count = white; race = 0; output;
  count = black; race = 1; output; drop white black other;
data new2; set new; do i = 1 to count; output; end; drop i;
proc genmod data = new2;
  model response = race / dist = negbin link = log;
proc genmod data = new2;
  model response = race / dist = poi  link = log scale = pearson;
data new; set new; case = _n_;
proc n1mixed data = new  qpoints = 400;
  parms alpha = -3.7 beta = 1.90 sigma = 1.6;
  eta = alpha + beta*race + u; mu = exp(eta);
  model response ~ poisson(mu);
  random u ~ normal(0, sigma*sigma) subject = case;
  replicate count;
```

fitting of binary and ordinal random effects models available from Don Hedeker (*www.uic.edu/~hedeker/mix.html*).

Chapter 13: Other Mixture Models for Categorical Data

PROC LOGISTIC provides two overdispersion approaches for binary data. The SCALE = WILLIAMS option uses variance function of the beta-binomial form (13.10), and SCALE = PEARSON uses the scaled binomial variance (13.11). Table A.27 illustrates for Table 4.5. That table also uses NLMIXED for adding litter random intercepts.

For Table 13.6, Table A.28 uses GENMOD to fit a negative binomial model and a quasi-likelihood model with scaled Poisson variance using the Pearson statistic, and NLMIXED to fit a Poisson GLMM. PROC NLMIXED can also fit negative binomial models.

Other

Latent GOLD (developed by J. Vermunt and J. Magidson for Statistical Innovations, Belmont, Massachusetts) can fit a wide variety of mixture models, including latent class models, nonparametric mixtures of logistic regression, and some Rasch mixture models.

Categorical Data Analysis, Second Edition. Alan Agresti

ISBN: 0-471-36093-7

APPENDIX B

Chi-Squared Distribution Values

	Right-Tailed Probability						
df	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	10.22	13.36	15.51	17.53	20.09	21.96	26.12
9	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	23.83	28.41	31.41	34.17	37.57	40.00	45.32
25	29.34	34.38	37.65	40.65	44.31	46.93	52.62
30	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	45.62	51.80	55.76	59.34	63.69	66.77	73.40
50	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	77.58	85.53	90.53	95.02	100.4	104.2	112.3
80	88.13	96.58	101.8	106.6	112.3	116.3	124.8
90	98.65	107.6	113.1	118.1	124.1	128.3	137.2
100	109.1	118.5	124.3	129.6	135.8	140.2	149.5

Source: Calculated using *StaTable*, Cytel Software, Cambridge, MA.

Categorical Data Analysis, Second Edition. Alan Agresti

ISBN: 0-471-36093-7

References

Adelbasit, K. M., and R. L. Plackett. 1983. Experimental design for binary data. *J. Amer. Statist. Assoc.* **78**: 90–98.

Agresti, A. 1984. *Analysis of Ordinal Categorical Data*. New York: Wiley.

Agresti, A. 1992. A survey of exact inference for contingency tables. *Statist. Sci.* **7**: 131–153.

Agresti, A. 1993. Computing conditional maximum likelihood estimates for generalized Rasch models using simple loglinear models with diagonal parameters. *Scand. J. Statist.* **20**: 63–71.

Agresti, A. 1997. A model for repeated measurements of a multivariate binary response. *J. Amer. Statist. Assoc.* **92**: 315–321.

Agresti, A. 1999. On logit confidence intervals for the odds ratio with small samples. *Biometrics* **55**: 597–602.

Agresti, A. 2001. Exact inference for categorical data: Recent advances and continuing controversies. *Statist. Medic.* **20**: 2709–2722.

Agresti, A., and B. Caffo. 2000. Simple and effective confidence intervals for proportions and difference of proportions result from adding two successes and two failures. *Amer. Statist.* **54**: 280–288.

Agresti, A., and B. A. Coull. 1998. Approximate is better than exact for interval estimation of binomial parameters. *Amer. Statist.* **52**: 119–126.

Agresti, A., and J. Hartzel. 2000. Strategies for comparing treatments on a binary response with multi-centre data. *Statist. Medic.* **19**(8): 1115–1139.

Agresti, A., and J. Lang. 1993a. A proportional odds model with subject-specific effects for repeated ordered categorical responses. *Biometrika* **80**: 527–534.

Agresti, A., and J. Lang. 1993b. Quasi-symmetric latent class models, with application to rater agreement. *Biometrics* **49**: 131–139.

Agresti, A., and I. Liu. 1999. Modeling a categorical variable allowing arbitrarily many category choices. *Biometrics* **55**: 936–943.

Agresti, A., and Y. Min. 2001. On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* **57**: 963–971.

Agresti, A., and R. Natarajan. 2001. Modeling clustered ordered categorical data: A survey. *Internal. Statist. Rev.* **69**: 345–371.

Agresti, A., D. Wackerly, and J. Boyett. 1979. Exact conditional tests for cross-classifications: Approximation of attained significance levels. *Psychometrika* **44**: 75–84.

Agresti, A., C. Chuang, and A. Kezouh. 1987. Order-restricted score parameters in association models for contingency tables. *J. Amer. Statist. Assoc.* **82**: 619–623.

Agresti, A., C. R. Mehta, and N. R. Patel. 1990. Exact inference for contingency tables with ordered categories. *J. Amer. Statist. Assoc.* **85**: 453–458.

Agresti, A., J. Booth, J. Hobert, and B. Caffo. 2000. Random-effects modeling of categorical response data. *Sociol. Methodol.* **30**: 27–81.

Aitchison, J., and C. G. G. Aitken. 1976. Multivariate binary discrimination by the kernel method. *Biometrika* **63**: 413–420.

Aitchison, J., and C. H. Cho. 1989. The multivariate Poisson-log normal distribution. *Biometrika* **76**: 643–653.

Aitchison, J., and S. M. Shen. 1980. Logistic-normal distributions: Some properties and uses. *Biometrika* **67**: 261–272.

Aitchison, J., and S. D. Silvey. 1957. The generalization of probit analysis to the case of multiple responses. *Biometrika* **44**: 131–140.

Aitchison, J., and S. D. Silvey. 1958. Maximum likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.* **29**: 813–828.

Aitkin, M. 1979. A simultaneous test procedure for contingency table models. *Appl. Statist.* **28**: 233–242.

Aitkin, M. 1980. A note on the selection of log-linear models. *Biometrics* **36**: 173–178.

Aitkin, M. 1999. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**: 117–128.

Aitkin, M., and D. Clayton. 1980. The fitting of exponential, Weibull, and extreme value distributions to complex censored survival data using GLIM. *Appl. Statist.* **29**: 156–163.

Aitkin, M., and M. Stasinopoulos. 1989. Likelihood analysis of a binomial sample size problem. Pp. 399–411 in *Contributions to Probability and Statistics*: *Essays in Honor of Ingram Olkin*, ed. L. J. Gleser, M. D. Perlman, S. J. Press, and A. R. Sampson. New York: Springer-Verlag.

Aitkin, M., D. Anderson, and J. Hinde. 1981. Statistical modelling of data on teaching styles. *J. Roy. Statist. Soc. Ser. A* **144**: 419–461.

Aitkin, M., D. Anderson, B. Francis, and J. Hinde. 1989. *Statistical Modeling in GLIM*. Oxford: Clarendon Press.

Albert, J. H. 1997. Bayesian testing and estimation of association in a two-way contingency table. *J. Amer. Statist. Assoc.* **92**: 685–693.

Albert, A., and J. A. Anderson. 1984. On the existence of maximum likelihood estimates in logistic models. *Biometrika* **71**: 1–10.

Albert, J. H., and S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88**: 669–679.

Albert, J. H., and A. K. Gupta. 1982. Mixtures of Dirichlet distributions and estimation in contingency tables. *Ann. Statist.* **10**: 1261–1268.

Allison, P. D. 1999. *Logistic Regression Using the SAS System*. Cary, NC: SAS Institute.

Altham, P. M. E. 1969. Exact Bayesian analysis of a 2×2 contingency table and Fisher's "exact" significance test. *J. Roy. Statist. Soc. Ser B* **31**: 261–269.

Altham, P. M. E. 1970. The measurement of association of rows and columns for an $r \times s$ contingency table. *J. Roy. Statist. Soc. Ser B* **32**: 63–73.

Altham, P. M. E. 1971. The analysis of matched proportions. *Biometrika* **58**: 561–576.

Altham, P. M. E. 1975. Quasi-independent triangular contingency tables. *Biometrics* **31**: 233–238.

Altham, P. M. E. 1978. Two generalizations of the binomial distribution. *Appl. Statist.* **27**: 162–167.

Altham, P. M. E. 1984. Improving the precision of estimation by fitting a model. *J. Roy. Statist. Soc. Ser B* **46**: 118–119.

Amemiya, T. 1981. Qualitative response models: A survey. *J. Econom. Literature* **19**: 1483–1536.

Andersen, E. B. 1970. Asymptotic properties of conditional maximum-likelihood estimators. *J. Roy. Statist. Soc. Ser B* **32**: 283–301.

Andersen, E. B. 1980. *Discrete Statistical Models with Social Science Applications*. Amsterdam: North-Holland.

Andersen, E. B. 1995. Polytomous Rasch models and their estimation. Pp. 272–291 in *Rasch Models: Foundations, Recent Developments, and Applications*, eds. G. Fischer and I. Molenaar. New York: Springer-Verlag.

Anderson, J. A. 1972. Separate sample logistic discrimination. *Biometrika* **59**: 19–35.

Anderson, J. A. 1975. Quadratic logistic discrimination. *Biometrika* **62**: 149–154.

Anderson, J. A. 1984. Regression and ordered categorical variables. *J. Roy. Statist. Soc. Ser B* **46**: 1–30.

Anderson, D. A., and M. Aitkin. 1985. Variance component models with binary response: Interviewer variability. *J. Roy. Statist. Soc. Ser B* **47**: 203–210.

Anderson, C. J., and U. Böckenholt. 2000. Graphical regression models for polytomous variables. *Psychometrika* **65**: 497–509.

Anderson, T. W., and L. A. Goodman. 1957. Statistical inference about Markov chains. *Ann. Math. Statist.* **28**: 89–110.

Anderson, J. A., and P. R. Philips. 1981. Regression, discrimination, and measurement models for ordered categorical variables. *Appl. Statist.* **30**: 22–31.

Anderson, C. J., and J. K. Vermunt. 2000. Log-multiplicative models as latent variable models for nominal and/or ordinal data. *Sociol. Methodol.* **30**: 81–121.

Aranda-Ordaz, F. J. 1981. On two families of transformations to additivity for binary response data. *Biometrics* **68**: 357–363.

Aranda-Ordaz, F. J. 1983. An extension of the proportional hazards model for grouped data. *Biometrics* **39**: 109–117.

Arminger, G., C. C. Clogg, and T. Cheng. 2000. Regression analysis of multivariate binary response variables using Rasch-type models and finite mixture methods. *Sociol. Methodol.* **30**: 1–26.

Armitage, P. 1955. Tests for linear trends in proportions and frequencies. *Biometrics* **11**: 375–386.

Ashford, J. R., and R. D. Sowden. 1970. Multivariate probit analysis. *Biometrics* **26**: 535–546.

Asmussen, S., and D. Edwards. 1983. Collapsibility and response variables in contingency tables. *Biometrika* **70**: 567–578.

Azzalini, A. 1994. Logistic regression for autocorrelated data with application to repeated measures. *Biometrika* **81**: 767–775.

Baglivo, J., D. Olivier, and M. Pagano. 1992. Methods for exact goodness-of-fit tests. *J. Amer. Statist. Assoc.* **87**: 464–469.

Baker, S. G. 1992. A simple method for computing the observed information matrix when using the EM algorithm with categorical data. *J. Comput. Graph. Statist.* **1**: 63–76.

Baker, S. G., and N. M. Laird. 1988. Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *J. Amer. Statist. Assoc.* **83**: 62–69.

Baker, R. J., M. R. B. Clarke, and P. W. Lane. 1985. Zero entries in contingency tables. *Comput. Statist. Data Anal.* **3**: 33–45.

Banerjee, C., M. Capozzoli, L. McSweeney, and D. Sinha. 1999. Beyond kappa: A review of interrater agreement measures. *Canad. J. Statist.* **27**: 3–23.

Baptista, J., and M. C. Pike. 1977. Algorithm AS115: Exact two-sided confidence limits for the odds ratio in a 2×2 table. *Appl. Statist.* **26**: 214–220.

Barnard, G. A. 1945. A new test for 2×2 tables. *Nature* **156**: 177.

Barnard, G. A. 1947. Significance tests for 2×2 tables. *Biometrika* **34**: 123–138.

Barnard, G. A. 1949. Statistical inference. *J. Roy. Statist. Soc. Ser B* **11**: 115–139.

Barnard, G. A. 1979. In contradiction to J. Berkson's dispraise: Conditional tests can be more efficient. *J. Statist. Plann. Inference* **3**: 181–188.

Barndorff-Nielsen, O. E., and B. Jörgensen. 1991. Some parametric models on the simplex. *J. Multivariate Anal.* **39**: 106–116.

Bartholomew, D. J. 1980. Factor analysis for categorical data. *J. Roy. Statist. Soc. Ser B* **42**: 293–321.

Bartholomew, D. J., and M. Knott. 1999. *Latent Variable Models and Factor Analysis*, 2nd ed. London: Edward Arnold.

Bartlett, M. S. 1935. Contingency table interactions. *J. Roy. Statist. Soc. Suppl.* **2**: 248–252.

Bartlett, M. S. 1937. Some examples of statistical methods of research in agriculture and applied biology. *J. Roy. Statist. Soc. Suppl.* **4**: 137–183.

Becker, M. 1989a. Models for the analysis of association in multivariate contingency tables. *J. Amer. Statist. Assoc.* **84**: 1014–1019.

Becker, M. 1989b. On the bivariate normal distribution and association models for ordinal categorical data. *Statist. Probab. Lett.* **8**: 435–440.

Becker, M. 1990. Maximum likelihood estimation of the RC(M) association model. *Appl. Statist.* **39**: 152–167.

Becker, M., and A. Agresti. 1992. Log-linear modelling of pairwise interobserver agreement on a categorical scale. *Statist. Medic.* **11**: 101–114.

Becker, M., and C. C. Clogg. 1989. Analysis of sets of two-way contingency tables using association models. *J. Amer. Statist. Assoc.* **84**: 142–151.

Bedrick, E. J. 1983. Chi-squared tests for cross-classified tables of survey data. *Biometrika* **70**: 591–595.

Bedrick, E. J. 1987. A family of confidence intervals for the ratio of two binomial proportions. *Biometrics* **43**: 993–998.

Begg, C. B., and R. Gray. 1984. Calculation of polytomous logistic regression parameters using individualized regressions. *Biometrika* **71**: 11–18.

Beitler, P. J., and J. R. Landis. 1985. A mixed-effects model for categorical data. *Biometrics* **41**: 991–1000.

Benedetti, J. K., and M. B. Brown. 1978. Strategies for the selection of loglinear models. *Biometrics* **34**: 680–686.

Benichou, J. 1998. Attributable risk. Pp. 216–229 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Benzécri, J.-P. 1973. *L'Analyse des Données*, Vol. 1, *La Taxonomie*; Vol. 2, *L'Analyse des Correspondances*. Paris: Dunod.

Berger, R., and D. D. Boos. 1994. p-Values maximized over a confidence set for the nuisance parameter. *J. Amer. Statist. Assoc.* **89**: 1012–1016.

Bergsma, W. P., and T. Rudas. 2002. Marginal models for categorical data. *Ann. Statist.* **30**: 140–159.

Berkson, J. 1938. Some difficulties of interpretation encountered in the application of the chi-square test. *J. Amer. Statist. Assoc.* **33**: 526–536.

Berkson, J. 1944. Application of the logistic function to bio-assay. *J. Amer. Statist. Assoc.* **39**: 357–365.

Berkson, J. 1951. Why I prefer logits to probits. *Biometrics* **7**: 327–339.

Berkson, J. 1953. A statistically precise and relatively simple method of estimating the bioassay with quantal response, based on the logistic function. *J. Amer. Statist. Assoc.* **48**: 565–599.

Berkson, J. 1955. Maximum likelihood and minimum logit χ^2 estimation of the logistic function. *J. Amer. Statist. Assoc.* **50**: 130–162.

Berkson, J. 1978. In dispraise of the exact test. *J. Statist. Plann. Inference* **2**: 27–42.

Berkson, J. 1980. Minimum chi-square, not maximum likelihood! *Ann. Statist.* **8**: 457–487.

Berry, G., and P. Armitage. 1995. Mid-*P* confidence intervals: A brief review. *The Statistician* **44**: 417–423.

Bhapkar, V. P. 1966. A note on the equivalence of two test criteria for hypotheses in categorical data. *J. Amer. Statist. Assoc.* **61**: 228–235.

Bhapkar, V. P. 1968. On the analysis of contingency tables with a quantitative response. *Biometrics* **24**: 329–338.

Bhapkar, V. P. 1973. On the comparison of proportions in matched samples. *Sankhya Ser A* **35**: 341–356.

Bhapkar, V. P. 1989. Conditioning on ancillary statistics and loss of information in the presence of nuisance parameters. *J. Statist. Plann. Inference.* **21**: 139–160.

Bhapkar, V. P., and G. G. Koch. 1968. On the hypothesis of "no interaction" in multidimensional contingency tables. *Biometrics* **24**: 567–594.

Bhapkar, V. P., and G. W. Somes. 1977. Distribution of Q when testing equality of matched proportions. *J. Amer. Statist. Assoc.* **72**: 658–661.

Biggeri, A. 1998. Negative binomial distribution. Pp. 2962–2967 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Billingsley, P. 1961. Statistical methods in Markov chains. *Ann. Math. Statist.* **32**: 12–40.

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

Birch, M. W. 1964a. A new proof of the Pearson–Fisher theorem. *Ann. Math. Statist.* **35**: 817–824.

Birch, M. W. 1964b. The detection of partial association I: The 2×2 case. *J. Roy. Statist. Soc. Ser. B* **26**: 313–324.

Birch, M. W. 1965. The detection of partial association II: The general case. *J. Roy. Statist. Soc. Ser B* **27**: 111–124.

Bishop, Y. M. M. 1971. Effects of collapsing multidimensional contingency tables. *Biometrics* **27**: 545–562.

Bishop, Y. M. M., and F. Mosteller. 1969. Smoothed contingency table analysis. Chap. IV-3 in *The National Halothane Study*. Washington, DC: U.S. Government Printing Office.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland. 1975. *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press.

Blaker, H. 2000. Confidence curves and improved exact confidence intervals for discrete distributions. *Canad. J. Statist.* **28**: 783–798.

Bliss, C. I. 1934. The method of probits. *Science* **79**: 38–39.

Bliss, C. I. 1935. The calculation of the dosage–mortality curve. *Ann. Appl. Biol.* **22**: 134–167.

Blyth, C. R. 1972. On Simpson's paradox and the sure-thing principle. *J. Amer. Statist. Assoc.* **67**: 364–366.

Blyth, C. R., and H. A. Still. 1983 Binomial confidence intervals. *J. Amer. Statist. Assoc.* **78**: 108–116.

Bock, R. D. 1970. Estimating multinomial response relations. Pp. 453–479 in *Contributions to Statistics and Probability*, ed. R. C. Bose. Chapel Hill, NC: University of North Carolina Press.

Bock, R. D., and M. Aitkin. 1981. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **46**: 443–459.

Bock, R. D., and L. V. Jones. 1968. *The Measurement and Prediction of Judgement and Choice*. San Francisco: Holden-Day.

Böckenholt, U., and W. Dillon. 1997. Modelling within-subject dependencies in ordinal paired comparison data. *Psychometrika* **62**: 411–434.

Bonney, G. E. 1987. Logistic regression for dependent binary observations. *Biometrics* **43**: 951–973.

Boos, D. D. 1992. On generalized score tests. *Amer. Statist.* **46**: 327–333.

Booth, J., and R. Butler. 1999. An importance sampling algorithm for exact conditional tests in log-linear models. *Biometrika* **86**: 321–332.

Booth, J. G., and J. P. Hobert. 1998. Standard errors of prediction in generalized linear mixed models. *J. Amer. Statist. Assoc.* **93**: 262–272.

Booth, J. G., and J. P. Hobert. 1999. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. Roy. Statist. Soc. Ser. B* **61**: 265–285.

Bowker, A. H. 1948. A test for symmetry in contingency tables. *J. Amer. Statist. Assoc.* **43**: 572–574.

Box, J. F. 1978. *R. A. Fisher*: *The Life of a Scientist*. New York: Wiley

Bradley, R. A. 1976. Science, statistics, and paired comparisons. *Biometrics* **32**: 213–240.

Bradley, R. A., and M. E. Terry. 1952. Rank analysis of incomplete block designs I. The method of paired comparisons. *Biometrika* **39**: 324–345.

Breslow, N. 1976. Regression analysis of the log odds ratio: A method for retrospective studies. *Biometrics* **32**: 409–416.

Breslow, N. 1981. Odds ratio estimators when the data are sparse. *Biometrika* **68**: 73–84.

Breslow, N. 1982. Covariance adjustment of relative-risk estimates in matched studies. *Biometrics* **38**: 661–672.

Breslow, N. 1984. Extra-Poisson variation in log-linear models. *Appl. Statist.* **33**: 38–44.

Breslow, N. 1996. Statistics in epidemiology: The case–control study. *J. Amer. Statist. Assoc.* **91**: 14–28.

Breslow, N., and D. G. Clayton. 1993. Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**: 9–25.

Breslow, N., and N. E. Day. 1980, 1987. *Statistical Methods in Cancer Research*, Vol. I, *The Analysis of Case–Control Studies*; Vol. II. *The Design and Analysis of Cohort Studies*. Lyon: IARC.

Breslow, N., and X. Lin. 1995. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82**: 81–91.

Breslow, N., and W. Powers. 1978. Are there two logistic regressions for retrospective studies? *Biometrics* **34**: 100–105.

Breslow, N., N. Day, K. Halvorsen, R. Prentice, and C. Sabai. 1978. Estimation of multiple relative risk functions in matched case–control studies. *Amer. J. Epidemiol.* **108**: 299–307.

Brier, S. S. 1980. Analysis of contingency tables under cluster sampling. *Biometrika* **67**: 591–596.

Brooks, S. P., B. J. T. Morgan, M. S. Ridout, and S. E. Pack. 1997. Finite mixture models for proportions. *Biometrics* **53**: 1097–1115.

Bross, I. D. J. 1958. How to use ridit analysis. *Biometrics* **14**: 18–38.

Brown, M. B. 1976. Screening effects in multidimensional contingency tables. *Appl. Statist.* **25**: 37–46.

Brown, M. B., and J. K. Benedetti. 1977. Sampling behavior of tests for correlation in two-way contingency tables. *J. Amer. Statist. Assoc.* **72**: 309–315.

Brown, P. J., and P. W. K. Rundell. 1985. Kernel estimates for categorical data. *Technometrics* **27**: 293–299.

Brown, L. D., T. T. Cai, and A. Das Gupta. 2001. Interval estimation for a binomial proportion. *Statist. Sci.* **16**: 101–133.

Brownstone, D., and K. F. Train. 1999. Forecasting new product penetration with flexible substitution patterns. *J. Econometrics* **89**: 109–129.

Bull, S. B., and A. Donner. 1987. The efficiency of multinomial logistic regression compared with multiple group discriminant analysis. *J. Amer. Statist. Assoc.* **82**: 1118–1122.

Burnham, K. P., and D. R. Anderson. 1998. *Model Selection and Inference: A Practical Information-Theoretic Approach*. New York: Springer-Verlag.

Burnham, K. P. and W. S. Overton. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* **65**: 625–633.

Burridge, J. 1981. A note on maximum likelihood estimation for regression models using grouped data. *J. Roy. Statist. Soc. Ser. B* **43**: 41–45.

Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge, U.K.: Cambridge University Press.

Carey, V., S. L. Zeger, and P. Diggle. 1993. Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80**: 517–526.

Carroll, R. J., S. Wang, and C. Y. Wang. 1995. Prospective analysis of logistic case–control pairs. *J. Amer. Statist. Assoc.* **90**: 157–169.

Casella, G., and R. Berger. 2001. *Statistical Inference*, 2nd ed. Pacific Grove, CA: Wadsworth.

Catalano, P. J., and L. M. Ryan. 1992. Bivariate latent variable models for clustered discrete and continuous outcomes. *J. Amer. Statist. Assoc.* **87**: 651–658.

Caussinus, H. 1966. Contribution à l'analyse statistique des tableaux de corrélation. *Ann. Fac. Sci. Univ. Toulouse* **29**: 77–182.

Chaloner, K., and K. Larntz. 1989. Optimal Bayesian design applied to logistic regression experiments. *J. Statist. Plann. Inference* **21**: 191–208.

Chamberlain, G. 1980. Analysis of covariance with qualitative data. *Rev. Econ. Stud.* **47**: 225–238.

Chambers, E. A., and D. R. Cox. 1967. Discrimination between alternative binary response models. *Biometrika* **54**: 573–578.

Chambers, R. L., and D. G. Steel. 2001. Simple methods for ecological inference in 2×2 tables. *J. Roy. Statist. Soc. Ser. A* **164**: 175–192.

Chan, I. 1998. Exact tests of equivalence and efficacy with non-zero lower bound for comparative studies. *Statist. Medic.* **17**: 1403–1413.

Chan, J. S. K., and A. Y. C. Kuk. 1997. Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics* **53**: 86–97.

Chao, A., P. K. Tsay, S.-H. Lin, W.-Y. Shau, and D.-Y. Chao. 2001. The applications of capture–recapture models to epidemiological data. *Statist. Medic.* **20**: 3123–3157.

Chapman, D. G., and R. C. Meng. 1966. The power of chi-square tests for contingency tables. *J. Amer. Statist. Assoc.* **61**: 965–975.

Chen, Z. and L. Kuo. 2001. A note on the estimation of the multinomial logit model with random effects. *Amer. Statist.* **55**: 89–95.

Christensen, R. 1997. *Log–Linear Models and Logistic Regression*. New York: Springer-Verlag.

Chuang, C., D. Gheva, and C. Odoroff. 1985. Methods for diagnosing multiplicative-interaction models for two-way contingency tables. *Commun. Statist. Ser. A* **14**: 2057–2080.

Clogg, C. C. 1995. Latent class models. Pp. 311–359 in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, ed. G. Arminger and C. C. Clogg. New York: Plenum Press.

Clogg, C. C., and S. R. Eliason. 1987. Some common problems in log-linear analysis. *Sociol. Methods Res.* **15**: 4–44.

Clogg, C. C., and L. A. Goodman. 1984. Latent structure analysis of a set of multidimensional contingency tables. *J. Amer. Statist. Assoc.* **79**: 762–771.

Clogg, C. C., and E. S. Shihadeh. 1994. *Statistical Models for Ordinal Variables*. Thousand Oaks, CA: Sage Publications.

Clopper, C. J., and E. S. Pearson. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**: 404–413.

Cochran, W. G. 1940. The analysis of variance when experimental errors follow the Poisson or binomial laws. *Ann. Math. Statist.* **11**: 335–347.

Cochran, W. G. 1943. Analysis of variance for percentages based on unequal numbers. *J. Amer. Statist. Assoc.* **38**: 287–301.

Cochran, W. G. 1950. The comparison of percentages in matched samples. *Biometrika* **37**: 256–266.

Cochran, W. G. 1952. The χ^2 test of goodness-of-fit. *Ann. Math. Statist.* **23**: 315–345.

Cochran, W. G. 1954. Some methods of strengthening the common χ^2 tests. *Biometrics* **10**: 417–451.

Cochran, W. G. 1955. A test of a linear function of the deviations between observed and expected numbers. *J. Amer. Statist. Assoc.* **50**: 377–397.

Coe, P. R., and A. C. Tamhane. 1993. Small sample confidence intervals for the difference, ratio and odds ratio of two success probabilities. *Commun. Statist. Ser. B* **22**: 925–938.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**: 37–46.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* **70**: 213–220.

Cohen, A., and H. B. Sackrowitz. 1991. Tests for independence in contingency tables with ordered alternatives. *J. Multivariate Anal.* **36**: 56–67.

Cohen, A., and H. B. Sackrowitz. 1992. An evaluation of some tests of trend in contingency tables. *J. Amer. Statist. Assoc.* **87**: 470–475.

Collett, D. 1991. *Modelling Binary Data*. London: Chapman & Hall.

Conaway, M. R. 1989. Analysis of repeated categorical measurements with conditional likelihood methods. *J. Amer. Statist. Assoc.* **84**: 53–62.

Cook, R. D., and S. Weisberg. 1999. *Applied Regression Including Computing and Graphics*. New York: Wiley.

Copas, J. B. 1973. Randomization models for the matched and unmatched 2×2 tables. *Biometrika* **60**: 467–476.

Copas, J. B. 1983. Plotting p against x. *Appl. Statist.* **32**: 25–31.

Copas, J. B. 1988. Binary regression models for contaminated data. *J. Roy. Statist. Soc. Ser B* **50**: 225–265.

Corcoran, C., L. Ryan, P. Senchaudhuri, C. Mehta, N. Patel, and G. Molenberghs. 2001. An exact trend test for correlated binary data. *Biometrics* **57**: 941–948.

Cormack, R. M. 1989. Log-linear models for capture–recapture. *Biometrics* **45**: 395–413.

Cornfield, J. 1951. A method of estimating comparative rates from clinical data: Applications to cancer of the lung, breast and cervix. *J. Natl. Cancer Inst.* **11**: 1269–1275.

Cornfield, J. 1956. A statistical problem arising from retrospective studies. In *Proc. 3rd Berkeley Symposium on Mathematics, Statistics and Probability*, ed. J. Neyman, **4**: 135–148.

Cornfield, J. 1962. Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: A discriminant function analysis. *Fed. Proc.* **21**, *Suppl.* **11**: 58–61.

Coull, B. A., and A. Agresti. 1999. The use of mixed logit models to reflect heterogeneity in capture–recapture studies. *Biometrics* **55**: 294–301.

Coull, B. A., and A. Agresti. 2000. Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. *Biometrics* **56**: 73–80.

Cox, C. 1984. An elementary introduction to maximum likelihood estimation for multinomial models: Birch's theorem and the delta method. *Amer. Statist.* **38**: 283–287.

Cox, C. 1995. Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach. *Statist. Medic.* **14**: 1191–1203.

Cox, C. 1996. Nonlinear quasi-likelihood models: Applications to continuous proportions. *Comput. Statist. Data Anal.* **21**: 449–461.

Cox, D. R. 1958a. The regression analysis of binary sequences. *J. Roy. Statist. Soc. Ser. B* **20**: 215–242.

Cox, D. R. 1958b. Two further applications of a model for binary regression. *Biometrika* **45**: 562–565.

Cox, D. R. 1970. *The Analysis of Binary Data* (2nd ed. 1989, by D. R. Cox and E. J. Snell). London: Chapman & Hall.

Cox, D. R. 1972. The analysis of multivariate binary data. *Appl. Statist.* **21**: 113–120.

Cox, D. R. 1983. Some remarks on overdispersion. *Biometrika* **70**: 269–274.

Cox, D. R., and D. V. Hinkley. 1974. *Theoretical Statistics*. London: Chapman & Hall.

Cramér, H. 1946. *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

Cressie, N., and T. R. C. Read. 1984. Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* **46**: 440–464.

Cressie, N., and T. R. C. Read. 1989. Pearson X^2 and the loglikelihood ratio statistic G^2: A comparative review. *Internat. Statist. Rev.* **57**: 19–43.

Croon, M., W. Bergsma, and J. Hagenaars. 2000. Analyzing change in categorical variables by generalized log-linear models. *Sociol. Methods Res.* **29**: 195–229.

Crouchley, R. 1995. A random-effects model for ordered categorical data. *J. Amer. Statist. Assoc.* **90**: 489–498.

Crowder, M. J. 1978. Beta-binomial ANOVA for proportions. *Appl. Statist.* **27**: 34–37.

D'Agostino, R. B., Jr. 1998. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statist. Medic.* **17**: 2265–2281.

Daniels, M. J., and C. Gatsonis. 1999. Hierarchical generalized linear models in the analysis of variations in health care utilization. *J. Amer. Statist. Assoc.* **94**: 29–42.

Dardanoni, V., and A. Forcina. 1998. A unified approach to likelihood inference on stochastic orderings in a nonparametric context. *J. Amer. Statist. Assoc.* **93**: 1112–1123.

Darroch, J. N. 1962. Interactions in multi-factor contingency tables. *J. Roy. Statist. Soc. Ser. B* **24**: 251–263.

Darroch, J. N. 1981. The Mantel–Haenszel test and tests of marginal symmetry; Fixed-effects and mixed models for a categorical response. *Internat. Statist. Rev.* **49**: 285–307.

Darroch, J. N., and P. I. McCloud. 1986. Category distinguishability and observer agreement. *Austral. J. Statist.* **28**: 371–388.

Darroch, J. N., and D. Ratcliff. 1972. Generalized iterative scaling for log-linear models. *Ann. Math. Statist.* **43**: 1470–1480.

Darroch, J. N., S. L. Lauritzen, and T. P. Speed. 1980. Markov fields and log-linear interaction models for contingency tables. *Ann. Statist.* **8**: 522–539.

Darroch, J. N., S. E. Fienberg, G. F. V. Glonek, and B. W. Junker. 1993. A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *J. Amer. Statist. Assoc.* **88**: 1137–1148.

Das Gupta, S., and M. D. Perlman. 1974. Power of the noncentral F-test: Effect of additional variates on Hotelling's T^2-test. *J. Amer. Statist. Assoc.* **69**: 174–180.

David, H. A. 1988. *The Method of Paired Comparisons*, 2nd ed. Oxford: Oxford University Press.

Davis, L. J. 1986a. Exact tests for 2 by 2 contingency tables. *Amer. Statist.* **40**: 139–141.

Davis, L. J. 1986b. Relationship between strictly collapsible and perfect tables. *Statist. Probab. Lett.* **4**: 119–122.

Davis, L. J. 1989. Intersection union tests for strictly collapsibility in three-dimensional contingency tables. *Ann. Statist.* **17**: 1693–1708.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application.* Cambridge, U.K. Cambridge University Press.

Dawson, R. B., Jr. 1954. A simplified expression for the variance of the χ^2-function on a contingency table. *Biometrika* **41**: 280.

Day, N. E., and D. P. Byar. 1979. Testing hypotheses in case–control studies: Equivalence of Mantel–Haenszel statistics and logit score tests. *Biometrics* **35**: 623–630.

de Falguerolles, A., S. Jmel, and J. Whittaker. 1995. Correspondence analysis and association models constrained by a conditional independence graph. *Psychometrika* **60**: 161–180.

Deming, W. E. 1964. *Statistical Adjustment of Data* (reprint of 1943 Wiley text). New York: Dover.

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

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**: 1–38.

Dey, D. K., S. K. Ghosh, and B. K. Mallick (editors). 2000. *Generalized Linear Models: A Bayesian Perspective*. New York: Marcel Dekker.

Diaconis, P., and B. Efron. 1985. Testing for independence in a two-way table: New interpretations of the chi-square statistic. *Ann. Statist.* **13**: 845–874.

Diaconis, P., and B. Sturmfels. 1998. Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.* **26**: 363–397.

Diggle, P. J., P. Heagerty, K.-Y. Liang, and S. L. Zeger. 2002. *Analysis of Longitudinal Data*, 2nd ed. Oxford: Clarendon Press.

Dittrich, R., R. Hatzinger, and W. Katzenbeisser. 1998. Modeling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Appl. Statist.* **47**: 511–525.

Dobson, A. J. 2001. *An Introduction to Generalized Linear Models*, 2*nd ed*. London: Chapman & Hall.

Dong, J. 1998. Simpson's paradox. Pp. 4108–4110 in *Encyclopedia of Biostatistics*, Vol. 5. Chichester, UK: Wiley.

Dong, J., and J. S. Simonoff. 1994. The construction and properties of boundary kernels for smoothing sparse multinomials. *J. Computat. Graph. Statist.* **3**: 57–66.

Dong, J., and J. S. Simonoff. 1995. A geometric combination estimator for d-dimensional ordinal sparse contingency tables. *Ann. Statist.* **23**: 1143–1159.

Donner, A., and W. W. Hauck. 1986. The large-sample efficiency of the Mantel–Haenszel estimator in the fixed-strata case. *Biometrics* **42**: 537–545.

Doolittle, M. H. 1888. Association ratios. *Bull. Philos. Soc. Washington* **10**: 83–87, 94–96.

Drost, F. C., W. C. M. Kallenberg, D. S. Moore, and J. Oosterhoff. 1989. Power approximations to multinomial tests of fit. *J. Amer. Statist. Assoc.* **84**: 130–141.

Ducharme, G. R., and Y. Lepage. 1986. Testing collapsibility in contingency tables. *J. Roy. Statist. Soc. Ser B* **48**: 197–205.

Dupont, W. D. 1986. Sensitivity of Fisher's exact test to minor perturbations in 2×2 contingency tables. *Statist. Medic.* **5**: 629–635.

Dyke, G. V., and H. D. Patterson. 1952. Analysis of factorial arrangements when the data are proportions. *Biometrics* **8**: 1–12.

Edwardes, M. D. deB. 1997. Univariate random cut-points theory for the analysis of ordered categorical data. *J. Amer. Statist. Assoc.* **92**: 1114–1123.

Edwards, A. W. F. 1963. The measure of association in a 2×2 table. *J. Roy. Statist. Soc. Ser A* **126**: 109–114.

Edwards, D. 2000. *Introduction to Graphical Modelling*, 2nd ed. New York: Springer-Verlag.

Edwards, D., and S. Kreiner. 1983. The analysis of contingency tables by graphical models. *Biometrika* **70**: 553–565.

Efron, B. 1975. The efficiency of logistic regression compared to normal discriminant analysis. *J. Amer. Statist. Assoc.* **70**: 892–898.

Efron, B. 1978. Regression and ANOVA with zero–one data: Measures of residual variation. *J. Amer. Statist. Soc.* **73**: 113–121.

Efron, B., and D. V. Hinkley. 1978. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65**: 457–482.

Efron, B., and C. Morris. 1975. Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70**: 311–319.

Ekholm, A., J. W. McDonald, and P. W. F. Smith. 2000. Association models for a multivariate binary response. *Biometrics* **56**: 712–718.

Escoufier, Y. 1982. L'analyse des tableaux de contingence simples et multiples. In *Proc. International Meeting on the Analysis of Multidimensional Contingency Tables* (Rome, 1981), ed. R. Coppi. *Metron* **40**: 53–77.

Espeland, M. A., and S. L. Handelman. 1989. Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics* **45**: 587–599.

Fahrmeir, L., and G. Tutz. 2001. *Multivariate Statistical Modelling based on Generalized Linear Models*, 2nd ed. New York: Springer-Verlag.

Farewell, V. T. 1979. Some results on the estimation of logistic models based on retrospective data. *Biometrika* **66**: 27–32.

Farewell, V. T. 1982. A note on regression analysis of ordinal data with variability of classification. *Biometrika* **69**: 533–538.

Fay, R. 1985. A jackknifed chi-squared test for complex samples. *J. Amer. Statist. Assoc.* **80**: 148–157.

Fay, R. 1986. Causal models for patterns of nonresponse. *J. Amer. Statist. Assoc.* **81**: 354–365.

Ferguson, T. S. 1967. *Mathematical Statistics*: *A Decision Theoretic Approach*. New York: Academic Press.

Fienberg, S. E. 1970a. An iterative procedure for estimation in contingency tables. *Ann. Math. Statist.* **41**: 907–917.

Fienberg, S. E. 1970b. Quasi-independence and maximum likelihood estimation in incomplete contingency tables. *J. Amer. Statist. Soc.* **65**: 1610–1616.

Fienberg, S. E. 1972. The analysis of incomplete multi-way contingency tables. *Biometrics* **28**: 177–202.

Fienberg, S. E. 1980. Fisher's contributions to the analysis of categorical data. Pp. 75–84 in *R. A. Fisher*: *An Appreciation*, ed. S. E. Fienberg and D. V. Hinkley. Berlin: Springer-Verlag.

Fienberg, S. E. 1984. The contributions of William Cochran to categorical data analysis. Pp. 103–118 in *W. G. Cochran's Impact on Statistics*, ed. P. S. R. S. Rao and J. Sedransk. New York: Wiley.

Fienberg, S. E., and P. W. Holland. 1973. Simultaneous estimation of multinomial cell probabilities. *J. Amer. Statist. Assoc.* **68**: 683–690.

Fienberg, S. E., and K. Larntz. 1976. Loglinear representation for paired and multiple comparison models. *Biometrika* **63**: 245–254.

Fienberg, S. E., M. A. Johnson, and B. J. Junker. 1999. Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *J. Roy. Statist. Soc. Ser. A* **162**: 383–405.

Finney, D. J. 1947. The estimation from individual records of the relationship between dose and quantal response. *Biometrika* **34**: 320–334.

Finney, D. J. 1971. *Probit Analysis*, 3rd ed. Cambridge: Cambridge University Press.

Firth, D. 1987. On the efficiency of quasi-likelihood estimation. *Biometrika* **74**: 233–245.

Firth, D. 1989. Marginal homogeneity and the superposition of Latin squares. *Biometrika* **76**: 179–182.

Firth, D. 1991. Generalized linear models. Pp. 55–82 in *Statistical Theory and Modelling. In Honour of Sir David Cox, FRS*, D. V. Hinkley, N. Reid, and E. J. Snell, eds. London: Chapman & Hall.

Firth, D. 1993a. Bias reduction of maximum likelihood estimates. *Biometrika* **80**: 27–38.

Firth, D. 1993b. Recent developments in quasi-likelihood methods. *Proc. ISI 49th Session*, pp. 341–358.

Firth, D., and J. Kuha. 2000. On the index of dissimilarity for lack of fit in log linear models. Unpublished manuscript.

Fischer, G. H., and I. W. Molenaar. 1995. *Rasch Models: Foundations, Recent Developments, and Applications*. New York: Springer-Verlag.

Fisher, R. A. 1922. On the interpretation of chi-square from contingency tables, and the calculation of *P*. *J. Roy. Statist. Soc.* **85**: 87–94.

Fisher, R. A. 1924. The conditions under which chi-square measures the discrepancy between observation and hypothesis. *J. Roy. Statist. Soc.* **87**: 442–450.

Fisher, R. A. 1926. Bayes' theorem and the fourfold table. *Eugenics Rev.* **18**: 32–33.

Fisher, R. A. 1934, 1970. *Statistical Methods for Research Workers* (originally published 1925, 14th ed., 1970.) Edinburgh: Oliver & Boyd.

Fisher, R. A. 1935a. *The Design of Experiments* (8th ed., 1966). Edinburgh: Oliver & Boyd.

Fisher, R. A. 1935b. Appendix to article by C. Bliss. *Ann. Appl. Biol.* **22**: 164–165.

Fisher, R. A. 1935c. The logic of inductive inference. *J. Roy. Statist. Soc.* **98**: 39–82.

Fisher, R. A. 1945. A new test for 2×2 tables (Letter to the Editor). *Nature* **156**: 388.

Fisher, R. A. 1956. *Statistical Methods for Scientific Inference*. Edinburgh: Oliver & Boyd.

Fisher, R. A., and F. Yates. 1938. *Statistical Tables*. Edinburgh: Oliver and Boyd.

Fitzmaurice, G. M., and N. M. Laird. 1993. A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80**: 141–151.

Fitzmaurice, G. M., N. M. Laird, and S. Lipsitz. 1994. Analysing incomplete longitudinal binary responses: A likelihood-based approach. *Biometrics* **50**: 601–612.

Fitzmaurice, G. M., N. M. Laird, and A. G. Rotnitzky. 1993. Regression models for discrete longitudinal responses. *Statist. Sci.* **8**: 284–299.

Fitzpatrick, S., and A. Scott. 1987. Quick simultaneous confidence intervals for multinomial proportions. *J. Amer. Statist. Assoc.* **82**: 875–878.

Fleiss, J. L. 1981. *Statistical Methods for Rates and Proportions*, 2nd ed. New York: Wiley.

Fleiss, J. L., and J. Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Meas.* **33**: 613–619.

Fleiss, J. L., J. Cohen, and B. S. Everitt. 1969. Large-sample standard errors of kappa and weighted kappa. *Psychol. Bull.* **72**: 323–327.

Follman, D. A., and D. Lambert. 1989. Generalizing logistic regression by nonparametric mixing. *J. Amer. Statist. Assoc.* **84**: 295–300.

Forster, J. J., and P. W. F. Smith. 1998. Model-based inference for categorical survey data subject to non-ignorable non-response. *J. Roy. Statist. Soc. Ser B* **60**: 57–70.

Forster, J. J., J. W. McDonald, and P. W. F. Smith. 1996. Monte Carlo exact conditional tests for log-linear and logistic models. *J. Roy. Statist. Soc. Ser B* **58**: 445–453.

Fowlkes, E. B. 1987. Some diagnostics for binary logistic regression via smoothing. *Biometrika* **74**: 503–515.

Fowlkes, E. B., A. E. Freeny, and J. Landwehr. 1988. Evaluating logistic models for large contingency tables. *J. Amer. Statist. Assoc.* **83**: 611–622.

Freedman, D., R. Pisani, and R. Purves. 1978. *Statistics*. New York: W. W. Norton.

Freeman, G. H., and J. H. Halton. 1951. Note on an exact treatment of contingency, goodness-of-fit and other problems of significance. *Biometrika* **38**: 141–149.

Freeman, D. H., Jr. and T. R. Holford. 1980. Summary rates. *Biometrics* **36**: 195–205.

Freeman, M. F., and J. W. Tukey. 1950. Transformations related to the angular and the square root. *Ann. Math. Statist.* **21**: 607–611.

Freidlin, B., and J. L. Gastwirth. 1999. Unconditional versions of several tests commonly used in the analysis of contingency tables. *Biometrics* **55**: 264–267.

Friendly, M. 2000. *Visualizing Categorical Data*. Cary, NC: SAS Institute.

Frome, E. L. 1983. The analysis of rates using Poisson regression models. *Biometrics* **39**: 665–674.

Fuchs, C. 1982. Maximum likelihood estimation and model selection in contingency tables with missing data. *J. Amer. Statist. Assoc.* **77**: 270–278.

Gabriel, K. R. 1966. Simultaneous test procedures for multiple comparisons on categorical data. *J. Amer. Statist. Assoc.* **61**: 1081–1096.

Gabriel, K. R. 1971. The biplot graphic display of matrices with applications to principal component analysis. *Biometrika* **58**: 453–467.

Gail, M. H., and J. J. Gart. 1973. The determination of sample sizes for use with the exact conditional test in 2×2 comparative trials. *Biometrics* **29**: 441–448.

Gail, M., and N. Mantel. 1977. Counting the number of $r \times c$ contingency tables with fixed margins. *J. Amer. Statist. Assoc.* **72**: 859–862.

Gart, J. J. 1966. Alternative analyses of contingency tables. *J. Roy. Statist. Soc. Ser B* **28**: 164–179.

Gart, J. J. 1969. An exact test for comparing matched proportions in crossover designs. *Biometrika* **56**: 75–80.

Gart, J. J. 1970. Point and interval estimation of the common odds ratio in the combination of 2×2 tables with fixed margins. *Biometrika* **57**: 471–475.

Gart, J. J. 1971. The comparison of proportions: A review of significance tests, confidence intervals and adjustments for stratification. *Rev. Internat. Statist. Rev.* **39**: 148–169.

Gart, J. J., and J. Nam. 1988. Approximate interval estimation of the ratio of binomial parameters: A review and corrections for skewness. *Biometrics* **44**: 323–338.

Gart, J. J., and J. R. Zweiful. 1967. On the bias of various estimators of the logit and its variance with applications to quantal bioassay. *Biometrika* **54**: 181–187.

Gelfand, A. E., and A. F. Smith. 1990. Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**: 398–409.

Genter, F. C., and V. T. Farewell. 1985. Goodness-of-link testing in ordinal regression models. *Canad. J. Statist.* **13**: 37–44.

Ghosh, B. K. 1979. A comparison of some approximate confidence intervals for the binomial parameter. *J. Amer. Statist. Assoc.* **74**: 894–900.

Ghosh, M., M. Chen, A. Ghosh, and A. Agresti. 2000. Hierarchical Bayesian analysis of binary matched pairs data. *Statist. Sin.* **10**: 647–657.

Gibbons, R. D., and D. Hedeker. 1997. Random-effects probit and logistic regression models for three-level data. *Biometrics* **53**: 1527–1537.

Gill, J. 2000. *Generalized Linear Models*: *A Unified Approach*. Thousand Oaks, CA: Sage Publications.

Gilmour, A. R., R. D. Anderson, and A. L. Rae. 1985. The analysis of binomial data by a generalized linear mixed model. *Biometrika* **72**: 593–599.

Gilula, Z., and S. Haberman. 1986. Canonical analysis of contingency tables by maximum likelihood. *J. Amer. Statist. Assoc.* **81**: 780–788.

Gilula, Z., and S. Haberman. 1988. The analysis of multivariate contingency tables by restricted canonical and restricted association models. *J. Amer. Statist. Assoc.* **83**: 760–771.

Gilula, Z., and S. Haberman. 1998. Chi-square, partition of. Pp. 622–627 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Gleser, L. J., and D. S. Moore. 1985. The effect of positive dependence on chi-squared tests for categorical data. *J. Roy. Statist. Soc. Ser B* **47**: 459–465.

Glonek, G. 1996. A class of regression models for multivariate categorical responses. *Biometrika* **83**: 15–28.

Glonek, G. F. V., and P. McCullagh. 1995. Multivariate logistic models. *J. Roy. Statist. Soc. Ser. B* **57**: 533–546.

Glonek, G., J. N. Darroch, and T. P. Speed. 1988. On the existence of maximum likelihood estimators for hierarchical loglinear models. *Scand. J. Statist.* **15**: 187–193.

Gokhale, D. V., and S. Kullback. 1978. *The Information in Contingency Tables*. New York: Marcel Dekker.

Goldstein, H. 1995. *Multilevel Statistical Models*, 2nd ed. London: Edward Arnold.

Goldstein, H., and J. Rasbash. 1996. Improved approximations for multilevel models with binary responses. *J. Roy. Statist. Soc. Ser A* **159**: 505–513.

Good, I. J. 1963. Maximum entropy for hypothesis formulation, especially for multi-dimensional contingency tables. *Ann. Math. Statist.* **34**: 911–934.

Good, I. J. 1965. *The Estimation of Probabilities*: *An Essay on Modern Bayesian Methods*. Cambridge, MA: MIT Press.

Good, I. J. 1976. On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Statist.* **4**: 1159–1189.

Good, I. J., and R. A. Gaskins. 1971. Nonparametric roughness penalties for probability densities. *Biometrika* **58**: 255–277.

Good, I. J., and Y. Mittal. 1987. The amalgamation and geometry of two-by-two contingency tables. *Ann. Statist.* **15**: 694–711.

Good, I. J., T. N. Gover, and G. J. Mitchell. 1970. Exact distributions for χ^2 and for the likelihood-ratio statistic for the equiprobable multinomial distribution. *J. Amer. Statist. Assoc.* **65**: 267–283.

Goodman, L. A. 1964a. Simultaneous confidence intervals for cross-product ratios in contingency tables. *J. Roy. Statist. Soc. Ser B* **26**: 86–102.

Goodman, L. A. 1964b. Interactions in multi-dimensional contingency tables. *Ann. Math. Statist.* **35**: 632–646.

Goodman, L. A. 1965. On simultaneous confidence intervals for multinomial proportions. *Technometrics* **7**: 247–254.

Goodman, L. A. 1968. The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries. *J. Amer. Statist. Assoc.* **63**: 1091–1131.

Goodman, L. A. 1969a. On partitioning chi-square and detecting partial association in three-way contingency tables. *J. Roy. Statist. Soc. Ser B* **31**: 486–498.

Goodman, L. A. 1969b. How to ransack social mobility tables and other kinds of cross-classification tables. *Amer. J. Sociol.* **75**: 1–40.

Goodman, L. A. 1970. The multivariate analysis of qualitative data: Interaction among multiple classifications. *J. Amer. Statist. Assoc.* **65**: 226–256.

Goodman, L. A. 1971a. The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics* **13**: 33–61.

Goodman, L. A. 1971b. The partitioning of chi-square, the analysis of marginal contingency tables, and the estimation of expected frequencies in multidimensional contingency tables. *J. Amer. Statist. Assoc.* **66**: 339–344.

Goodman, L. A. 1973. The analysis of multidimensional contingency tables with some variables are posterior to others: A modified path analysis approach. *Biometrika* **60**: 179–192.

Goodman, L. A. 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61**: 215–231.

Goodman, L. A. 1979a. Simple models for the analysis of association in cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* **74**: 537–552.

Goodman, L. A. 1979b. Multiplicative models for square contingency tables with ordered categories. *Biometrika* **66**: 413–418.

Goodman, L. A. 1981a. Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* **76**: 320–334.

Goodman, L. A. 1981b. Association models and the bivariate normal for contingency tables with ordered categories. *Biometrika* **68**: 347–355.

Goodman, L. A. 1983. The analysis of dependence in cross-classification having ordered categories, using log-linear models for frequencies and log-linear models for odds. *Biometrics* **39**: 149–160.

Goodman, L. A. 1985. The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Ann. Statist.* **13**: 10–69.

Goodman, L. A. 1986. Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *Internat. Statist. Rev.* **54**: 243–309.

Goodman, L. A. 1996. A single general method for the analysis of cross-classified data: Reconciliation and synthesis of some methods of Pearson, Yule, and Fisher, and also some methods of correspondence analysis and association analysis. *J. Amer. Statist. Assoc.* **91**: 408–427.

Goodman, L. A. 2000. The analysis of cross-classified data: Notes on a century of progress in contingency table analysis, and some comments on its prehistory and its future. Pp. 189–231 in *Statistics for the 21st Century*, ed. C. R. Rao and G. J. Székely. New York: Marcel Dekker.

Goodman, L. A., and W. H. Kruskal. 1979. *Measures of Association for Cross Classifications*. New York: Springer-Verlag (contains articles appearing in *J. Amer. Statist. Assoc.* in 1954, 1959, 1963, 1972).

Gould, S. J. 1981. *The Mismeasure of Man*. New York: W. W. Norton.

Gourieroux, C., A. Monfort, and A. Trognon. 1984. Pseudo maximum likelihood methods: Theory. *Econometrica* **52**: 681–700.

Graubard, B. I., and E. L. Korn. 1987. Choice of column scores for testing independence in ordered $2 \times K$ contingency tables. *Biometrics* **43**: 471–476.

Green, P. J. 1984. Iteratively weighted least squares for maximum likelihood estimation and some robust and resistant alternatives. *J. Roy. Statist. Soc. Ser B* **46**: 149–192.

Greenacre, M. J. 1993. *Correspondence Analysis in Practice*. New York: Academic Press.

Greenland, S. 1991. On the logical justification of conditional tests for two-by-two contingency tables. *Amer. Statist.* **45**: 248–251.

Greenland, S., and J. M. Robins. 1985. Estimation of a common effect parameter from sparse follow-up data. *Biometrics* **41**: 55–68.

Greenwood, M., and G. U. Yule. 1920. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *J. Roy. Statist. Soc. Ser A* **83**: 255–279.

Greenwood, P. E., and M. S. Nikulin. 1996. *A Guide to Chi-Squared Testing*. New York: Wiley.

Grizzle, J. E., C. F. Starmer, and G. G. Koch. 1969. Analysis of categorical data by linear models. *Biometrics* **25**: 489–504.

Gross, S. T. 1981. On asymptotic power and efficiency of tests of independence in contingency tables with ordered classifications. *J. Amer. Statist. Assoc.* **76**: 935–941.

Gueorguieva, R., and A. Agresti. 2001. A correlated probit model for joint modeling of clustered binary and continuous responses. *J. Amer. Statist. Assoc.* **96**: 1102–1112.

Haber, M. 1980. A comparison of some continuity corrections for the chi-squared test on 2×2 tables. *J. Amer. Statist. Assoc.* **75**: 510–515.

Haber, M. 1982. The continuity correction and statistical testing. *Internat. Statist. Rev.* **50**: 135–144.

Haber, M. 1985. Maximum likelihood methods for linear and log-linear models in categorical data. *Comput. Statist. Data Anal.* **3**: 1–10.

Haber, M. 1986. An exact unconditional test for the 2×2 comparative trial. *Psychol. Bull.* **99**: 129–132.

Haber, M. 1989. Do the marginal totals of a 2×2 contingency table contain information regarding the table proportions? *Commun. Statist. Ser A* **18**: 147–156.

Haberman, S. J. 1973a. The analysis of residuals in cross-classification tables. *Biometrics* **29**: 205–220.

Haberman, S. J. 1973b. Log-linear models for frequency data: Sufficient statistics and likelihood equations. *Ann. Statist.* **1**: 617–632.

Haberman, S. J. 1974a. *The Analysis of Frequency Data*. Chicago: University of Chicago Press.

Haberman, S. J. 1974b. Log-linear models for frequency tables with ordered classifications. *Biometrics* **36**: 589–600.

Haberman, S. J. 1977a. Log-linear models and frequency tables with small expected cell counts. *Ann. Statist.* **5**: 1148–1169.

Haberman, S. J. 1977b. Maximum likelihood estimation in exponential response models. *Ann. Statist.* **5**: 815–841.

Haberman, S. J. 1978, 1979. *Analysis of Qualitative Data*, Vols. 1 and 2. New York: Academic Press.

Haberman, S. J. 1981. Tests for independence in two-way contingency tables based on canonical correlation and on linear-by-linear interaction. *Ann. Statist.* **9**: 1178–1186.

Haberman, S. J. 1982. The analysis of dispersion of multinomial responses. *J. Amer. Statist. Assoc.* **77**: 568–580.

Haberman, S. J. 1988. A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *J. Amer. Statist. Assoc.* **83**: 555–560.

Haberman, S. J. 1995. Computation of maximum likelihood estimates in association models. *J. Amer. Statist. Assoc.* **90**: 1438–1446.

Hagenaars, J. A. 1998. Categorical causal modeling: Latent class analysis and directed log-linear models with latent variables. *Sociol. Methods Res.* **26**: 436–486.

Hald, A. 1998. *A History of Mathematical Statistics from* 1750 *to* 1930. New York: Wiley.

Haldane, J. B. S. 1940. The mean and variance of χ^2, when used as a test of homogeneity, when expectations are small. *Biometrika* **31**: 346–355.

Haldane, J. B. S. 1956. The estimation and significance of the logarithm of a ratio of frequencies. *Ann. Human Genet.* **20**: 309–311.

Hall, P., and D. M. Titterington. 1987. On smoothing sparse multinomial data. *Austral. J. Statist.* **29**: 19–37.

Hamada, M., and C. F. J. Wu. 1990. A critical look at accumulation analysis and related methods. *Technometrics* **32**: 119–130.

Hansen, L. P. 1982. Large sample properties of generalized-method of moments estimators. *Econometrica* **50**: 1029–1054.

Harkness, W. L., and L. Katz. 1964. Comparison of the power functions for the test of independence in 2×2 contingency tables. *Ann. Math. Statist.* **35**: 1115–1127.

Harrell F. E., R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. 1982. Evaluating the yield of medical tests. *J. Amer. Medic. Assoc.* **247**: 2543–2546.

Hartzel, J., I.-M. Liu, and A. Agresti. 2001a. Describing heterogeneous effects in stratified ordinal contingency tables, with application to multi-center clinical trials. *Computat. Statist. Data Anal.* **35**: 429–449.

Hartzel, J., A. Agresti, and B. Caffo. 2001b. Multinomial logit random effects models. *Statistical Modelling* **1**: 81–102.

Haslett, S. 1990. Degrees of freedom and parameter estimability in hierarchical models for sparse complete contingency tables. *Computat. Statist. Data Anal.* **9**: 179–195.

Hastie, T., and R. Tibshirani. 1987. Non-parametric logistic and proportional odds regression. *Appl. Statist.* **36**: 260–276.

Hastie, T., and R. Tibshirani. 1990. *Generalized Additive Models*. London: Chapman & Hall.

Hatzinger, R. 1989. The Rasch model, some extensions and their relation to the class of generalized linear models. *Statistical Modelling*: *Lecture Notes in Statistics*, Vol. 57. Berlin: Springer-Verlag.

Hauck, W. W. 1979. The large sample variance of the Mantel–Haenszel estimator of a common odds ratio. *Biometrics* **35**: 817–819.

Hauck, W. W. 1983. A note on confidence bands for the logistic response curve. *Amer. Statist.* **37**: 158–160.

Hauck, W. W., and A. Donner. 1977. Wald's test as applied to hypotheses in logit analysis. *J. Amer. Statist. Assoc.* **72**: 851–853.

Heagerty, P. J. 1999. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics* **55**: 688–698.

Heagerty, P. J., and S. L. Zeger. 1996. Marginal regression models for clustered ordinal measurements. *J. Amer. Statist. Assoc.* **91**: 1024–1036.

Heagerty, P. J., and S. L. Zeger. 2000. Marginalized multilevel models and likelihood inference. *Statist. Sci.* **15**: 1–19.

Hedeker, D., and R. D. Gibbons. 1994. A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50**: 933–944.

Heinen, T. 1996. *Latent Class and Discrete Latent Trait Models*. Thousand Oaks, CA: Sage Publications.

Heyde, C. C. 1997. *Quasi-likelihood and Its Application.* New York: Springer-Verlag.

Hinde, J. 1982. Compound Poisson regression models. Pp. 109–121 in *GLIM* 82: *Proc. International Conference on Generalised Linear Models*, ed. R. Gilchrist. New York: Springer-Verlag.

Hinde, J., and C. G. B. Demétrio. 1998. Overdispersion: Models and estimation. *Comput. Statist. Data Anal.* **27**: 151–170.

Hirji, K. F. 1991. A comparison of exact, mid-*P*, and score tests for matched case-control studies. *Biometrics* **47**: 487–496.

Hirji, K. F., C. R. Mehta, and N. R. Patel. 1987. Computing distributions for exact logistic regression. *J. Amer. Statist. Assoc.* **82**: 1110–1117.

Hirotsu, C. 1982. Use of cumulative efficient scores for testing ordered alternatives in discrete models. *Biometrika* **69**: 567–577.

Hirschfeld, H. O. 1935. A connection between correlation and contingency. *Cambridge Philos. Soc. Proc.* (*Math. Proc.*) **31**: 520–524.

Hodges, J. L., Jr. 1958. Fitting the logistic by maximum likelihood. *Biometrics* **14**: 453–461.

Hoem, J. M. 1987. Statistical analysis of a multiplicative model and its application to the standardization of vital rates: A review. *Internat. Statist. Rev.* **5**: 119–152.

Holford, T. R. 1980. The analysis of rates and of survivorship using log-linear models. *Biometrics* **36**: 299–305.

Holt, D., A. J. Scott, and P. D. Ewings. 1980. Chi-squared tests with survey data. *J. Roy. Statist. Soc. Ser. A* **143**: 303–320.

Hook, E. B., and R. R. Regal. 1995. Capture–recapture methods in epidemiology: Methods and limitations. *Epidemiol. Rev.* **17**: 243–264.

Hosmer, D. W., and S. Lemeshow. 1980. A goodness-of-fit test for multiple logistic regression model. *Commun. Statist. Ser A* **9**: 1043–1069.

Hosmer, D. W., and S. Lemeshow. 2000. *Applied Logistic Regression*, 2nd ed. New York: Wiley.

Hosmer, D. W., T. Hosmer, S. le Cessie, and S. Lemeshow. 1997. A comparison of goodness-of-fit tests for the logistic regression model. *Statist. Medic.* **16**: 965–980.

Hout, M., O. D. Duncan, and M. E. Sobel. 1987. Association and heterogeneity: Structural models of similarities and differences. *Sociol. Methodol.* **17**: 145–184.

Howard, J. V. 1998. The 2×2 table: A discussion from a Bayesian viewpoint. *Statist. Sci.* **13**: 351–367.

Hsieh, F. Y. 1989. Sample size tables for logistic regression. *Statist. Medic.* **8**: 795–802.

Hsieh, F. Y., D. A. Bloch, and M. D. Larsen. 1998. A simple method of sample size calculation for linear and logistic regression. *Statist. Medic.* **17**: 1623–1634.

Hwang, J. T. G., and M. T. Wells. 2002. Optimality results for mid *P*-values. To appear.

Hwang, J. T. G., and M.-C. Yang. 2001. An optimality theory for mid *P*-values in 2×2 contingency tables. *Statist. Sin.* **11**: 807–826.

Imrey, P. B. 1998. Bradley–Terry model. Pp. 437–443 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Imrey, P. B., W. D. Johnson, and G. G. Koch. 1976. An incomplete contingency table approach to paired-comparison experiments. *J. Amer. Statist. Assoc.* **71**: 614–623.

Imrey, P. B., G. G. Koch, and M. E. Stokes. 1981. Categorical data analysis: Some reflections on the log linear model and logistic regression. I: Historical and methodological overview. *Internat. Statist. Rev.* **49**: 265–283.

Ireland, C. T., and S. Kullback. 1968a. Minimum discrimination information estimation. *Biometrics* **24**: 707–713.

Ireland, C. T., and S. Kullback. 1968b. Contingency tables with given marginals. *Biometrika* **55**: 179–188.

Ireland, C. T., H. H. Ku, and S. Kullback. 1969. Symmetry and marginal homogeneity of an $r \times r$ contingency table. *J. Amer. Statist. Assoc.* **64**: 1323–1341.

Irwin, J. O. 1935. Tests of significance for differences between percentages based on small numbers. *Metron* **12**: 83–94.

Jennison, C., and B. W. Turnbull. 2000. *Group Sequential Methods with Applications to Clinical Trials*. London: Chapman & Hall.

Johnson, B. M. 1971. On the admissible estimators for certain fixed sample binomial problems. *Ann. Math. Statist.* **42**: 1579–1587.

Johnson, W. 1985. Influence measures for logistic regression: Another point of view. *Biometrika* **72**: 59–65.

Johnson, N. L., S. Kotz, and A. W. Kemp. 1992. *Univariate Discrete Distributions*, 2nd ed. New York: Wiley.

Jones, B., and M. G. Kenward. 1987. Modelling binary data from a three-period cross-over trial. *Statist. Medic.* **6**: 555–564.

Jones, M. P., T. W. O'Gorman, J. H. Lemke, and R. F. Woolson. 1989. A Monte Carlo investigation of homogeneity tests of the odds ratio under various sample size considerations. *Biometrics* **45**: 171–181.

Jørgensen, B. 1983. Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika* **70**: 19–28.

Jørgensen, B. 1987. Exponential dispersion models. *J. Roy. Statist. Soc. Ser. B* **49**: 127–162.

Kalbfleisch, J. D., and J. F. Lawless. 1985. The analysis of panel data under a Markov assumption. *J. Amer. Statist. Assoc.* **80**: 863–871.

Kastner, C., A. Fieger, and C. Heumann. 1997. MAREG and WinMAREG: A tool for marginal regression models. *Comput. Statist. Data Anal.* **24**: 237–241.

Kauermann, G., and R. J. Carroll, 2001. A note on the efficiency of sandwich covariance matrix estimation. *J. Amer. Statist. Assoc.* **96**: 1387–1397.

Kauermann, G., and G. Tutz. 2001. Testing generalized linear and semiparametric models against smooth alternatives. *J. Roy. Statist. Soc. Ser. B* **63**: 147–166.

Kelderman, H. 1984. Loglinear Rasch model tests. *Psychometrika* **49**: 223–245.

Kempthorne, O. 1979. In dispraise of the exact test: Reactions. *J. Statist. Plann. Inference* **3**: 199–213.

Kendall, M. G. 1945. The treatment of ties in rank problems. *Biometrika* **33**: 239–251.

Kendall, M., and A. Stuart. 1979. *The Advanced Theory of Statistics*, Vol. 2; *Inference and Relationship*, 4th ed. New York: Macmillan.

Kenward, M. G., and B. Jones. 1991. The analysis of categorical data from cross-over trials using a latent variable model. *Statist. Medic.* **10**: 1607–1619.

Kenward, M. G., and B. Jones. 1994. The analysis of binary and categorical data from crossover trials. *Statist. Methods Medic. Res.* **3**: 325–344.

Kenward, M. G., E. Lesaffre, and G. Molenberghs. 1994. An application of maximum likelihood and estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics* **50**: 945–953.

Khamis, H. J. 1983. Log-linear model analysis of the semi-symmetric intraclass contingency table. *Commun. Statist. Ser. A* **12**: 2723–2752.

Kim, D., and A. Agresti. 1995. Improved exact inference about conditional association in three-way contingency tables. *J. Amer. Statist. Assoc.* **90**: 632–639.

Kim, D., and A. Agresti. 1997. Nearly exact tests of conditional independence and marginal homogeneity for sparse contingency tables. *Comput. Statist. Data Anal.* **24**: 89–104.

King, G. 1997. *A Solution to the Ecological Inference Problem*. Princeton, NJ: Princeton University Press.

Knuiman, M. W., and T. P. Speed. 1988. Incorporating prior information into the analysis of contingency tables. *Biometrics* **44**: 1061–1071.

Koch, G. G., and V. P. Bhapkar. 1982. Chi-square tests. Pp. 442–457 in *Encyclopedia of Statistical Sciences*, Vol. 1. New York: Wiley.

Koch, G. G., J. R. Landis, J. L. Freeman, D. H. Freeman, and R. G. Lehnen. 1977. A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* **33**: 133–158.

Koch, G. G., I. A. Amara, G. W. Davis, and D. B. Gillings. 1982. A review of some statistical methods for covariance analysis of categorical data. *Biometrics* **38**: 563–595.

Koch, G. G., P. B. Imrey, J. M. Singer, S. S. Atkinson, and M. E. Stokes. 1985. *Lecture Notes for Analysis of Categorical Data*. Montreal: Les Presses de L'Université de Montréal.

Koehler, K. 1986. Goodness-of-fit tests for log-linear models in sparse contingency tables. *J. Amer. Statist. Assoc.* **81**: 483–493.

Koehler, K. 1998. Chi-square tests. Pp. 608–622 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Koehler, K., and K. Larntz. 1980. An empirical investigation of goodness-of-fit statistics for sparse multinomials. *J. Amer. Statist. Assoc.* **75**: 336–344.

Koehler, K., and J. Wilson. 1986. Chi-square tests for comparing vectors of proportions for several cluster samples. *Commun. Statist. Ser. A* **15**: 2977–2990.

Koopman, P. A. R. 1984. Confidence limits for the ratio of two binomial proportions. *Biometrics* **40**: 513–517.

Kraemer, H. C. 1979. Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika* **44**: 461–472.

Kreiner, S. 1987. Analysis of multidimensional contingency tables by exact conditional tests: Techniques and strategies. *Scand. J. Statist.* **14**: 97–112.

Kreiner, S. 1998. Interaction models. Pp. 2063–2068 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Kruskal, W. H. 1958. Ordinal measures of association. *J. Amer. Statist. Assoc.* **53**: 814–861.

Ku, H. H., R. N. Varner, and S. Kullback. 1971. Analysis of multidimensional contingency tables. *J. Amer. Statist. Assoc.* **66**: 55–64.

Kuha, J., and C. Skinner. 1997. Categorical data analysis and misclassification. Pp. 633–670 in *Survey Measurement and Process Quality*, ed. L. Lyberg et al. New York: Wiley.

Kuha, J., C. Skinner, and J. Palmgren. 1998. Misclassification error. Pp. 2615–2621 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Kullback, S. 1959. *Information Theory and Statistics*. New York: Wiley.

Kullback, S., M. Kupperman, and H. H. Ku. 1962. Tests for contingency tables and Markov chains. *Technometrics* **4**: 573–608.

Kupper, L. L., and J. K. Haseman. 1978. The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics* **34**: 69–76.

Kupper, L. L., C. Portier, M. D. Hogan, and E. Yamamoto. 1986. The impact of litter effects on dose–response modeling in teratology. *Biometrics* **42**: 85–98.

Läärä, E., and J. N. S. Matthews. 1985. The equivalence of two models for ordinal data. *Biometrika* **72**: 206–207.

Lachin, J. M. 1977. Sample-size determinations for $r \times c$ comparative trials. *Biometrics* **33**: 315–324.

Laird, N. M. 1978. Empirical Bayes methods for two-way contingency tables. *Biometrika* **65**: 581–590.

Laird, N. M. 1998. EM algorithm. Pp. 1300–1313 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Laird, N. M., and D. Olivier. 1981. Covariance analysis of censored survival data using log-linear analysis techniques. *J. Amer. Statist. Assoc.* **76**: 231–240.

Lancaster, H. O. 1949. The derivation and partition of χ^2 in certain discrete distributions. *Biometrika* **36**: 117–129.

Lancaster, H. O. 1951. Complex contingency tables treated by partition of χ^2. *J. Roy. Statist. Soc. Ser. B* **13**: 242–249.

Lancaster, H. O. 1961. Significance tests in discrete distributions. *J. Amer. Statist. Assoc.* **56**: 223–234.

Lancaster, H. O. 1969. *The Chi-Squared Distribution*. New York: Wiley.

Lancaster, H. O., and M. A. Hamdan. 1964. Estimation of the correlation coefficient in contingency tables with possible nonmetrical characters. *Psychometrika* **29**: 383–391.

Landis, J. R., and G. G. Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* **33**: 363–374.

Landis, J. R., E. R. Heyman, and G. G. Koch. 1978. Average partial association in three-way contingency tables: A review and discussion of alternative tests. *Internat. Statist. Rev.* **46**: 237–254.

Landis, J. R., T. J. Sharp, S. J. Kuritz, and G. G. Koch. 1998. Mantel-Haenszel methods. Pp. 2378–2691 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Landwehr, J. M., D. Pregibon, and A. C. Shoemaker. 1984. Graphical methods for assessing logistic regression models. *J. Amer. Statist. Assoc.* **79**: 61–71.

Lang, J. B. 1992. Obtaining the observed information matrix for the Poisson log linear model with incomplete data. *Biometrika* **79**: 405–407.

Lang, J. B. 1996a. Maximum likelihood methods for a generalized class of log-linear models. *Ann. Statist.* **24**: 726–752.

Lang, J. B. 1996b. On the partitioning of goodness-of-fit statistics for multivariate categorical response models. *J. Amer. Statist. Assoc.* **91**: 1017–1023.

Lang, J. B. 1996c. On the comparison of multinomial and Poisson log-linear models. *J. Roy. Statist. Soc. Ser. B* **58**: 253–266.

Lang, J. B., and A. Agresti. 1994. Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *J. Amer. Statist. Assoc.* **89**: 625–632.

Lang, J. B., J. W. McDonald, and P. W. F. Smith. 1999. Association-marginal modeling of multivariate categorical responses: A maximum likelihood approach. *J. Amer. Statist. Assoc.* **94**: 1161–1171.

Laplace, P. S. 1812. *Théorie Analytique des Probabilités*. Paris: Courcier.

Larntz, K. 1978. Small-sample comparison of exact levels for chi-squared goodness-of-fit statistics. *J. Amer. Statist. Assoc.* **73**: 253–263.

Larsen, K., J. H. Petersen, E. Budtz-Jørgensen, and L. Endahl. 2000. Interpreting parameters in the logistic regression model with random effects. *Biometrics* **56**: 909–914.

Larson, M. G. 1984. Covariate analysis of competing-risks data with log-linear models. *Biometrics* **40**: 459–469.

Lauritzen, S. L. 1996. *Graphical Models*. New York: Oxford University Press.

Lauritzen, S. L., and N. Wermuth. 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17**: 31–57.

LaVange, L. M., G. G. Koch, and T. A. Schwartz. 2001. Applying sample survey methods to clinical trials data. *Statist. Medic.* **20**: 2609–2623.

Lawal, H. B. 1984. Comparisons of the X^2, Y^2, Freeman–Tukey and Williams improved G^2 test statistics in small samples of one-way multinomials. *Biometrika* **71**: 415–418.

Lawless, J. F. 1987. Negative binomial and mixed Poisson regression. *Canad. J. Statist.* **15**: 209–225.

Lazarsfeld, P. F., and N. W. Henry. 1968. *Latent Structure Analysis*. Boston: Houghton Mifflin.

Lee, S. K. 1977. On the asymptotic variances of û terms in loglinear models of multidimensional contingency tables. *J. Amer. Statist. Assoc.* **72**: 412–419.

Lee, Y., and J. A. Nelder. 1996. Hierarchical generalized linear models. *J. Roy. Statist. Soc. Ser B* **58**: 619–678.

Lefkopoulou, M., D. Moore, and L. Ryan. 1989. The analysis of multiple correlated binary outcomes: Application to rodent teratology experiments. *J. Amer. Statist. Assoc.* **84**: 810–815.

Lehmann, E. L. 1966. Some concepts of dependence. *Ann. Math. Statist.* **37**: 1137–1153.

Lehmann, E. L. 1986. *Testing Statistical Hypotheses*, 2nd ed. New York: Wiley.

Leonard, T. 1975. Bayesian estimation methods for two-way contingency tables. *J. Roy. Statist. Soc. Ser. B* **37**: 23–37.

Leonard, T. and J. S. J. Hsu. 1994. The Bayesian analysis of categorical data: A selective review. Pp. 283–310 in *Aspects of Uncertainty: A Tribute to D. V. Lindley*. P. R. Freeman and A. F. M. Smith, eds. New York: Wiley.

Lesaffre, E., and A. Albert. 1989. Multiple-group logistic regression diagnostics. *Appl. Statist.* **38**: 425–440.

Lesaffre, E., and G. Molenberghs. 1991. Multivariate probit analysis: A neglected procedure in medical statistics. *Statist. Medic.* **10**: 1391–1403.

Lesaffre, E., and B. Spiessens. 2001. On the effect of quadrature points in a logistic random-effects model: An example. *Appl. Statist.* **50**: 325–335.

Lewis, T., I. W. Saunders, and M. Westcott. 1984. The moments of the Pearson chi-squared statistic and the minimum expected value in two-way tables. *Biometrika* **71**: 515–522.

Liang, K. Y. 1984. The asymptotic efficiency of conditional likelihood methods. *Biometrika* **71**: 305–313.

Liang, K. Y., and J. Hanfelt. 1994. On the use of the quasi-likelihood method in teratological experiments. *Biometrics* **50**: 872–880.

Liang, K. Y., and P. McCullagh. 1993. Case studies in binary dispersion. *Biometrics* **49**: 623–630.

Liang, K. Y., and S. G. Self. 1985. Tests for homogeneity of odds ratios when the data are sparse. *Biometrika* **72**: 353–358.

Liang, K. Y., and S. L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* **73**: 13–22.

Liang, K. Y., and S. L. Zeger. 1988. On the use of concordant pairs in matched case–control studies. *Biometrics* **44**: 1145–1156.

Liang, K. Y, and S. L. Zeger. 1995. Inference based on estimating functions in the presence of nuisance parameters. *Statist. Sci.* **10** 158–173.

Liang, K. Y., S. L. Zeger, and B. Qaqish. 1992. Multivariate regression analyses for categorical data. *J. Roy. Statist. Soc.Ser. B* **54**: 3–24.

Lin, X. 1997. Variance component testing in generalized linear models with random effects. *Biometrika* **84**: 309–326.

Lindley, D. V. 1964. The Bayesian analysis of contingency tables. *Ann. Math. Statist.* **35**: 1622–1643.

Lindsay, B., C. Clogg, and J. Grego. 1991. Semi-parametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *J. Amer. Statist. Assoc.* **86**: 96–107.

Lindsey, J. K. 1999. *Models for Repeated Measurements*, 2nd ed. Oxford: Oxford University Press.

Lindsey, J. K., and P. M. E. Altham. 1998. Analysis of the human sex ratio by using overdispersion models. *Appl. Statist.* **47**: 149–157.

Lindsey, J. K., and G. Mersch. 1992. Fitting and comparing probability distributions with log linear models. *Comput. Statist. Data Anal.* **13**: 373–384.

Lipsitz, S. 1992. Methods for estimating the parameters of a linear model for ordered categorical data. *Biometrics* **48**: 271–281.

Lipsitz, S. R., and G. Fitzmaurice. 1996. The score test for independence in $R \times C$ contingency tables with missing data. *Biometrics* **52**: 751–762.

Lipsitz, S., N. Laird, and D. Harrington. 1990. Finding the design matrix for the marginal homogeneity model. *Biometrika* **77**: 353–358.

Lipsitz, S., N. Laird, and D. Harrington. 1991. Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika* **78**: 153–160.

Lipsitz, S. R., K. Kim, and L. Zhao. 1994. Analysis of repeated categorical data using generalized estimating equations. *Statist. Medic.* **13**: 1149–1163.

Little, R. J. 1989. Testing the equality of two independent binomial proportions. *Amer. Statist.* **43**: 283–288.

Little, R. J. 1998. Missing data. Pp. 2622–2635 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Little, R. J., and D. B. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Little, R. J. A., and M.-M. Wu. 1991. Models for contingency tables with known margins when target and sampled populations differ. *J. Amer. Statist. Assoc.* **86**: 87–95.

Liu, Q., and D. A. Pierce. 1993. Heterogeneity in Mantel–Haenszel-type models. *Biometrika* **80**: 543–556.

Liu, Q., and D. A. Pierce. 1994. A note on Gauss–Hermite quadrature. *Biometrika* **81**: 624–629.

Lloyd, C. J. 1988a. Some issues arising from the analysis of 2×2 contingency tables. *Austral. J. Statist.* **30**: 35–46.

Lloyd, C. J. 1988b. Doubling the one-sided *P*-value in testing independence in 2×2 tables against a two-sided alternative. *Statist. Medic.* **7**: 1297–1306.

Lloyd, C. J. 1999. *Statistical Analysis of Categorical Data*. New York: Wiley.

Longford, N. T. 1993. *Random Coefficient Models*. New York: Oxford University Press.

Loughin, T. M., and P. N. Scherer. 1998. Testing for association in contingency tables with multiple column responses. *Biometrics* **54**: 630–637.

Louis, T. A. 1982. Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44**: 226–233.

Luce, R. D. 1959. *Individual Choice Behavior*. New York: Wiley.

Madansky, A. 1963. Tests of homogeneity for correlated samples. *J. Amer. Statist. Assoc.* **58**: 97–119.

Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Magnus, J. R., and H. Neudecker. 1988. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New York: Wiley.

Mantel, N. 1963. Chi-square tests with one degree of freedom: Extensions of the Mantel–Haenszel procedure. *J. Amer. Statist. Assoc.* **58**: 690–700.

Mantel, N. 1966. Models for complex contingency tables and polychotomous dosage response curves. *Biometrics* **22**: 83–95.

Mantel, N. 1973. Synthetic retrospective studies and related topics. *Biometrics* **29**: 479–486.

Mantel, N. 1985. Maximum likelihood vs. minimum chi-square. *Biometrics* **41**: 777–781.

Mantel, N. 1987a. Understanding Wald's test for exponential families. *Amer. Statist.* **41**: 147–148.

Mantel, N. 1987b. Exact tests for 2×2 contingency tables (Letter). *Amer. Statist.* **41**: 159.

Mantel, N., and D. P. Byar. 1978. Marginal homogeneity, symmetry and independence. *Commun. Statist. Ser. A* **7**: 953–976.

Mantel, N., and W. Haenszel. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22**: 719–748.

Martín Andrés, A., and Silva Mato, A. 1994. Choosing the optimal unconditional test for comparing two independent proportions. *Comput. Statist. Data Anal.* **17**: 555–574.

Matthews, J. N. S., and K. P. Morris. 1995. An application of Bradley–Terry-type models to the measurement of pain. *Appl. Statist.* **44**: 243–255.

McCullagh, P. 1978. A class of parametric models for the analysis of square contingency tables with ordered categories. *Biometrika* **65**: 413–418.

McCullagh, P. 1980. Regression models for ordinal data. *J. Roy. Statist. Soc. Ser. B* **42**: 109–142.

McCullagh, P. 1982. Some applications of quasisymmetry. *Biometrika* **69**: 303–308.

McCullagh, P. 1983. Quasi-likelihood functions. *Ann. Statist.* **11**: 59–67.

McCullagh, P. 1986. The conditional distribution of goodness-of-fit statistics for discrete data. *J. Amer. Statist. Assoc.* **81**: 104–107.

McCullagh, P., and J. A. Nelder. 1983; 2nd ed., 1989. *Generalized Linear Models*. London: Chapman & Hall.

McCulloch, C. E. 1994. Maximum likelihood variance components estimation for binary data. *J. Amer. Statist. Assoc.* **89**: 330–335.

McCulloch, C. E. 1997. Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* **92**: 162–170.

McCulloch, C. E. 2000. Generalized linear models. *J. Amer. Statist. Assoc.* **95**: 1320–1324.

McCulloch, C. E., and S. Searle. 2001. *Generalized, Linear, and Mixed Models*. New York: Wiley.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. Pp. 105–142 in *Frontiers in Econometrics*, ed. P. Zarembka. New York: Academic Press.

McFadden, D. 1982. Qualitative response models. Pp. 1–37 in *Advances in Econometrics*, ed. W. Hildebrand. Cambridge: Cambridge University Press.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**: 153–157.

Mee, R. W. 1984. Confidence bounds for the difference between two probabilities (letter). *Biometrics* **40**: 1175–1176.

Meeden, G., C. Geyer, J. Lang, and E. Funo. 1998. The admissibility of the maximum likelihood estimator for decomposable log-linear interaction models for contingency tables. *Commun. Statist. Ser. A* **27**: 473–493.

Mehta, C. R. 1994. The exact analysis of contingency tables in medical research. *Statist. Methods Medic. Res.* **3**: 135–156.

Mehta, C. R., and N. R. Patel. 1983. A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *J. Amer. Statist. Assoc.* **78**: 427–434.

Mehta, C. R., and N. R. Patel. 1995. Exact logistic regression: Theory and examples. *Statist. Medic.* **14**: 2143–2160.

Mehta, C. R., and S. J. Walsh. 1992. Comparison of exact, mid-*P*, and Mantel–Haenszel confidence intervals for the common odds ratio across several 2×2 contingency tables. *Amer. Statist.* **46**: 146–150.

Mehta, C. R., N. R. Patel, and R. Gray. 1985. Computing an exact confidence interval for the common odds ratio in several 2 by 2 contingency tables. *J. Amer. Statist. Assoc.* **80**: 969–973.

Mehta, C. R., N. R. Patel, and P. Senchaudhuri. 1988. Importance sampling for estimating exact probabilities in permutational inference. *J. Amer. Statist. Assoc.* **83**: 999–1005.

Mehta, C. R., N. R. Patel, and P. Senchaudhuri. 2000. Efficient Monte Carlo methods for conditional logistic regression. *J. Amer. Statist. Assoc.* **95**: 99–108.

Michailidis, G., and J. de Leeuw. 1998. The Gifi system of descriptive multivariate analysis. *Statist. Sci.* **13**: 307–336.

Miettinen, O. S. 1969. Individual matching with multiple controls in the case of all-or-none responses. *Biometrics* **25**: 339–355.

Miettinen, O. S., and M. Nurminen. 1985. Comparative analysis of two rates. *Statist. Medic.* **4**: 213–226.

Miller, M. E., C. S. Davis, and J. R. Landis. 1993. The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares. *Biometrics* **49**: 1033–1044.

Minkin, S. 1987. On optimal design for binary data. *J. Amer. Statist. Assoc.* **82**: 1098–1103.

Mirkin, B. 2001. Eleven ways to look at the chi-squared coefficient for contingency tables. *Amer. Statist.* **55**: 111–120.

Mitra, S. K. 1958. On the limiting power function of the frequency chi-square test. *Ann. Statist.* **29**: 1221–1233.

Molenberghs, G., and E. Goetghebeur. 1997. Simple fitting algorithms for incomplete categorical data. *J. Roy. Statist. Soc. Ser. B* **59**: 401–414.

Molenberghs, G., and E. Lesaffre. 1994. Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. *J. Amer. Statist. Assoc.* **89**: 633–644.

Molenberghs, G., M. G. Kenward, and E. Lesaffre. 1997. The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika* **84**: 33–44.

Moore, D. F. 1986a. Asymptotic properties of moment estimates for overdispersed counts and proportions. *Biometrika* **35**: 583–588.

Moore, D. S. 1986b. Tests of chi-squared type. Pp. 63–95 in *Goodness-of-Fit Techniques*, ed. R. D'Agostino and M. A. Stephens. New York: Marcel Dekker.

Moore, D. F., and A. Tsiatis. 1991. Robust estimation of the variance in moment methods for extra-binomial and extra-Poisson variation. *Biometrics* **47**: 383–401.

Morgan, B. J. T. 1992. *Analysis of Quantal Response Data*. London: Chapman & Hall.

Morgan, W. M., and B. A. Blumenstein. 1991. Exact conditional tests for hierarchical models in multidimensional contingency tables. *Appl. Statist.* **40**: 435–442.

Mosimann, J. E. 1962. On the compound multinomial distribution, the multivariate β-distribution and correlations among proportions. *Biometrika* **49**: 65–82.

Mosteller, F. 1951. Remarks on the method of paired comparisons I: The least-squares solution assuming equal standard deviations and equal correlations. *Psychometrika* **16**: 3–9.

Mosteller, F. 1952. Some statistical problems in measuring the subjective response to drugs. *Biometrics* **8**: 220–226.

Mosteller, F. 1968. Association and estimation in contingency tables. *J. Amer. Statist. Assoc.* **63**: 1–28.

Nair, V. N. 1987. Chi-squared-type tests for ordered alternatives in contingency tables. *J. Amer. Statist. Assoc.* **82**: 283–291.

Natarajan, R., and C. McCulloch. 1995. A note on the existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika* **82**: 639–643.

Natarajan, R., and C. McCulloch. 1998. Gibbs sampling with diffuse proper priors: A valid approach to data-driven inference? *J. Comput. Graph. Statist.* **7**: 267–277.

Nelder, J., and D. Pregibon. 1987. An extended quasi-likelihood function. *Biometrika* **74**: 221–232.

Nelder, J., and R. W. M. Wedderburn. 1972. Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135**: 370–384.

Nerlove, M., and S. J. Press. 1973. Univariate and multivariate log-linear and logistic models. Technical Report R-1306-EDA/NIH, Rand Corporation, Santa Monica, CA.

Neuhaus, J. M. 1992. Statistical methods for longitudinal and clustered designs with binary responses. *Statist. Methods Medic. Res.* **1**: 249–273.

Neuhaus, J. M., and N. P. Jewell. 1990a. Some comments on Rosner's multiple logistic model for clustered data. *Biometrics* **46**: 523–534.

Neuhaus, J. M., and N. P. Jewell. 1990b. The effect of retrospective sampling on binary regression models for clustered data. *Biometrics* **46**: 977–990.

Neuhaus, J. M., and M. L. Lesperance. 1996. Estimation efficiency in a binary mixed-effects model setting. *Biometrika* **83**: 441–446.

Neuhaus, J. M., J. D. Kalbfleisch, and W. W. Hauck. 1991. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Internat. Statist. Rev.* **59**: 25–35.

Neuhaus, J. M., W. W. Hauck, and J. D. Kalbfleisch. 1992. The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79**: 755–762.

Neuhaus, J. M., J. D. Kalbfleisch, and W. W. Hauck. 1994. Conditions for consistent estimation in mixed-effects models for binary matched-pairs data. *Canad. J. Statist.* **22**: 139–148.

Newcombe, R. 1998a. Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statist. Medic.* **17**: 857–872.

Newcombe, R. 1998b. Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statist. Medic.* **17**: 873–890.

Newcombe, R. 2001. Logit confidence intervals and the inverse sinh transformation. *Amer. Statist.* **55**: 200–202.

Neyman, J. 1935. On the problem of confidence limits. *Ann. Math. Statist.* **6**: 111–116.

Neyman, J. 1949. Contributions to the theory of the χ^2 test. Pp. 239–273 in *Proc. First Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley, CA: University of California Press.

Nurminen, M. 1986. Confidence intervals for the ratio and difference of two binomial proportions. *Biometrics* **42**: 675–676.

O'Brien, P. C. 1988. Comparing two samples: Extensions of the *t*, rank-sum, and log-rank tests. *J. Amer. Statist. Assoc.* **83**: 52–61.

O'Brien, R. G. 1986. Using the SAS system to perform power analyses for log-linear models. Pp. 778–784 in *Proc. 11th Annual SAS Users Group Conference*. Cary, NC: SAS Institute.

Ochi, Y., and R. Prentice. 1984. Likelihood inference in a correlated probit regression model. *Biometrika* **71**: 531–543.

O'Gorman, T. W., and R. F. Woolson. 1988. Analysis of ordered categorical data using the SAS system. Pp. 957–963 in *Proc. 13th Annual SAS Users Group Conference*. Cary, NC: SAS Institute.

Paik, M. 1985. A graphic representation of a three-way contingency table: Simpson's paradox and correlation. *Amer. Statist.* **39**: 53–54.

Palmgren, J. 1981. The Fisher information matrix for log-linear models arguing conditionally in the observed explanatory variables. *Biometrika* **68**: 563–566.

Palmgren, J., and A. Ekholm. 1987. Exponential family non-linear models for categorical data with errors of observation. *Appl. Stochastic Models Data Anal.* **3**: 111–124.

Park, T., and M. B. Brown. 1994. Models for categorical data with nonignorable nonresponse. *J. Amer. Statist. Assoc.* **89**: 44–52.

Parr, W. C., and H. D. Tolley. 1982. Jackknifing in categorical data analysis. *Austral. J. Statist.* **24**: 67–79.

Parzen, E. 1997. Concrete statistics. Pp. 309–332 in *Statistics of Quality*. New York: Marcel Dekker.

Patefield, W. M. 1982. Exact tests for trends in ordered contingency tables. *Appl. Statist. Ser B* **31**: 32–43.

Patnaik, P. B. 1949. The non-central χ^2 and *F*-distributions and their applications. *Biometrika* **36**: 202–232.

Paul, S. R., K. Y. Liang, and S. G. Self. 1989. On testing departure from the binomial and multinomial assumptions. *Biometrics* **45**: 231–236.

Pearson, E. S. 1947. The choice of a statistical test illustrated on the interpretation of data classified in 2×2 tables. *Biometrika* **34**: 139–167.

Pearson, K. 1900. On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag. Ser.* 5 **50**: 157–175. (Reprinted in *Karl Pearson's Early Statistical Papers*, ed. E. S. Pearson. Cambridge: Cambridge University Press, 1948.)

Pearson, K. 1904. Mathematical contributions to the theory of evolution XIII: On the theory of contingency and its relation to association and normal correlation. *Draper's Co. Research Memoirs, Biometric Series*, no. 1. (Reprinted in *Karl Pearson's Early Papers*, ed. E. S. Pearson, Cambridge: Cambridge University Press, 1948.)

Pearson, K. 1913. On the probable error of a correlation coefficient as found from a fourfold table. *Biometrika* **9**: 22–27.

Pearson, K. 1917. On the general theory of multiple contingency with special reference to partial contingency. *Biometrika* **11**: 145–158.

Pearson, K. 1922. On the χ^2 test of goodness of fit. *Biometrika* **14**: 186–191.

Pearson, K., and D. Heron. 1913. On theories of association. *Biometrika* **9**: 159–315.

Peduzzi, P., J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein. 1996. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* **49**: 1373–1379.

Pendergast, J. F., S. J. Gange, M. A. Newton, M. J. Lindstrom, M. Palta, and M. R. Fisher. 1996. A survey of methods for analyzing clustered binary response data. *Internat. Statist. Rev.* **64**: 89–118.

Pepe, M. S. 2000. Receiver operating characteristic methodology. *J. Amer. Statist. Assoc.* **95**: 308–311.

Peterson, B., and F. E. Harrell, Jr. 1990. Partial proportional odds models for ordinal response variables. *Appl. Statist.* **39**: 205–217.

Pierce, D. A., and D. Peters. 1992. Practical use of higher order asymptotics for multiparameter exponential families. *J. Roy. Statist. Soc. Ser. B* **54**: 701–725.

Pierce, D. A., and D. Peters. 1999. Improving on exact tests by approximate conditioning. *Biometrika* **86**: 265–277.

Pierce, D. A., and B. R. Sands. 1975. Extra-Bernoulli variation in regression of binary data. Technical Report 46, Statistics Deptartment, Oregon State University, Cornwallis, OR.

Pierce, D. A., and D. W. Schafer. 1986. Residuals in generalized linear models. *J. Amer. Statist. Assoc.* **81**: 977–983.

Plackett, R. L. 1962. A note on interactions in contingency tables. *J. Roy. Statist. Soc. Ser. B* **24**: 162–166.

Plackett, R. L. 1964. The continuity correction in 2×2 tables. *Biometrika* **51**: 327–337.

Plackett, R. L. 1983. Karl Pearson and the chi-squared test. *Internat. Statist. Rev.* **51**: 59–72.

Podgor, M. J., J. L. Gastwirth, and C. R. Mehta. 1996. Efficiency robust tests of independence in contingency tables with ordered classifications. *Statist. Medic.* **15**: 2095–2105.

Poisson, S.-D. 1837. *Recherches sur la probabilité des jugements en matieère criminelle et en matière civile, précédées des règles générales du calcul des probabilités*. Paris: Bachelier.

Pratt, J. W. 1981. Concavity of the log likelihood. *J. Amer. Statist. Assoc.* **76**: 103–106.

Pregibon, D. 1980. Goodness of link tests for generalized linear models. *Appl. Statist.* **29**: 15–24.

Pregibon, D. 1981. Logistic regression diagnostics. *Ann. Statist.* **9**: 705–724.

Pregibon, D. 1982. Score tests in GLIM with application. Pp. 87–97 in *Lecture Notes in Statistics*, 14: *GLIM* 82, *Proc. International Conference on Generalised Linear Models*, ed. R. Gilchrist. New York: Springer-Verlag.

Prentice, R. 1976a. Use of the logistic model in retrospective studies. *Biometrics* **32**: 599–606.

Prentice, R. 1976b. Generalization of the probit and logit methods for dose response curves. *Biometrics* **32**: 761–768.

Prentice, R. 1986. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *J. Amer. Statist. Assoc.* **81**: 321–327.

Prentice, R., and N. Breslow. 1978. Retrospective studies and failure time models. *Biometrika* **65**: 153–158.

Prentice, R., and L. A. Gloeckler. 1978. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* **34**: 57–67.

Prentice, R., and R. Pyke. 1979. Logistic disease incidence models and case-control studies. *Biometrika* **66**: 403–412.

Prentice, R., and L. P. Zhao. 1991. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47**: 825–839.

Press, S. J., and S. Wilson. 1978. Choosing between logistic regression and discriminant analysis. *J. Amer. Statist. Assoc.* **73**: 699–705.

Qu, A., B. G. Lindsay, and B. Li. 2000. Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87**: 823–836.

Quine, M. P., and E. Seneta. 1987. Bortkiewicz's data and the law of small numbers. *Internat. Statist. Rev.* **5**: 173–181.

Rabe-Hesketh, S., and A. Skrondal. 2001. Parameterisation of multivariate random effects models for categorical data. *Biometrics* **57**:–.

Raftery, A. E. 1986. Choosing models for cross-classification. *Amer. Sociol. Rev.* **51**: 145–146.

Rao, C. R. 1957. Maximum likelihood estimation for the multinomial distribution. *Sankhya* **18**: 139–148.

Rao, C. R. 1963. Criteria of estimation in large samples. *Sankhya* **25**: 189–206.

Rao, C. R. 1973. *Linear Statistical Inference and Its Applications*, 2nd ed. New York: Wiley.

Rao, C. R. 1982. Diversity: Its measurement, decomposition, apportionment, and analysis. *Sankhya Ser. A* **44**: 1–22.

Rao, J. N. K., and A. J. Scott. 1987. On simple adjustments to chi-square tests with sample survey data. *Ann. Statist.* **15**: 385–397.

Rao, J. N. K., and D. R. Thomas. 1988. The analysis of cross-classified categorical data from complex sample surveys. *Sociol. Methodol.* **18**: 213–270.

Rasch, G. 1961. On general laws and the meaning of measurement in psychology. Pp. 321–333 in *Proc. 4th Berkeley Symposium on Mathematics, Statistics, and Probability*, Vol. 4, ed. J. Neyman. Berkeley, CA: University of California Press.

Rayner, J. C. W., and D. J. Best. 2001. *A Contingency Table Approach to Nonparametric Testing*. London: Chapman & Hall.

Read, T. R. C., and N. A. C. Cressie. 1988. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York: Springer-Verlag.

Rice, W. R. 1988. A new probability model for determining exact P-values for 2×2 contingency tables when comparing binomial proportions. *Biometrics* **44**: 1–22.

Ritov, Y., and Z. Gilula. 1991. The order-restricted RC model for ordered contingency tables: Estimation and testing for fit. *Ann. Statist.* **19**: 2090–2101.

Robins, J., N. Breslow, and S. Greenland. 1986. Estimators of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* **42**: 311–323.

Robins, J., A. Rotnitzky, and L. P. Zhao. 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90**: 106–121.

Röhmel, J., and U. Mansmann. 1999. Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biometrical J.* **41**: 149–170.

Rosenbaum, P. R., and D. R. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**: 41–55.

Rosner, B. 1984. Multivariate methods in ophthalmology with application to other paired-data situations. *Biometrics* **40**: 1025–1035.

Rosner, B. 1989. Multivariate methods for clustered binary data with more than one level of nesting. *J. Amer. Statist. Assoc.* **84**: 373–380.

Rotnitzky, A., and N. P. Jewell. 1990. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* **77**: 485–497.

Routledge, R. D. 1992. Resolving the conflict over Fisher's exact test. *Canad. J. Statist.* **20**: 201–209.

Routledge, R. D. 1994. Practicing safe statistics with the mid-P^*. *Canad. J. Statist.* **22**: 103–110.

Roy, S. N., and M. A. Kastenbaum. 1956. On the hypothesis of no "interaction" in a multiway contingency table. *Ann. Math. Statist.* **27**: 749–757.

Roy, S. N., and S. K. Mitra. 1956. An introduction to some nonparametric generalizations of analysis of variance and multivariate analysis. *Biometrika* **43**: 361–376.

Rudas, T., C. C. Clogg, and B. G. Lindsay. 1994. A new index of fit based on mixture methods for the analysis of contingency tables. *J. Roy. Statist. Soc.* **56**: 623–639.

Ryan, L. 1992. Quantitative risk assessment for developmental toxicity. *Biometrics* **48**: 163–174.

Ryan, L. 1995. Comment on article by Liang and Zeger. *Statist. Sci.* **10**: 189–193.

Samuels, M. L. 1993. Simpson's paradox and related phenomena. *J. Amer. Statist. Assoc.* **88**: 81–88.

Santner, T. J., and M. K. Snell. 1980. Small-sample confidence intervals for p_1-p_2 and p_1/p_2 in 2×2 contingency tables. *J. Amer. Statist. Assoc.* **75**: 386–394.

Santner, T. J., and S. Yamagami. 1993. Invariant small sample confidence intervals for the difference of two success probabilities. *Commun. Statist. Ser. B* **22**: 33–59.

Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schluchter, M. D., and K. L. Jackson. 1989. Log-linear analysis of censored survival data with partially observed covariates. *J. Amer. Statist. Assoc.* **84**: 42–52.

Scott, A., and C. Wild. 2001. Case–control studies with complex sampling. *Appl. Statist.* **50**: 389–401.

Seeber, G. 1998. Poisson regression. Pp. 3404–3412 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Sekar, C. C., and W. E. Deming. 1949. On a method of estimating birth and death rates and the extent of registration. *J. Amer. Statist. Assoc.* **44**: 101–115.

Self, S. G., and K.-Y. Liang. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.* **82**: 605–610.

Sen, P. K., and J. M. Singer. 1993. *Large Sample Methods in Statistics: An Introduction with Applications*. London: Chapman & Hall.

Shapiro, S. H. 1982. Collapsing contingency tables: A geometric approach. *Amer. Statist.* **36**: 43–46.

Shuster, J., and D. Downing. 1976. Two-way contingency tables for complex sampling schemes. *Biometrika* **63**: 271–276.

Silvapulle, M. J. 1981. On the existence of maximum likelihood estimators for the binomial response models. *J. Roy. Statist. Soc. Ser. B* **43**: 310–313.

Simon, G. 1973. Additivity of information in exponential family probability laws.)*J. Amer. Statist. Assoc.* **68**: 478–482.

Simon, G. 1974. Alternative analyses for the singly-ordered contingency table. *J. Amer. Statist. Assoc.* **69**: 971–976.

Simon, G. 1978. Efficacies of measures of association for ordinal contingency tables. *J. Amer. Statist. Assoc.* **73**: 545–551.

Simonoff, J. 1983. A penalty function approach to smoothing large sparse contingency tables. *Ann. Statist.* **11**: 208–218.

Simonoff, J. 1986. Jackknifing and bootstrapping goodness-of-fit statistics in sparse multinomials. *J. Amer. Statist. Assoc.* **81**: 1005–1111.

Simonoff, J. S. 1996. *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Simonoff, J. S. 1998. Three sides of smoothing: Categorical data smoothing, nonparametric regression, and density estimation. *Internat. Statist. Rev.* **66**: 137–156.

Simpson, E. H. 1949. The measurement of diversity. *Nature* **163**: 699.

Simpson, E. H. 1951. The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. Ser. B* **13**: 238–241.

Skellam, J. G. 1948. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *J. Roy. Statist. Soc. Ser. B* **10**: 257–261.

Skene, A. M., and J. C. Wakefield. 1990. Hierarchical models for multicentre binary response studies. *Statist. Medic.* **9**: 919–929.

Slaton, T. L., W. W. Piegorsch, and S. D. Durham. 2000. Estimation and testing with overdispersed proportions using the beta-logistic regression model of Heckman and Willis. *Biometrics* **56**: 125–133.

Small, K. A. 1987. A discrete choice model for ordered alternatives. *Econometrica* **55**: 409–424.

Smith, K. W. 1976. Table standardization and table shrinking: Aids in the traditional analysis of contingency tables. *Social Forces* **54**: 669–693.

Smith, P. W. F., J. J. Forster, and J. W. McDonald. 1996. Monte Carlo exact tests for square contingency tables. *J. Roy. Statist. Soc. Ser. A* **159**: 309–321.

Snell, E. J. 1964. A scaling procedure for ordered categorical data. *Biometrics* **20**: 592–607.

Somers, R. H. 1962. A new asymmetric measure of association for ordinal variables. *Amer. Sociol. Rev.* **27**: 799–811.

Speed, T. 1998. Iterative proportional fitting. Pp. 2116–2119 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Spiegelhalter, D. J., and A. F. M. Smith. 1982. Bayes factors for linear and log-linear models with vague prior information. *J. Roy. Statist. Soc. Ser. B* **44**: 377–387.

Spitzer, R. L., J. Cohen, J. L. Fleiss, and J. Endicott. 1967. Quantification of agreement in psychiatric diagnosis. *Arch. Gen. Psychiatry* **17**: 83–87.

Sprott, D. A. 2000. *Statistical Inference in Science*. New York: Springer-Verlag.

Stern, S. 1997. Simulation-based estimation. *J. Econ. Literature* **35**: 2006–2039.

Sterne, T. E. 1954. Some remarks on confidence or fiducial limits. *Biometrika* **41**: 275–278.

Stevens, S. S. 1951. Mathematics, measurement, and psychophysics. Pp. 1–49 in *Handbook of Experimental Psychology*, ed. S. S. Stevens. New York: Wiley.

Stevens, W. L. 1950. Fiducial limits of the parameter of a discontinuous distribution. *Biometrika* **37**: 117–129.

Stigler, S. 1986. *The History of Statistics*: *The Measurement of Uncertainty before* 1900. Cambridge, MA: Harvard University Press.

Stigler, S. 1994. Citation patterns in the journals of statistics and probability. *Statist. Sci.* **9**: 94–108.

Stigler, S. 1999. *Statistics on the Table*. Cambridge, MA: Harvard University Press.

Stiratelli, R., N. Laird, and J. H. Ware. 1984. Random-effects models for serial observations with binary response. *Biometrics* **40**: 1025–1035.

Stokes, M. E., C. S. Davis, and G. G. Koch. 2000. *Categorical Data Analysis Using the SAS System*, 2nd ed. Cary, NC: SAS Institute.

Strawderman, R. L., and M. T. Wells. 1998. Approximately exact inference for the common odds ratio in several 2×2 tables. *J. Amer. Statist. Assoc.* **93**: 1294–1307.

Stuart, A. 1955. A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika* **42**: 412–416.

Stukel, T. A. 1988. Generalized logistic models. *J. Amer. Statist. Assoc.* **83**: 426–431.

Suissa, S., and J. J. Shuster. 1984. Are uniformly most powerful unbiased tests really best? *Amer. Statist.* **38**: 204–206.

Suissa, S., and J. J. Shuster. 1985. Exact unconditional samples sizes for the 2 by 2 binomial trial. *J. Roy. Statist. Soc. Ser. A* **148**: 317–327.

Suissa, S., and J. J. Shuster. 1991. The 2×2 matched-pairs trial: Exact unconditional design and analysis. *Biometrics* **47**: 361–372.

Sundberg, R. 1975. Some results about decomposable (or Markov-type) models for multidimensional contingency tables: Distribution of marginals and partitioning of tests. *Scand. J. Statist.* **2**: 71–79.

Tango, T. 1998. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statist. Medic.* **17**: 891–908.

Tanner, M. A., and M. A. Young. 1985. Modelling agreement among raters. *J. Amer. Statist. Assoc.* **80**: 175–180.

Tarone, R. E. 1985. On heterogeneity tests based on efficient scores. *Biometrika* **72**: 91–95.

Tarone, R. E., and J. J. Gart. 1980. On the robustness of combined tests for trends in proportions. *J. Amer. Statist. Assoc.* **75**: 110–116.

Tarone, R. E., J. J. Gart, and W. W. Hauck. 1983. On the asymptotic relative efficiency of certain noniterative estimators of a common relative risk or odds ratio. *Biometrika* **70**: 519–522.

Tavaré, S., and P. M. E. Altham. 1983. Serial dependence of observations leading to contingency tables, and corrections to chi-squared statistics. *Biometrika* **70**: 139–144.

Ten Have, T. R. 1996. A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses. *Biometrics* **52**: 473–491.

Ten Have, T. R., and A. R. Localio. 1999. Empirical Bayes estimation of random effects parameters in mixed effects logistic regression models. *Biometrics* **55**: 1022–1029.

Ten Have, T. R., and A. Morabia. 1999. Mixed effects models with bivariate and univariate association parameters for longitudinal bivariate binary response data. *Biometrics* **55**: 85–93.

Ten Have, T. R., and D. H. Uttal. 1994. Subject-specific and population-averaged continuation ratio logit models for multiple discrete time survival profiles. *Appl. Statist.* **43**: 371–384.

Theil, H. 1969. A multinomial extension of the linear logit model. *Internat. Econ. Rev.* **10**: 251–259.

Theil, H. 1970. On the estimation of relationships involving qualitative variables. *Amer. J. Sociol.* **76**: 103–154.

Thompson, R., and R. J. Baker. 1981. Composite link functions in generalized linear models. *Appl. Statist.* **30**: 125–131.

Thompson, W. A. 1977. On the treatment of grouped observations in life studies. *Biometrics* **33**: 463–470.

Thurstone, L. L. 1927. The method of paired comparisons for social values. *J. Abnormal Social Psych.* **21**: 384–400.

Tjur, T. 1982. A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scand. J. Statist.* **9**: 23–30.

Tocher, K. D. 1950. Extension of the Neyman–Pearson theory of tests to discontinuous variates. *Biometrika* **37**: 130–144.

Toledano, A., and C. Gatsonis. 1996. Ordinal regression methodology for ROC curves derived from correlated data. *Statist. Medic.* **15**: 1807–1826.

Train, K. 1986. *Qualitative Choice Analysis*: *Theory*, *Econometrics*, *and an Application*. Cambridge, MA: MIT Press.

Tsiatis, A. A. 1980. A note on the goodness-of-fit test for the logistic regression model. *Biometrika* **67**: 250–251.

Tutz, G. 1989. Compound regression models for ordered categorical data. *Biometrical J.* **31**: 259–272.

Tutz, G. 1991. Sequential models in categorical regression. *Comput. Statist. Data Anal.* **11**: 275–295.

Tutz, G., and W. Hennevogl. 1996. Random effects in ordinal regression models. *Comput. Statist. Data Anal.* **22**: 537–557.

Uebersax, J. S. 1993. Statistical modeling of expert ratings on medical treatment appropriateness. *J. Amer. Statist. Assoc.* **88**: 421–427.

Uebersax, J. S., and W. M. Grove. 1990. Latent class analysis of diagnostic agreement. *Statist. Medic.* **9**: 559–572.

Uebersax, J. S., and W. M. Grove. 1993. A latent trait finite mixture model for the analysis of rating agreement. *Biometrics* **49**: 823–835.

van der Heijden, P. G. M., and J. de Leeuw. 1985. Correspondence analysis: A complement to log-linear analysis. *Psychometrika* **50**: 429–447.

van der Heijden, P. G. M., A. de Falguerolles, and J. de Leeuw. 1989. A combined approach to contingency table analysis using correspondence analysis and log-linear analysis. *Appl. Statist.* **38**: 249–292.

Verbeke, G., and E. Lesaffre. 1996. A linear mixed-effects model with heterogeneity in the random-effects population. *J. Amer. Statist. Assoc.* **91**: 217–221.

Verbeke, G., and G. Molenberghs. 2000. *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Wald, A. 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* **54**: 426–482.

Walker, S. H., and D. B. Duncan. 1967. Estimation of the probability of an event as a function of several independent variables. *Biometrika* **54**: 167–179.

Walley, P. 1996. Inferences from multinomial data: Learning about a bag of marbles. *J. Roy. Statist. Soc. Ser. B* **58**: 3–34.

Wardrop, R. L. 1995. Simpson's paradox and the hot hand in basketball. *Amer. Statist.* **49**: 24–28.

Ware, J. H., S. Lipsitz, and F. E. Speizer. 1988. Issues in the analysis of repeated categorical outcomes. *Statist. Medic.* **7**: 95–107.

Watson, G. S. 1956. Missing and "mixed up" frequencies in contingency tables. *Biometrics* **12**: 47–50.

Watson, G. S. 1959. Some recent results in chi-square goodness-of-fit tests. *Biometrics* **15**: 440–468.

Wedderburn, R. W. M. 1974. Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61**: 439–447.

Wedderburn, R. W. M. 1976. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63**: 27–32.

Wermuth, N. 1976. Model search among multiplicative models. *Biometrics* **32**: 253–263.

Wermuth, N. 1987. Parametric collapsibility and the lack of moderating effects in contingency tables with a dichotomous response variable. *J. Roy. Statist. Soc. Ser. B* **49**: 353–364.

Westfall, P. H., and R. D. Wolfinger. 1997. Multiple tests with discrete distributions. *Amer. Statist.* **51**: 3–8.

Westfall, P. H., and S. S. Young. 1993. *Resampling-Based Multiple Testing*: *Examples and Methods for p-Value Adjustment*. New York: Wiley.

White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* **50**: 1–26.

White, A. A., J. R. Landis, and M. M. Cooper. 1982. A note on the equivalence of several marginal homogeneity test criteria for categorical data. *Internat. Statist. Rev.* **50**: 27–34.

Whitehead, J. 1993. Sample size calculations for ordered categorical data. *Statist. Medic.* **12**: 2257–2271.

Whittaker, J. 1990. *Graphical Models in Applied Multivariate Statistics*. New York: Wiley.

Whittaker, J., and M. Aitkin. 1978. A flexible strategy for fitting complex log-linear models. *Biometrics* **34**: 487–495.

Whittemore, A. S. 1978. Collapsibility of multidimensional tables. *J. Roy. Statist. Soc. Ser. B* **40**: 328–340.

Whittemore, A. S. 1981. Sample size for logistic regression with small response probability. *J. Amer. Statist. Assoc.* **76**: 27–32.

Wilks, S. S. 1935. The likelihood test of independence in contingency tables. *Ann. Math. Statist.* **6**: 190–196.

Wilks, S. S. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.* **9**: 60–62.

Williams, D. A. 1975. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31**: 949–952.

Williams, D. A. 1982. Extra-binomial variation in logistic linear models. *Appl. Statist.* **31**: 144–148.

Williams, D. A. 1987. Generalized linear model diagnostics using the deviance and single-case deletions. *Appl. Statist.* **36**: 181–191.

Williams, D. A. 1988. Comments on "The impact of litter effects on dose–response modeling in teratology." *Biometrics* **44**: 305–308.

Williams, E. J. 1952. Use of scores for the analysis of association in contingency tables. *Biometrika* **39**: 274–289.

Williams, O. D., and J. E. Grizzle. 1972. Analysis for contingency tables having ordered response categories. *J. Amer. Statist. Assoc.* **67**: 55–63.

Wilson, E. B. 1927. Probable inference, the law of succession, and statistical inference. *J. Amer. Statist. Assoc.* **22**: 209–212.

Wolfinger, R., and M. O'Connell. 1993. Generalized linear mixed models: A pseudo-likelihood approach. *J. Statist. Comput. Simul.* **48**: 233–243.

Wong, G. Y., and W. M. Mason. 1985. The hierarchical logistic regression model for multilevel analysis. *J. Amer. Statist. Assoc.* **80**: 513–524.

Woolf, B. 1955. On estimating the relation between blood group and disease. *Ann. Human Genet.* (*London*) **19**: 251–253.

Woolson, R. F., and W. R. Clarke. 1984. Analysis of categorical incomplete longitudinal data. *J. Roy. Statist. Soc. Ser. A* **147**: 87–99.

Wu, C. F. J. 1985. Efficient sequential designs with binary data. *J. Amer. Statist. Soc.* **80**: 974–984.

Yang, I., and M. P. Becker. 1997. Latent variable modeling of diagnostic accuracy. *Biometrics* **53**: 948–958.

Yates, F. 1934. Contingency tables involving small numbers and the χ^2 test. *J. Roy. Statist. Soc. Suppl.* **1**: 217–235.

Yates, F. 1948. The analysis of contingency tables with grouping based on quantitative characters. *Biometrika* **35**: 176–181.

Yates, F. 1984. Tests of significance for 2×2 contingency tables. *J. Roy. Statist. Soc. Ser. A* **147**: 426–463.

Yee, T. W., and C. J. Wild. 1996. Vector generalized additive models. *J. Roy. Statist. Soc. Ser. B* **58**: 481–493.

Yerushalmy, J. 1947. Statistical problems in assessing methods of medical diagnosis, with special reference to x-ray techniques. *Public Health Rep.* **62**: 1432–1449.

Yule, G. U. 1900. On the association of attributes in statistics. *Philos. Trans. Roy. Soc. London Ser. A* **194**: 257–319.

Yule, G. U. 1903. Notes on the theory of association of attributes in statistics. *Biometrika* **2**: 121–134.

Yule, G. U. 1906. On a property which holds good for all groupings of a normal distribution of frequency for two variables, with application to the study of contingency tables for the inheritance of unmeasured qualities. *Proc. Roy. Soc. Ser A* **77**: 324–336.

Yule, G. U. 1912. On the methods of measuring association between two attributes. *J. Roy. Statist. Soc.* **75**: 579–642.

Zeger, S. L., and M. R. Karim. 1991. Generalized linear models with random effects: A Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86**: 79–86

Zeger, S. L., K.-Y. Liang, and P. S. Albert. 1988. Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**: 1049–1060.

Zelen, M. 1971. The analysis of several 2×2 contingency tables. *Biometrika* **58**: 129–137.

Zelen, M. 1991. Multinomial response models. *Comput. Statist. Data Anal.* **12**: 249–254.

Zellner, A., and P. E. Rossi. 1984. Bayesian analysis of dichotomous quantal response models. *J. Economet.* **25**: 365–393.

Zelterman. D. 1987. Goodness-of-fit tests for large sparse multinomial distributions. *J. Amer. Statist. Soc.* **82**: 624–629.

Zermelo, E. 1929. Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Math. Z.* **29**: 436–460.

Zhang, H., J. Crowley, H. Sox, and R. Olshen. 1998. Tree-structured statistical methods. Pp. 4561–4573 in *Encyclopedia of Biostatistics*. Chichester, UK: Wiley.

Zheng, B., and A. Agresti. 2000. Summarizing the predictive power of a generalized linear model. *Statist. Medic.* **19**: 1771–1781.

Zhu, Y., and N. Reid. 1994. Information, ancillarity, and sufficiency in the presence of nuisance parameters. *Canad. J. Statist.* **22**: 111–123.

Categorical Data Analysis, Second Edition. Alan Agresti

ISBN: 0-471-36093-7

Author Index

Adelbasit, K. M., 196
Agresti, A., 27, 32, 33, 60, 100, 101, 102, 104, 111, 156, 227, 255, 258, 266, 298, 301, 379, 384, 397, 399, 422, 426, 435, 443, 445, 453, 465, 481, 485, 491, 502, 511, 513, 517, 518, 526, 527, 533, 536, 551, 552, 565, 567, 596, 630, 651
Aitchison, J., 265, 301, 465, 561, 613, 625
Aitken, C. G. G., 613
Aitkin M., 155, 388, 398, 399, 495, 520, 526, 545, 565, 633
Albert, A., 195, 197
Albert, J. H., 607, 609
Albert, P. S., 688
Allison, P. D., 633, 643
Altham, P. M. E., xv, 59, 103, 104, 240, 442, 443, 555, 566, 573, 587, 608
Amemiya, T., 227, 258, 300
Andersen, E. B., 255, 399, 450, 496, 526, 576, 631
Anderson, C. J., 398, 399
Anderson, D. A., 520, 526
Anderson, D. R., 216
Anderson, J. A., 171, 195, 196, 197, 207, 277
Anderson, R. L., 631
Anderson, T. W., 478, 482, 490
Aranda-Ordaz, F. J., 250, 399
Arminger, G., 549
Armitage, P., 104, 181
Ashford, J. R., 258, 379
Asmussen, S., 360
Azzalini, A., 480

Baglivo, J., 346
Baker, R. J., 283
Baker, S. G., 393, 482, 541
Banerjee, C., 443
Baptista, J., 100
Barnard, G. A., 95, 104, 114
Barndorff-Nielsen, O. E., 266
Bartholomew, D. J., 526, 565
Bartlett, M. S., 265, 623–624, 631
Becker, M., 370, 399, 435, 443, 544, 644
Bedrick, E. J., 77, 103
Begg, C. B., 273
Beitler, P. J., 508
Benedetti, J. K., 102, 398
Benichou, J., 66
Benzécri, J. P., 399, 624
Berger, R., 18, 33, 95, 594
Bergsma, W. P., 481
Berkson, J., 80, 104, 166, 197, 612, 624, 631
Berry, G., 104
Berry, S. M., 207
Best, D. J., 103
Bhapkar, V. P., 27, 103, 104, 291, 422, 453, 488, 612, 615, 616, 629
Bickel, P., 63
Biggeri, A., 566
Billingsley, P., 482
Birch, M. W., 255, 263, 295, 298, 336, 339, 340, 341, 346, 369, 392, 576, 585, 627, 631
Bishop, Y. M. M., 347, 360, 366, 452, 482, 526, 576, 582, 587, 591, 594, 627, 629, 631
Blaker, H., 20, 27, 93, 635
Bliss, C., 246, 247, 560, 623
Blyth, C. R., 20, 27, 32, 59
Bock, R. D., 300, 301, 495, 526, 624, 625
Böckenholt, U., 398, 399, 443
Bonney, G. E., 479, 480
Boos, D. D., 95, 467, 481
Booth, J., xv, 104, 223, 397, 443, 523, 525, 567, 630
Bowker, A. H., 424
Box, J. F., 23, 92, 623, 624
Bradley, R. A., 302, 436, 443
Breslow, N., 51, 59, 155, 156, 171, 234, 235, 255, 258, 399, 419, 493, 523, 524, 563, 625, 631
Brier, S. S., 515
Brooks, S. P., 566
Bross, I. D. J., 111

Brown, L. D., 15, 27, 33, 606
Brown, M. B., 102, 398
Brown, P. J., 613, 614
Brownstone, D., 302, 527
Bull, S. B., 196
Burnham, K. P., 216, 526, 552
Burridge, J., 283
Butler, R., 104, 397, 443, 630
Byar, D. P., 232, 295, 414, 481, 625

Caffo, B., xv, 102, 656
Cameron, A. C., 131, 155, 561, 566, 574
Carey, V., 474
Carroll, R. J., 171, 467
Casella, G., 18, 33, 594
Catalano, P. J., 527
Caussinus, H., 425, 427, 428, 443, 451, 627, 631
Chaloner, K., 196, 609
Chamberlain, G., 419, 420, 526
Chambers, E. A., 258
Chambers, J. M., 633
Chambers, R. L., 527
Chan, I., 104
Chan, J. S. K., 523
Chao, A., 513, 526, 533
Chapman, D. G., 258
Chen, Z., 527, 643, 651
Chib, S., 609
Christensen, R., 196
Chuang, C., 399, 462
Clayton D. G., 388, 399, 493, 523, 563, 625, 631
Clogg, C. C., 103, 391, 399, 565, 627
Clopper, C. J., 18
Cochran W. G., 27, 80, 88, 163, 181, 232, 239, 396, 459, 488, 596, 626, 627, 631
Coe, P. R., 101
Cohen, A., 103, 104
Cohen, J., 434, 435, 443
Coleman, J. S., 516, 532
Collett, D., 196, 204
Conaway, M. R., 426, 482, 526, 565
Cook, R. D., 225
Copas, J. B., 156, 257, 442, 616
Corcoran, C., 197, 573
Cormack, R. M., 511, 526, 551
Cornfield, J., 42, 47, 51, 71, 77, 99, 100, 171, 196, 208, 221, 624
Coull, B. A., xv, 27, 32, 33, 513, 518, 526, 533, 552, 567, 655, 662
Cox, C., 266, 282, 286, 576, 587, 641
Cox, D. R., 12, 104, 133, 138, 196, 197, 258, 415, 482, 493, 497, 624, 625, 631
Cramér, H., 112, 576, 587, 625–626
Cressie, N., 27, 112, 258, 396, 612
Croon, M., 481
Crouchley, R., 527
Crowder, M. J., 555, 566

D'Agostino, R. B., Jr., 196
Dalal, S. R., 199
Daniels, M. J., 524, 609
Dardanoni, V., 301
Darroch, J. N., 347, 357, 398, 414, 426, 443, 459, 481, 513, 526, 551, 552, 565, 626, 629, 631
Das Gupta, S., 237
David, H. A., 443
Davis, L. J., 59, 93, 398
Davison, A. C., 156, 594
Dawson, R. B., Jr., 103
Dawson, R. J. M., 61
Day, N. E., 51, 171, 232, 235, 258, 399, 625
de Falguerolles, A., 399, 664, 686
de Leeuw, J., 399
Demétrio, C. G. B., 156, 555, 566
Deming, W. E., 343, 347, 511
Dempster, A. P., 522
Dey, D. K., 609
Diaconis, P., 103, 104
Diggle, P., 471, 625
Dillon, W., 443
Dittrich, R., xv, 443
Dobson, A. J., 155
Doll, R., 42, 62, 64, 404
Dong, J., xv, 59, 614, 616
Donner, A., 172, 196, 258
Doolittle, M. H., 621
Downing, D., 103
Drost, F. C., 112, 258, 595
Ducharme, G. R., 398
Duncan, D. B., 195, 197, 277, 301, 624
Dupont, W. D., 93
Dyke, G. V., 624

Edwardes, M. D., 301
Edwards, A. W. F., 59
Edwards, D., 360, 398
Efron, B., 103, 146, 196, 227, 258, 526, 605, 610
Ekholm, A., 156, 481
Eliason, S. R., 103, 391
Escoufier, Y., 383, 399
Espeland, M. A., 544, 571
Everitt, B. S., 633

Fahrmeir, L., 155, 300, 615
Farewell, V. T., 171, 301, 527, 625

Fay, R., 103, 482, 594
Fechner, G., 623
Ferguson, T. S., 605, 617
Fienberg, S. E., 344, 347, 392, 438, 443, 513, 526, 610, 615, 616, 626, 627, 629, 631
Finney, D., 151, 258, 556, 623
Firth, D., 155, 156, 196, 330, 467, 481, 482
Fischer, G. H., 526
Fisher, R. A., 12, 22, 23, 29, 51, 79, 91, 92, 95, 99, 104, 114, 146, 156, 162, 237, 247, 560, 576, 589, 622–624, 625, 626, 628, 631
Fitzmaurice, G. M., 103, 466, 474, 481, 482, 649
Fitzpatrick, S., 35
Fleiss, J. L., 104, 110, 111, 242, 258, 347, 435, 436, 443
Follman, D. A., 546, 547, 548, 566
Forcina, A., 301
Forster, J. J., 104, 346, 397, 482, 616, 630
Forthofer, R. N., 378
Fowlkes, E. B., 199, 226, 257
Francom, S., 462
Freedman, D., 23, 63
Freeman, D. H. Jr., 399
Freeman, G. H., 97
Freeman, M. F., 112
Freidlin, B., 104
Friendly, M., 59, 399, 633
Frome, E. L., 155, 399
Fuchs, C., 482

Gabriel, K. R., 263, 399
Gaddum, J. H., 623
Gail, M. H., 104, 625
Gart, J. J., 70, 71, 77, 102, 104, 197, 255, 258, 397, 442
Gaskins, R. A., 614
Gastwirth, J., 104, 197
Gatsonis, C., xv, 230, 481, 524, 609
Gelfand, A. E., 609
Genter, F. C., 301
Geyer, C., 678
Ghosh, B. K., 27
Ghosh, M., 442, 609
Gibbons, R. D., 520
Gilbert, G. N., 217
Gill, J., 155
Gilmour, A. R., 526
Gilula, Z., 83, 382, 384
Gini, C., 329
Glass, P. V., 447
Gleser, L. J., 103
Glonek, G. F. V., 393, 466
Godambe, V. P., 104, 482
Goetghebeur, E., 482
Gokhale, D. V., 112, 612, 616
Goldstein, H., 520, 524
Good, I. J., 24, 60, 104, 605, 607, 608, 612, 614, 616, 626, 630
Goodman L. A., 35, 59, 68, 69, 83, 84, 102, 110, 213, 217, 228, 340, 346, 365, 366, 369, 370, 374, 379, 380, 381, 382, 383, 384, 397, 398, 399, 406, 407, 408, 425, 428, 431, 443, 478, 482, 490, 516, 527, 540, 565, 566, 572, 621, 622, 627, 628, 629, 631
Gould, S. J., 544
Gourieroux, C., 467, 482
Graubard, B. I., 89, 103
Gray, R., 273
Green, P. J., 156
Greenacre, M. J., 384, 399
Greene, G., 28
Greenland, S., 96, 234, 258
Greenwood, M., 566
Greenwood, P. E., 27
Grego, J., 686
Grizzle, J. E., 291, 301, 457, 601, 615, 624, 629, 631
Gross, S. T., 197
Grove, W. M., 544
Gueorguieva, R., xv, 527, 670
Gupta, A. K., 607

Haber, M., 95, 96, 103, 291, 465
Haberman, S. J., 69, 81, 83, 113, 195, 224, 258, 268, 300, 349, 346, 347, 364, 367, 369, 374, 380, 382, 392, 393, 396, 399, 408, 440, 526, 540, 565, 572, 576, 589, 591, 592, 595, 627, 629, 631
Hagenaars, J., 565
Hald, A., 623
Haldane, J. B. S., 70, 103, 196
Hall, P., 616
Halton, J. H., 97
Hamada, M., 301
Handelman, S. L., 544, 571
Hanfelt, J., 566, 571
Hansen, L. P., 467, 482
Harkness, W. L., 258
Harrell, F. E., 229, 282, 301
Hartzel, J., 511, 513, 514, 516, 534, 651
Haslett, S., 394
Hastie, T., 153, 199, 301, 633
Hatzinger, R., 565
Hauck, W. W., 172, 234, 258
Haynam, G. F., 243
Heagerty, P., 481, 527, 548
Hedeker, D., 520, 653

Heinen, T., 565
Hennevogl, W., 513
Henry, N. W., 565
Heyde, C. C., 156, 481
Hill, A. B., 42, 64, 111, 404
Hinde, J., 155, 156, 555, 563, 566
Hinkley, D., 12, 104, 133, 138, 146, 156, 594
Hirji, K. F., 104, 258, 625
Hirotsu, C., 406
Hirschfeld, H., 399
Hoadley, B., 199
Hobert, J., xv, 523, 525, 630
Hodges, J. L., 197
Hoem, J. M., 347, 399
Holford, T. R., 389, 390, 399
Holland, P. W., 610, 615, 629, 631
Hollander, M., 443
Holt, D., 103
Holtbrügge, 306
Hook, E. B., 526
Hosmer, D. W., 177, 196, 197, 257, 258
Hout, M., 65, 428, 443
Howard, J. V., 104, 608
Hsieh, F. Y., 242, 243
Hsu, J. S. J., 608
Hwang, J. T. G., 104

Imrey, P. B., 346, 443, 615
Ireland, C. T., 616
Irwin, J., 91

Jennison, C., 103
Jewell, N. P., 467, 496, 566
Johnson, B. M., 605
Johnson, N. L., 566, 574
Johnson, W., 257
Jones, B., 442, 484, 536
Jones, M. P., 258
Jørgensen, B., 136, 155, 156, 266, 470

Kalbfleisch, J. D., 482
Karim, M. R., 524, 609
Kastenbaum, M. A., 627
Kastner, C., 466, 649
Katzenbeisser, W., xv, 672
Kauerman, G., 156, 467
Kelderman, H., 565
Kempthorne, O., 96, 104
Kendall M. G., 27, 56, 60, 68, 399, 631
Kenward, M. G., 442, 475, 484, 536
Khamis, H. J., xv, 332, 443
Kim, D., 104, 255, 298, 379, 397
King, G., 527
Knott, M., 565
Knuiman, M. W., 616
Koch, G. G., 27, 302, 436, 447, 459, 460, 481, 532, 601, 615, 616, 629, 631, 670, 673, 674, 675
Koehler, K., 27, 103, 396, 397
Koopman, P. A. R., 77
Korn, E. L., 89, 103
Kraemer, H. C., 443
Kreiner, S., xv, 358, 398
Kruskal, W. H., 59, 60, 68, 69, 102, 110, 621, 631
Ku, H. H., 616
Kuha, J., 330, 347
Kuk, A. Y. C., 523
Kullback, S., 112, 399, 612, 616
Kuo, L., 527, 643, 651
Kupper, L. L., 566

Läärä, F., 301, 313
Lachin, J., 258
Laird, N. M., 385, 386, 389, 466, 482, 522, 541, 609, 610, 649
Lambert, D., 546, 547, 548, 566
Lancaster, H., 20, 27, 83, 84, 113, 399, 626
Landis, J. R., 111, 295, 297, 301, 302, 436, 447, 462, 508, 532
Landrum, M., 609
Landwehr, J. M., 226, 257
Lang, J. B., xv, 301, 340, 399, 465, 481, 537, 541, 551, 643, 644, 649, 655, 675, 678
Laplace, P. S., 15
Larntz, K., 196, 396, 397, 438, 443, 609
Larsen, K., 498
Larson, M. G., 399
Lauritzen, S. L., 346, 398, 399
LaVange, L. M., 103, 197, 481
Lawal, H. B., 396
Lawless, J. F., 155, 482, 560, 561, 566
Lazarsfeld, P. F., 565
Lee, E., 198
Lee, S. K., 596
Lee, Y., 559, 566, 574
Lefkopoulou, M., 566
Lehmann, E., 67, 104, 263, 406
Lehnen, R. G., 378
Lemeshow, S., 177, 196, 197, 257, 258
Leonard, T., 608, 609
Lesaffre, E., 258, 300, 466, 522, 545
Lesperance, M. L., 526, 548
Liang, K.-Y., 104, 258, 442, 467, 469, 471, 473, 481, 482, 525, 556, 566, 571, 573, 625, 631
Lin, X., 524, 525
Lindley, D. V., 609, 630
Lindsay, B., 494, 545, 549

Lindsey, J. K., 400, 467, 566, 573
Lipsitz, S., 103, 291, 422, 456, 469, 473, 474, 481, 645
Little, R. J., 114, 346, 347, 475, 476, 482
Liu, I., xv, 485, 655, 671
Liu, Q., 510, 522
Lloyd, C., 93, 104, 156, 615
Localio, A. R., 526
Longford, N. T., 520
Loughin, T., 484
Louis, T., 541
Luce, R., 299, 302, 443

Madansky, A., 422, 456
Maddala, G. S., 258, 264, 302
Magidson, J., 653
Magnus, J. R., 602
Mansmann, U., 104
Mantel, N., 87, 93, 104, 171, 197, 209, 230, 231, 232, 234, 238, 260, 295, 296, 297, 300, 379, 414, 481, 612, 618, 624, 625, 627, 631
Martin Andres, A., 104
Mason, W. M., 524, 609
Matthews, J. N. S., 301, 313, 443
Maxwell, A. E., 631
McArdle, J. J., 202
McCloud, P. I., 443
McCullagh, P., 132, 155, 156, 257, 276, 277, 283, 286, 290, 301, 308, 312, 340, 378, 397, 431, 443, 466, 471, 481, 556, 566, 625, 631
McCulloch, C. E., 522, 523, 524, 527, 548, 555, 623, 625
McDonald, J. W., 667, 684
McFadden, D., 228, 264, 299, 300, 302, 302, 624, 631
McNemar, Q., 411
Mee, R. W., 77
Meeden, G., 605
Mehta, C. R., 98, 104, 254, 255, 258, 298, 397, 625, 630
Mendel, G., 22
Mendenhall, W. M., 107
Mersch, G., 400
Michailidis, G., 399
Miettinen, O. S., 77, 442
Miller, M. E., 481, 604
Min, Y., 100, 101
Minkin, S., 196
Mirkin, B., 112
Mitra, S. K., 79, 258, 346, 591, 627, 631
Mittal, Y., 60
Molenaar, I. W., 526
Molenberghs, G., 258, 466, 482
Moore, D. F., 152, 556, 566
Moore, D. S., 27, 103
Morabia, A., 527
Morgan, B. J. T., 196, 207
Morgan, W. M., 346
Morris, C., 526, 605, 610
Mosimann, J. E., 566
Mosteller, F., 345, 412, 443, 627, 629, 631

Nair, V. N., 103, 301
Nam, J., 77
Natarajan, R., xv, 481, 502, 524
Nelder, J., 116, 132, 148, 149, 155, 156, 257, 290, 301, 312, 340, 378, 559, 566, 574, 625, 631
Nerlove, M., 300, 624
Neudecker, H., 602
Neuhaus, J. M., 417, 494, 496, 499, 502, 526, 547, 548, 566
Newcombe, R., 27, 109, 110
Neyman, J., 18, 112, 611, 612, 616, 626, 631
Nikulin, M. S., 27
Normand, S.-L., 609
Norusis, M. J., 633
Nurminen, M., 77

O'Brien, P. C., 207
O'Brien, R. G., 244, 258, 640
Ochi, Y., 258
Odoroff, C., 661
O'Gorman, T. W., 596
Olivier, D., 385, 386, 389
Overton, W. S., 526, 552

Pagano M., 61, 657
Paik, M., 59
Palmgren, J., 156, 340
Park, T., 482
Parr, W. C., 594
Parzen, E., 34
Patefield, W. M., 104
Patel, N. R., 98, 258, 625, 630
Patnaik, P. B., 258
Paul, S. R., 566
Pearson, E. S., 18, 104, 626, 631
Pearson, K., 22, 79, 112, 399, 576, 589, 620, 621, 622, 628, 631
Peduzzi, P., 212
Pendergast, J. F., xv, 502
Pepe, M. S., 258
Perlman, M., 237
Peters, D., 104, 630
Peterson, B., 282, 301
Peto, R., 62
Piccarreta, R., 206

Pierce, D. A., 104, 143, 156, 497, 502, 522, 526, 630
Pike, M. C., 100
Piegorsch, W. W., 684
Plackett, R. L., 103, 196, 399, 623, 627, 631
Podgor, M. J., 197
Poisson, S.-D., 7
Pratt, J. W., 283
Pregibon, D., 143, 156, 197, 225, 257, 258, 566, 638
Prentice, R. L., 171, 196, 258, 283, 399, 482, 555, 566, 625
Presnell, B., 156
Press, S. J., 196, 300, 624
Pyke, R., 171, 625

Qaqish, B., 676
Qu, A., 482
Quetelet, A., 68
Quine, M. P., 29

Rabe-Hesketh, S., 527, 633
Radelet, M., 48, 65
Raftery, A., 257
Rao, C. R., 10, 12, 576, 582, 585, 587, 589, 591, 596, 616, 626, 631
Rao, J. N. K., 103, 515
Rasbash, J., 520, 524
Rasch, G., 399, 415, 493, 495, 624
Rayner, J. C. W., 103
Read, T. R. C., 27, 112, 258, 396, 612
Regal, R. R., 526
Reid, N., 96
Rice, W. R., 104
Ripley, B., 633
Ritov, Y., 384
Robins, J., 234, 258, 475
Röhmel, J., 104
Rosenbaum, P. R., 196
Rosner, B., 566
Rossi, P. E., 609
Rotnitzky, A., 467
Routledge, R. D., 104, 607
Roy, S. N., 79, 346, 627, 631
Rubin, D., 196, 475, 482
Rudas, T., 481, 565
Rundell, P. W. K., 613, 614
Ryan, L., 290, 527, 566

Sackrowitz, H. B., 103, 104
Samuels, M. L., 60
Santner, T. J., 101
Schafer, D. W., 143, 156
Schafer, J. L., 103, 347, 482
Schluchter, M. D., 399
Schumacher, M., 306
Scott, A. J., 35, 103, 197
Searle, S., 527, 555
Seeber, G., 155
Sekar, C. C., 511
Self, S. G., 258, 525
Sen, P. K., 594
Seneta E., 29
Silvey, S. D., 301, 465, 625
Singer, J. M., 594
Shapiro, S. H., 398
Shen, S. M., 265
Shihadeh, E. S., 399
Shuster, J. J., 95, 103, 104, 442
Silva Mato, A., 104
Silvapulle, M. J., 195
Silvey, S. D., 301, 465, 625
Simon, G., 197, 301, 374, 399, 612, 624, 629
Simonoff, J., 594, 614, 615, 616
Simpson, E. H., 51, 60, 398, 596, 621
Singer, J. M., 594
Skellam, J. G., 566
Skene, A. M., 502, 609
Skinner, C., 347
Skrondal, A., 527
Slaton, T. L., 566
Small, K. A., xv, 302
Smith A. F. M., 609, 616
Smith, K. W., 345
Smith, P. W. F., 443, 482, 616
Snell, E. J., 196, 301, 624
Snell, M. K., 101
Sobel, M. E., 672
Somers, R. H., 68
Somes, G. W., 488
Speed, T., 347, 616
Spiegelhalter, D., 616
Spitzer, R. L., 435
Sprott, D. A., 95, 114, 453
Starmer, C. F., 601, 629
Stasinopoulos, M., 526
Stern, S., 302
Sterne, T. E., 20
Stevens, S. S., 26
Stigler, S., 22, 443, 448, 623, 631
Still, H. A., 20, 27, 32
Stiratelli, R., 482, 526
Stokes M. E., xv, 282, 302, 399, 476, 482, 633, 640, 649
Strawderman, R. L., 104, 630
Stuart, A., 27, 56, 399, 422
Stukel, T. A., 196, 250
Sturmfels, B., 104

Suissa, S., 95, 104, 442
Sundberg, R., 346, 366

Tamhane, A. C., 101
Tango, T., 411
Tanner, M. A., 443
Tarone, R. E., 197, 234, 258
Tavaré, S., 103
Ten Have, T. R., xv, 517, 526, 527, 527
Theil, H., 57, 228, 300, 624
Thomas, D. R., 103, 515
Thompson, R., 283
Thompson, W. A., 399
Thurstone, L. L., 443
Tibshirani, R., 153, 199, 301
Titterington, D. M., 616
Tjur, T., 426, 552, 553
Tocher, K. D., 94
Toledano, A., 230, 481
Tolley, H. D., 594
Train, K., 302, 527
Trivedi, P. K., 131, 155, 561, 566, 574
Tsiatis, A. A., 152, 197, 556, 566
Tukey, J., 112
Turing, A., 631
Turnbull, B. W., 103
Tutz, G., 155, 156, 289, 290, 300, 301, 513, 615

Uebersax, J. S., 544
Uttal, D. H., 517

van der Heijden, P. G., 399
Venables, W. N., 633
Verbeke, G., 482, 545
Vermunt, J. K., 399, 653

Wainer, H., 63
Wakefield, J. C., 502, 609
Wald, A., 11, 172
Walker, S. H., 195, 197, 277, 301, 624
Walley, P., 616
Walsh, S. J., 104
Wardrop, R. L., 105
Ware, J. H., 478, 480, 482
Watson, G. S., 79, 103, 576, 590, 627, 631
Wedderburn, R. W. M., 116, 148, 149, 150, 155, 156, 195, 258, 265, 266, 466, 470, 625, 631
Weisberg, S., 226
Wells, M. T., 104, 630
Wermuth, N., 398, 399, 401
Westfall, P. H., 214, 360
White, A. A., 481
White, H., 467, 471, 482
Whitehead, J., 301
Whittaker, J., 346, 358, 398, 399
Whittemore, A. S., 243, 398
Wild, C., 103, 197, 301
Wilks, S. S., 12
Williams, D. A., 156, 225, 397, 555, 566, 653
Williams, E. J., 103, 399
Williams, O. D., 291, 301, 624
Wilson, E. B., 16
Wilson, J., 103
Winner, L., 445
Wolfinger, R. D., 214, 360, 527
Wong, G. Y., 524, 609
Woolf, B., 71
Woolson, R. F., 487, 596
Wu, C. F. J., 196, 301
Wu, M., 346, 347

Yamagami, S., 101
Yang, I., 544
Yang, M., 104
Yates F., 91, 93, 96, 98, 103, 104, 114, 239, 624
Yee, T. W., 301
Yerushalmy, J., 38
Young, S. S., 214, 360
Yule, G. U., 44, 53, 59, 68, 110, 346, 406, 566, 620–621, 628, 631

Zeger, S. L., 442, 467, 469, 471, 473, 481, 482, 499, 500, 524, 548, 609, 625, 631
Zelen, M., 255, 625, 631
Zellner, A., 609
Zelterman, D., 397
Zermelo, E., 443
Zhang, H., 257
Zhao, L., 482
Zheng, B., 227, 258, 266
Zhu, Y., 96
Zweiful, J. R., 70, 397

Categorical Data Analysis, Second Edition. Alan Agresti

ISBN: 0-471-36093-7

Examples Index

Abortion and education, 345
Abortion opinions, 29, 205–206, 441, 486, 504–506, 553
Admissions into Berkeley, 62–63
Admissions into Florida, 223–224, 529
Afterlife, belief in, 302–303
AIDS and AZT use, 184–187
AIDS, measures to deal with, 347
Air pollution and breathing, 377–378
Alcohol, cigarettes, and marijuana use, 322–326, 361–363, 367, 482–483, 528
Alcohol consumption and malformation, 89–90, 158, 179–180, 182
Alcohol and driving, 203
Alligator food choice, 268–274, 304
Alzheimer's disease and cognitive impairment, 310
Aspirations by income, 107,
Aspirin and heart attacks, 37, 46, 71–72
Automobile collisions and seat belts, 40–41, 61, 305–306, 327–329, 331, 349, 361

Baseball complete games, 157–158
Baseball standings, 437–438
Beetle mortality, 247–250
Birth control, teenage, 352
Blood pressure and heart disease, 221–223
Breast cancer, 38, 105, 107
Breathing test and smoking, 307, 377–378
Breathlessness, wheeze, and age, 378
Buchanan vote in Palm Beach County, 156–157
Busing and race, 348

Calves and pneumonia, 25–26, 34
Cancer of larynx and radiation therapy, 107
Cancer remission, 197–199, 261
Capture–recapture, hepatitis, 533
Capture–recapture of snowshoe hares, 511–513, 544–545, 551–552
Carcinoma of uterine cervix, 431–435, 532, 541–544, 549–551
Chlorophyll inheritance, 29
Cholesterol and cereal, 309
Claritin, 109
Clinical trials, 230–236, 507–510
Coffee drinking, 446
Cola drink taste test, 448
Condoms and adolescents, 202
Coronary deaths and smoking, 404
Credit card and income (Italy), 206
Crime and race, 63
Crossover drug trial, 457, 483–484

Death penalty and race, 48–52, 63, 65, 201
Depression, mental, 459–461, 468–469, 506–507
Developmental toxicity study, 290–291, 517–521
Diabetes, case-control study, 418–419
Diagnostic tests, 60, 66
Diarrhea, 255
Draft position in sports, 207
Dumping severity, 308–309
Dysmenorrhea, 483–484, 572

Esophageal cancer, 203

Fish egg hatching, 568–569
Free throws, 105, 160–161

Gambler's ruin, 489–490
Genetics, 165
Government spending, 349–351, 449, 530–531
Graduate admissions at Florida, 223–224, 529
Graduate admissions at Berkeley, 62–63

Graham Greene, 28
Gun-related deaths, 61

Heart attacks and aspirin use, 37, 46
Heart catheterization and race, 62
Heart disease and blood pressure, 221–223
Heart disease and snoring, 121–123
Heart valve replacement and survival, 385–387
Hepatitis outbreaks, 533
Home team advantage in baseball, 437–438
Homicide victims, number, 561–563, 564–565, 571
Horseshoe crab mating, 126–131, 154–155, 159, 168–170, 173–176, 188–192, 212–216, 570

Income by year, 308
Infant survival, gestation, smoking, and age, 400–401
Insomnia, 462–464, 469, 487, 514–515, 531

Job satisfaction and income, 57–59, 87–88, 287–288, 295, 297, 308
Job satisfaction and race, gender, age, and location, 205
Journal citations, 448

Kyphosis and spinal surgery, 199–200

Labelling index and remission, 197–199, 261
Larry Bird free throws, 105
Leading crowd, 516–517, 532
Leprosy, 239
Life table, 284
Lung cancer and chemotherapy, 306,
Lung cancer and smoking, 42, 61, 62, 64
Lung cancer survival, 390–391

Malformation of infants, 89–90, 158, 179–180, 182
Mendel's theories, 22–23
Mental health, and parents SES, 381, 383–384
Mental impairment, life events and SES, 279–282
Migration, 423, 427–428
Missing people in London, 202
Mixture for two protozoan genuses, 546
Motor vehicle accident rates, 403
Movie reviewers, 445–446
Multicenter clinical trial, infection cream, 230–235, 508–510
Multicenter clinical trial, fungal infections, 394–395, 530
Multiple sclerosis and neurologist ratings, 447
Murder rates in U.S., 62, 63
Myocardial infarction and aspirin, 37, 46, 71–72
Myocardial infarction and diabetes, 418–419

NCAA graduation rates, 202
Nervousness and Claritin, 109

Obesity, occasion and gender, 487
Occupational aspirations, 206
Occupational status, father and son, 447
Oral contraceptive use, 200
Osteosarcoma, 262–263

Palm Beach County vote for Buchanan, 156–157
Party identification by race and by gender, 105–106, 303
Party identification and protestors, 307
Pathologists ratings of carcinoma, 431–435, 532, 541–544, 549–551
Penicillin and rabbits, 259–260
Pig farmer survey, 484–485
Pneumonia infections, 25–26, 34
Poison dose for protozoa, 546–547
Political ideology and party affiliation, 305, 375–377
Pregnancy rates, 567
Presidential approval rating, 409–412
Presidential vote, by state, 503–504, 534
Promotion discrimination, 254–255
Prussian army and mule kicks, 30
Psychiatric patients and prescribed drugs, 106–107

Religious fundamentalism and education, 80, 81–82
Religious services, frequency of attendance, 352
Respiratory illness, age and maternal smoking, 480–481
Respiratory illness in children, 478–479

Satisfaction with housing, 310
Satisfaction with job, 205
Schizophrenia origin, 83–84
Seat belts and injury, 40–41, 61, 305–306, 327–329, 331, 349, 361
Sex, frequency of, 569–570
Sex opinions, 65, 217–219, 368, 371–373, 421, 430, 431, 530
Sexual intercourse, gender and race, 201

Shopping choice, 300
Snowshoe hares, 511–513, 544–545, 551–552
Soccer and arrests, 403
Sore throat in surgery, 204
Space shuttle, 199
Student survey (alcohol, marijuana, cigarettes), 322–326, 361–363, 367, 482–483, 528

Tea drinker, 92, 100
Teenage birth control, 368, 371–373
Tennis rankings, 449
Teratology studies, 151–153
Titanic, 61
Toxicity study, 517–520
Train accidents, 403, 569

UFOs, 106

Vegetarianism, 16–17, 29
Veterinary information sources, 484–485
Voting, proportion by state, 503–504, 534

Categorical Data Analysis, Second Edition. Alan Agresti

ISBN: 0-471-36093-7

Subject Index

Adjacent categories logit, 286–288, 370–371, 374–376, 642
Adjusted residual, *see* Standardized Pearson residual
Agreement, 431–436, 443, 453–454, 541–544, 549–551
AIC, 216–217, 324
Alternating logistic regressions, 474
Ancillary statistic, 104
Arc sine transformation, 596
Armitage test, *see* Cochran–Armitage trend test
Association, *see* Measures of association
Association graphs, 357–360, 539
Association models, 373–381, 399
Asymptotic covariance matrix, 137–138, 577–581, 594
Asymptotic normality, 73–77, 577–581
Attributable risk, 66, 110

Backward elimination, 214–216
BAN, 611, 626
Baseline-category logits, 267–274, 300, 310–311, 426, 515, 640–643
Bayesian inference, 604–610, 616, 630–631
  binomial parameters, 605–607, 617
  generalized linear mixed models, 524, 609
  kernel smoothing, connection, 614
  multinomial proportions, 607–610, 618
Bernoulli distribution, 117
Beta-binomial distribution, 30, 553–559, 566, 572, 573, 653
Beta distribution, 554, 572, 605–606
Bias, 70, 85, 196, 450, 496, 524, 548, 595, 615
BIC, 257
Binary data
  correlated, 409–420, 455–482, 491–527, 538–559
  generalized linear models, 120–125, 137, 140
  matched pairs, 409–420
Binomial distribution, 5–6
  admissible estimator, 605
  confidence interval for proportion, 15–17, 32–33, 635
  exact inference, 18–20
  exponential family, 117, 134
  GLM likelihood equations, 137
  likelihood function, 9
  matched pairs, 409–420
  moment generating function, 31
  overdispersion, 8, 30
  tests for proportion, 14–15
  variance stabilizing, 596
Binomial models
  deviance, 140
  GLMs, 120–125
  likelihood equations, 137, 265
  overdispersion, 151–153, 291, 573, 653
Birch's results, 336
Bootstrap, 75, 156, 525, 531, 594
Bradley–Terry model, 436–439, 443, 647
Breslow–Day test, 258

Calibration, 207
Canonical correlation, 382, 399, 408, 624
Canonical link, 117, 148–149, 193, 257, 472, 496
Capture–recapture, 511–513, 526, 544–545, 551–552
CART, 257
Case-control study, 42–43, 46–47, 59, 233,
  and logistic regression, 170–171, 418–420, 625
  several controls per case, 233, 442
Categorical data analysis, 1–688
Causal diagram, 217–218
Censoring, 386, 400
Centering, 167, 175
Chi-squared distribution
  df, 12, 79, 175, 589

Chi-squared distribution (*Continued*)
mgf, 35
moments, 27
noncentral, 237, 258, 408, 591–592, 595, 597
reproductive property, 82
table of percentage points, 654
Chi-squared statistics
likelihood-ratio, *see* Likelihood-ratio statistic
partitioning, *see* Partitioning
Pearson, *see* Pearson chi-squared statistic
Classification methods, 196, 257, 228–230, 258
Clinical trials, 42, 230–236, 507–510
Clopper–Pearson confidence interval, 18–20, 33, 606
Cluster sampling, 103, 481, 515
Clustered data, 455, 491–527, 556–558
Cochran, W. G., 626
Cochran–Armitage trend test, 181–182, 197, 237, 253, 640
Cochran–Mantel–Haenszel test, 231–234, 639
exact test, 254, 298
and marginal homogeneity, 413, 458–459, 481
and McNemar test, 413–414
matched pairs, 413
nominal and ordinal cases, 295–298, 302, 379, 642–643
score test for logit model, 232, 297–298
Cochran's Q, 459, 488
Collapsibility, 358–360, 398
Complementary log–log model
binary response, 248–250, 640
ordinal response, 283–284, 301, 313, 527, 641
Computer software, *see* Software
Concentration coefficient, 69
Concordance index, 229
Concordant pair, 57–59
Conditional distribution, 37, 48
Conditional independence, 52
$I \times J \times K$ tables, 293–298, 302, 318–319, 325
logit models, 183–184, 230–234, 263, 293–295, 359–360
versus marginal independence, 53, 365–366
power and sample size, 244–245
small-sample test, 254, 298
Conditional inference, 91–101, 250–257, 416–420, 495–496, 630
Conditional logistic regression, 250–258, 414–420, 495–496, 526, 625, 640, 645
Conditional logit, 299
Conditional ML, 100, 417, 494–496, 526
Conditional symmetry, 431, 452
Confidence intervals
likelihood-based, 13, 77–78
tail method, 18, 99
Wald, 13
score, 15–16, 77
Confounding, 47–51, 230
Conjugate mixture model, 558–559
Constraint equations, 612
Constraints, parameter, 178–179, 317, 352–353
Contingency coefficient, 112, 620
Contingency table, 36, 47–54
Continuation-ratio logit, 289–291, 301, 517–520
Continuity correction, 27,
Continuous proportions, 265–266, 624
Contrasts, 82, 317, 340, 344, 603, 636, 639
Correlation, 87, 226, 296, 634
Correlation models, 381–384, 399, 408
Correspondence analysis, 382–384, 399, 624, 644
Cramér's V^2, 112
Credit scoring, 165, 263, 631
Cross-classification table, *see* Contingency table
Crossover study, 444, 457, 483, 498, 501, 572
Cross-product ratio, 44
Cross validation, 266
Cumulant function, 155
Cumulative link models, 282–286, 313
Cumulative logit models, 274–282, 301, 624, 641
dispersion effects, 285–286
marginal models, 420–421, 462–463, 469
proportional odds property, 275–276, 282
random effects, 514–515, 536
score test and ranks, 301
Cumulative odds ratio, 67, 276
Cumulative probit model, 278, 283, 301, 312, 624–625, 641

Data mining, 219, 631
Decomposable model, 346, 360
Degrees of freedom, 12, 79, 175, 589, 622
Delta method, 73–77, 577–581, 594
Dependent proportions, 410–412
Design, 196, 609
Design matrix, *see* Model matrix
Deviance, 118–119, 139–142
grouped vs. ungrouped binary data, 208
likelihood-ratio tests, 141–142, 186–187, 363–365
residual, 142, 220, 638
R-squared measures, 228

Diagnostics, 142–143, 219–230, 257–258, 366–367
Diagonals-parameter symmetry, 443
Difference of proportions, 43
  collapsibility, 398
  dependent, 410–412, 645
  homogeneity, 258
  large-sample confidence interval, 72, 77, 102, 110, 410–411
  sample size determination, 240–242, 258
  small-sample confidence interval, 101
  z test and Pearson statistic, 111
Directed alternatives, 88–90, 236–239, 373
Dirichlet distribution, 607, 610
Discordant pair, 57–59
Discrete choice models, 298–300, 302, 527, 624
Discreteness and conservatism, 18–20, 93–94, 257
Discriminant analysis, 196
Dispersion parameters, 131, 133, 285–286, 560
Dissimilarity index, 329–330
Diversity index, 596
Dummy variables, 178–179

Ecological inference, 527
Effect modifier, 54
EM algorithm, 522–523, 540–541
Empirical Bayes, 526, 610
Empirical logit, 168
Empty cells, 392
Entropy, 57, 613
Estimated expected frequencies, 25, 78, 315
Estimating equations, 470, 481–482
Exact confidence intervals, 18–20, 99–101, 255
Exact tests
  binomial parameter, 18, 412
  conditional independence, 254, 298
  Fisher, 91–97, 253
  $I \times J$ tables, 97–98, 104
  logistic regression, 251–257
  matched pairs, 412
  ordinal variables, 114
  StatXact and LogXact, 633, 635, 640, 643
  trend in proportions, 98
  unconditional test, 94–96, 104, 114
Expected frequencies, 22, 25,
Exponential dispersion family, 133, 310
Exponential distribution, 313, 388
Exponential family, 116, 133
Extreme-value distribution, 249–250, 264

Fisher, R. A., 22–23, 622–624, 626, 628
  df argument with Pearson, 622–623
  variance test, 163
  Fisher scoring, 145–149, 156, 247, 623, 625
Fisher's exact test, 91–97, 99, 253, 623
  and Bayes approach, 608
  conservativism, 93–94
  controversy, 95–96, 104
  software, 635
  UMPU, 104
  versus unconditional test, 95–96, 104, 114
Fitted values, 121
  asymptotic distribution, 194, 341, 585–586, 593
Freeman–Tukey chi-squared, 112, 594

G^2 statistic, *see* Likelihood-ratio statistic
$G^2(M_0 \mid M_1)$, 187, 363
Gamma, 58–59, 88, 110, 596–597
Gamma distribution, 559–560, 574
Gauss–Hermite quadrature, 521–522, 651
Generalized estimating equations (GEE), 466–475, 481–482, 501, 557–558, 649
Generalized additive models, 153–155, 156, 301, 630, 636
Generalized linear mixed model (GLMM), 417, 492
  Bayesian approach, 524, 609
  binary data, 492–527
  correlation nonnegative, 497, 564
  count data, 563–565
  heterogeneity, interpretation, 497–498
  marginal effects, comparison, 498–502, 535, 563–564
  marginal model, corresponding, 527, 563–564, 574–575
  misspecification, 547–548
  model fitting, 520–526, 527
  multinomial data, 513–516
  software, 649–653
Generalized linear model (GLM), 116–119, 625
  canonical link, 117, 148–149, 193, 257, 472, 496
  covariance matrix, 137–138
  exponential dispersion family, 133
  inference using, 139–143
  likelihood equations, 135–136, 148
  model fitting, 143–149
  moments, 132–134
  multivariate, 274
  variance function, 136
Generalized loglinear model, 332–333, 464, 481, 602
Gini concentration index, 68
Goodman, L. A., 627–629

Goodman and Kruskal tau and lambda, 68–69
Goodness-of-fit statistics
  continuous explanatory variables, 176–177, 197
  deviance for GLMs, 118–119, 139–142
  likelihood-ratio test, 141–142, 186–187, 363–365
  logistic regression, 174–177, 186–187, 208
  loglinear models, 324
  mixture summary, 565
  Pearson chi-squared, 22–26
  uninformative for ungrouped data, 162
Graphical models, 357–360, 398, 629
Grouped versus ungrouped data, 140–141, 162, 174–177, 208, 228
GSK method, 601
Gumbel distribution, 249

Hat matrix, 143, 225, 589
Hazard function, 301, 388, 399–400
Heterogeneity, 130, 235–236, 291, 377, 492–493, 497, 499–500, 507–510, 538
Hierarchical models, 316, 520, 609
History, 619–631
Homogeneity of odds ratios, 54, 183, 234–236, 255, 258
Homogeneous association, 54, 320, 377, 407, 623
Hosmer–Lemeshow statistic, 177, 639
Hypergeometric distribution, 91
  and binomial, 113
  moments, 103, 232
  multiple hypergeometric, 97
  noncentral, 99

Identity link, 117, 120, 124, 128, 385, 387, 562, 565
Incomplete table, 392
Independence
  conditional, *see* Conditional independence
  estimated expected frequencies, 78
  exact test, *see* Fisher's exact test
  from irrelevant alternatives, 299, 302
  joint, 318, 319
  likelihood-ratio test, 79
  loglinear model, 132, 314–315, 336, 352
  mutual, 318–319, 353, 354
  Pearson test, 78–79
  quasi, 426–428, 432–433, 443
  residuals, 81, 111–112
  smoothing using, 85–86
  two-way table, 38–39, 78–79, 111
  variance of proportion estimator, 113
Independent multinomial sampling, 40, 67, 339–340
Influence diagnostics, 224–226, 638
Information matrix, 9
  GLM, 138, 145–146
  logistic regression, 193
  loglinear model, 339
  observed versus expected, 145–146, 247
Interaction, 210
  and odds ratios, 54
  three-factor, 320
  uniform, 407
Isotropy, 406
Item response models, 495
Iterative proportional fitting, 343–345, 347
Iterative reweighted least squares, 147, 156, 195, 343

Joint independence, 318, 319

Kappa, 434–435, 443, 453, 645
Kendall's tau and tau-*b*, 60, 68
Kernel smoothing, 613–615, 616

Lambda (measure of association), 69
Laplace approximation, 523
Latent class models, 538–545, 565, 571–572, 653
Latent variable, 277–278, 399
LD 50, 167
Leverage, 143, 589
Likelihood function, 9
  generalized linear model, 133, 135
  marginal likelihood, 521
Likelihood-ratio statistic, 11–12, 24
  asymptotic chi-squared distribution, 590–591
  and confidence intervals, 13, 16, 17, 78, 638
  difference of deviances, 141–142, 187, 363–364
  independence, 79
  minimized by ML estimate, 590–591
  monotone property, 141
  nested models, 363–365
  noncentrality, 243
  nonnegative, 34, 141
  partitioning, 82–84, 363–365, 399, 405
  Pearson statistic, comparison, 24, 80, 364
  as power divergence statistic, 112
  sparse data, 80, 395–397
Linear-by-linear association, 369–373, 643–644
  and bivariate normal, 370, 399
  and correlation model, 408
  heterogeneous, 377
  homogeneous, 377–379, 407
  score statistic, 406

Linear logit model, 180–182
directed inference, 236–237
efficiency, 197
exact test, 253
likelihood equations, 209
and trend test, 197, 237–239
Linear predictor, 116
Linear probability model, 120–121, 291
and trend test, 181–182
Link function, 116, 135
canonical, 117, 148–149, 193, 257, 472, 496
cumulative, 282–286, 301
goodness of link, 257–258, 301
inverse cdf, 124–125, 163, 282
Litter effects, 151–153, 291, 556–558, 566
Local odds ratio, 55, 312, 369–370
asymptotic covariances, 597
conditional, 321–322, 377
exponential family for multinomial, 310–311
Logistic distribution, 125, 162, 197, 246
Logistic-normal distribution, 265
Logistic-normal model, 496–513, 516–527
Logistic regression, 121–125, 165–196
case-control studies, 170–171, 418–420, 625
categorical predictors, 177–186
conditional, 250–258, 414–420, 495–496, 526, 625, 640, 645
conditional independence, 183–184, 231
covariance matrix, 193–194
design, 196, 609
diagnostics, 219–230, 257–258
existence of ML estimates, 195–196, 394–395
fitting model, 192–196
generalized linear model, 117, 121–125
goodness-of-fit, 174–177, 186–187, 197
inference, 172–177
interpretation, 166–171, 191
likelihood equations, 192–193
linear logit model, *see* Linear logit model
loglinear models, connection, 315, 330–332, 367, 593–594
marginal models, 414, 456–476
matched pairs, 414–420, 493–496
model-building, 211–225
multiple predictors, 182–195
nonparametric mixture, 546–547, 653
normal distribution connection, 171, 207–208
and odds ratio, 124, 166
perfect discrimination, 195–196
probability estimators, 166–167, 191, 194
random effects, 496–513, 516–527
regressive logistic model, 479–481
repeated binary response, 414–420, 456–476, 496–513, 516–527
repeated multinomial response, 461–464, 469, 474–475, 513–516
residuals, 219–223
sample size determination, 242–243
sample size and number of predictors, 212
software, 637–643, 645, 649–651
Logit transform, 75, 117, 624
bias, 196
confidence interval, 109
in logistic regression, 123
standard error, 74–75
Wald test of proportion, 208–209
Loglinear models, 117–118, 314–347, 627–629
covariance matrix, 138–139, 338, 341, 593, 598
existence of estimates, 341, 392–395
fitting, 342–344
four dimensions, 326–330, 355
generalized loglinear model, 332–333, 464, 481
generalized linear model, 117–118, 125–132
goodness of fit, 337–338
homogeneous association, 320, 377
independence, 232, 314–315, 318–319, 336, 352, 365–366
likelihood equations, 334–336
linear-by-linear association, 369–373, 377–379
logit models, connection, 315, 330–332, 367, 593–594
ordinal variables, 367–377
parameter definition, 316–317, 352–353
Poisson-multinomial connection, 317–318, 339–340
probability estimates, 340–341
rates, 385–391
saturated, 316, 380
selection, 360–366
software, 643–644
square-tables, 424–431
three-factor interaction, 320
(X, Y, Z) type symbols, 320–321
Log link, 118, 124, 125, 132, 138, 140, 314, 560, 563
Log-log models, 248–250, 283
Longitudinal studies, *see* Repeated response
Lowess, 154

Mann–Whitney statistic, 90, 301, 452–453
Mantel, N., 625
Mantel–Haenszel estimator, 234–235, 417, 639

Mantel–Haenszel test, *see* Cochran–Mantel–Haenszel test
Mantel score test, 87, 88, 89, 379
Marginal distribution, 37. *See also* Marginal models
Marginal likelihood, 521
Marginal homogeneity
  binary matched pairs, 410–413
  and independence, 111
  nominal tests, 422–423, 457–459
  ordinal tests, 421, 452–453, 458
  multi-way table, 439–442, 456–459, 647–649
Marginal models, 414, 420–423, 439–442, 456–476
  conditional models, comparison, 498–502
  GEE approach, 466–475
  ML fitting, 464–466, 481
  odds ratio, 451, 494
  software, 644–649
Marginal symmetry, 442
Marginal table, 48
  same association as partial table, 358–360, 398
Markov chains, 477–481, 482, 489–490
Matched pairs, 409–454
  Cochran–Mantel–Haenszel approach, 413
  dependent proportions, 410–412
  logistic models, 414–420, 493–496, 516–517
  McNemar test, 411–413, 424, 442, 644–645
  odds ratio estimates, 417, 451, 494
  ordinal data, 420–421, 429–431, 439, 443, 452–453, 462–464, 536
  random effects, 417–418, 493–494, 535
Maximum likelihood, 9
  conditional, 100, 417, 494–496, 526
  inconsistent estimator, 450
  iterative reweighted least squares, 147, 156, 195, 343
  likelihood function, *see* Likelihood function
  versus other methods, 468, 603–605, 612
McNemar test, 411–413, 424, 442, 644–645
Mean response model, 291–294
Measurement error, 347, 493
Measures of association, 43–47, 54–60, 68–69, 620–622
  asymptotic normality, 110
  comparing several values, 599
Mendel, 22–23, 623
Mid-distribution function, 34
Mid-P-value, 20, 27, 33, 104
Midranks, 89, 90, 302
Minimum chi-squared, 112, 611–612, 616, 618, 629
Minimum discrimination information, 112, 612–613, 616
Misclassification error, 347
Missing data, 103, 347, 463, 475–476, 482
Mixture models, 538–566. *See also* Generalized linear mixed models
ML, *see* Maximum likelihood
Model-based inference
  improved precision of estimation, 85, 112, 174, 239–240, 264
  model-based tests, 141–142, 172, 363–365, 396, 399
Model matrix, 135
Monotone trends, 88. *See also* Trend tests
Monte Carlo methods, 114, 522–525, 609, 629–630, 635
Multicollinearity, 212
Multilevel models, 520, 609, 651
Multinomial distribution, 6–7
  binomial factorization, 289
  exponential family, 310–311
  inference, 21–26, 35
  mean, correlation, covariance, 7, 31, 579–580, 596
  and Poisson, 8–9, 40
  sampling models, 40–41, 67
Multinomial logit models, 267–291, 298–300, 302, 624, 640–643, 651–653
Multinomial loglinear model, 317–318, 339–341
Multinomial response models, 267–300, 640–643
Mutual independence, 318–319, 353, 354

National Halothane Study, 627, 629
Natural exponential family, 116, 133, 155
Natural parameter, 133
Negative binomial
  distribution, 31, 161, 163, 560, 566, 574
  regression model, 131, 560–563, 565, 566, 653
Nested models
  likelihood-ratio comparison, 141–142, 187, 363–364
  simultaneous tests, 263
  using X^2, 364
Newton–Raphson, 143–146, 163–164
  and Fisher scoring, 145, 247
  IPF, comparison, 344–345
  logistic regression, 194–195
  loglinear models, 342–345
Neyman, J., 626
Nominal variable, 2–3
  baseline-category logit models, 267–274, 300, 310–311, 426, 515, 640–643

Nominal variable (*Continued*)
matched pairs, 422–423
measures of association, 55–57, 68–69
square table models, 425–433, 439–442
Noncentral chi-squared distribution, 237, 258
asymptotic representation, 591–592, 595
noncentrality parameter, 237, 243–245, 408, 597
power and df, 237–239
Nonparametric random effects, 545–553, 565–566, 653
Normal distribution
asymptotic normality, *see* Delta method
and chi-squared, 82
and logistic regression, 171, 207–208
underlying categorical data, 112, 264, 370, 620

O, o rates of convergence, 577, 595
Observational study, 43
Odds, 44
Odds ratio, 44, 620
bias, 70, 595
case-control studies, 46–47
conditional, 51–54, 255, 321, 417, 451
conditional ML estimate, 255, 417
confidence interval, 71, 77–78, 99–102, 255, 256
cumulative, 67
exact inference, 99–101, 253, 255
homogeneity, in $2 \times 2 \times K$ tables, 54, 183, 234–236, 255
$I \times J$ tables, 55–56, 581, 597
invariance properties, 45–46, 59
local, *see* Local odds ratio
Mantel–Haenszel estimator, 234–235
marginal, 451, 494
matched pairs, 415–418, 451
logistic regression parameters, 124, 166, 171, 179, 183, 331, 415, 497–500
loglinear model parameters, 315, 316, 321, 331, 369
ordinal variables, *see* Local odds ratio
relation to relative risk, 47, 124, 624
standard error, 71, 75–77, 581, 597
Offset, 385
Ordinal variables, 2–3
cumulative link models, 282–286
cumulative logit models, 274–282, 301, 420–421
efficiency, 197, 301
exact tests, 98, 253
improved power, 88–90, 236–239, 373
loglinear models, 367–377, 399
marginal models, 420–421, 429–430, 440–441, 462–464
matched pairs, 420–421, 429–431, 439, 443, 452–454, 462–464
mean response model, 291–294
measures of association, 57–59, 67, 68
multinomial response models, 274–295
ordinal quasi symmetry, 429–430, 440–441, 647
repeated response, 461–464, 469, 474–475, 514–515, 517–520
scores, choice of, 88–90, 383–384
testing independence, 86–91, 373
Overdispersion, 493
binomial, 8, 30, 151–153, 291, 555–558, 573, 653
litter effects, 151–153, 291, 556–558, 566
Poisson, 7–8, 130–131, 636
quasi-likelihood, 151–153, 291, 555–558, 653

Paired comparisons, *see* Bradley–Terry model
Parallel odds models, 374–375
Partial tables, 48
Partitioning
chi-squared statistic, 82–84, 112–113, 365, 399, 405
and combining rows, 112
$I \times J$ tables, 82–83
nested models, 365
trend test, 181, 373
Pattern mixture model, 476
Pearson, Karl, 619–623, 628
arguments with Fisher, Yule, 79, 619–623
goodness of fit, 22–24, 79
Pearson chi-squared statistic, 22–26, 79, 111–112
asymptotic chi-squared distribution, 589–590
asymptotic conditional distribution, 103
continuity correction, 103
degrees of freedom, 25, 79, 622
and z for difference of proportions, 111
goodness of fit, 22–26
independence, 78–79, 111–112, 622
and likelihood-ratio, comparison, 24, 80, 364
minimizing, 112, 611–612, 616, 618, 629
moments, 103
multinomial parameters, 22–26
nested models, 364
noncentral chi-squared distribution, *see* Noncentral chi-squared distribution
score statistic, 24
sparse data, 80, 395–397
with ungrouped data, 162
upper bound, 112

Pearson residual, 81, 142, 588–589, 593
  binomial GLM, 220, 555, 638
  Poisson GLM, 142, 366, 588
Penalized likelihood, 614–615
Penalized quasi likelihood (PQL), 523–524
Perfect contingency tables, 398
Perfect discrimination, 195–196
Phi-squared, 112
Poisson distribution, 7
  comparing means, 31
  exponential family, 117, 134
  moments, 7, 31
  and multinomial, 8–9, 40,
  and negative binomial, 131, 559–560, 566, 574
  overdispersion, 7–8, 130–131, 636
  Poisson sampling, 39
  variance test, 163
Poisson models
  counts, 125–132, 155, 563–565
  deviance, 140
  loglinear model, 117–118, 125–132, 138–139, 232, 314–347
  overdispersion, 130–131, 150–151, 636
  random effects, 563–565
  rates, 385–391, 399–400
Polytomous logit models, 267–291
Population-averaged effects, 414, 495, 499–501
Positive likelihood-ratio dependence, 406
Power
  calculating, 240–245, 640
  increased, for directed alternatives, 88–90, 236–239, 373
  and noncentrality, 237–239, 243–245
  and number of ordinal categories, 301
Power-divergence statistic, 112, 613
Prediction, 525–526
Probit model, 124–125, 246–247, 258, 623, 640
  discrete choice, 302
  likelihood equations, 265
  normal parameters, 163, 246, 264
  ordinal data, 278, 283, 301, 312, 641
  random effects, 535
  threshold and utility motivations, 264
Profile likelihood confidence interval, 78, 512, 638
Propensity score, 196
Proportional hazards model, 283–284, 301, 389, 643
Proportional odds, *see* Cumulative logit models
Proportional reduction in variation, 56–57, 67–68
Proportions
  admissible estimator, 605
  asymptotic distribution, 585–588, 593
  Bayesian inference, 605–607
  confidence interval, 15–17, 32–33, 635
  dependent, 410–412
  difference, *see* Difference of proportions
  ratio, *see* Relative risk
  standard error, 11, 340–341
P-value
  mid-P-value, 20, 27, 33, 104
  randomized, 27, 32
  UMVU estimator, 162

Qualitative variable, 3–4
Quantitative variable, 3–4
Quasi-association, 431, 453–454
Quasi-independence, 426–428, 432–433, 443
Quasi-likelihood
  binary models, 151–153, 291, 555–558
  count models, 150–151
  GLM, 149–153, 156
  multivariate (GEE), 466–475, 481–482, 625
  overdispersion, 150–153, 291, 555–558
Quasi-symmetry, 425–431, 433–434, 451, 454, 646–647
  and Bradley–Terry model, 438–439
  and marginal homogeneity, 428–430
  multiway tables, 440–441
  and Rasch model, 552–553, 565

Raking a table, 345–346, 347, 643
Random component of GLM, 116, 133
Random effects, 417, 492–527
Random intercept, 493
Ranks, 89, 90, 298, 301, 302
Rasch mixture model, 548–551, 653
Rasch model, 495–496, 517, 526, 535, 565, 624
Rates, 385–391, 399–400
RC model, 379–381, 399–400
Regressive logistic model, 479–481
Relative risk, 43–44
  asymptotic standard error, 73
  collapsibility, 398
  confidence interval, 73, 77
  homogeneity, 258
  in model, 124
  and odds ratio, 47, 624
Repeated response, 409–517. *See also* Generalized linear mixed models; Marginal models; Matched pairs
Residuals, 142–143, 156
  asymptotic distribution, 587–589
  binomial GLMs, 219–223
  deviance, *see* Deviance residual
  Pearson, *see* Pearson residual
  Poisson GLMs, 143, 366–367

standardized Pearson, *see* Standardized Pearson residual
Retrospective study, 42–43. *See also* Case-control study
  logistic regression, 170–171
  odds ratio, 46–47
Ridits, 111, 406
ROC curve, 228–230, 258
Row and column effects model, *see* RC model
Row effects model, 374–376, 643–644
R-squared type measure
  logistic regression, 226–228, 258
  nominal association, 56–57, 67–68

Sample size determination, 240–245
Sampling methods, 39–43
Sampling zero, 392
Sandwich estimator, 471–474
SAS, 632–643
Saturated model, 119, 139, 382
  logit models, 178,
  loglinear models, 316, 380
Scaled deviance, 140
Scores
  choice of, 88–90, 383–384
  efficiency, 197, 301
  in loglinear models, 369–379, 407
  in trend test, 88–89, 181–182, 406
Score statistic, 12, 26–27
  confidence intervals, 15–16, 77
  logistic regression, 232, 297–298
  Pearson statistic, 24
  and standardized residuals, 156
  trend test, 182
Selection model, 475–476
Sensitivity, 38, 60, 228–230
Simpson diversity index, 596
Simpson's paradox, 51, 59–60, 224, 354, 621
Small-area estimation, 502–504
Small samples
  adding constants to cells, 397–398
  alternative asymptotics, 233, 396–397
  exact inference, 18–20, 91–101, 104, 251–257
  existence of estimates, 195–196, 341, 392–395
  model-based tests, 187, 251–257
  X^2 and G^2, 24, 80, 364, 395–397
  zeros, 392–398
Smoothing
  Bayes, 606–610
  generalized additive model, 153–155
  improved estimation with model, 85, 112, 174, 239–240, 264
  kernel, 613–615, 616
  penalized likelihood, 614–615
Software, 632–653
  SAS, 632–643
  StatXact and LogXact, 633, 635, 640, 643
Somers' d, 68
Sparse data, 391–398, 187, 250–257, 591
  asymptotics, 233, 396–397
Spearman's rho, 90
Specificity, 38, 60, 228–230
Square tables, 409–454
Standardized table, 345–346
Standardized parameter estimate, 191–192, 197
Standardized Pearson residual, 81, 143, 589
  binomial GLMs, 220, 638
  and Pearson statistic, 112
  Poisson GLMs, 143, 367, 634
  as score statistic, 156
StatXact, 633, 635, 640, 643
Stepwise model-building, 213–216
Stochastic ordering, 33, 67, 301
Structural zero, 25, 392
Subject-specific effects, 414–420, 491, 498–500
Sufficient statistics, 148, 250–257, 273, 334, 336
Suppressor variable, 67
Survival data, 385–391
Symmetric association, 425
Symmetry, 424–425, 644–647
  complete, 440
  multiway, 439–442
Systematic component of GLM, 116

Tetrachoric correlation, 620
Three-factor interaction, 320
Threshold model, 264, 277–279
Tolerance distribution, 245–246
Transformations, 595, 596
Transition probabilities, 477, 490
Transitional model, 464, 476–481, 482
Tree-structured methods, 257, 631
Trend tests, 86–90, 103, 296, 373, 379
  Cochran–Armitage for proportions, 90, 181–182, 237–239
  efficiency, 197, 301
  exact, 253
  software, 634, 635

Uncertainty coefficient, 57
Uniform association model, 312, 369–370, 377
Uniform interaction model, 407
Uniqueness of ML estimate, 341
Utility, 264

Variance
  asymptotic, *see* Delta method
  components, 492, 525
  in exponential family, 134
  stabilizing, 596, 626
  test for Poisson, 163
  variance function, 136, 149–150

Wald statistic, 11, 27
  and power, 172, 208–209
Wald confidence intervals, 13
  adjusted intervals, 33, 102
Weight matrix, 138, 155, 164
Weighted kappa, 435, 443, 645
Weighted observation, 391
Weighted least squares, 481, 600–604, 615, 629
  and minimum modified chi-squared, 611, 612
  and ML estimation, 146–148, 603–604
Wilcoxon test, 90, 301
WLS, *see* Weighted least squares

X^2 statistic, *see* Pearson chi-squared statistic
$X^2(M_0|M_1)$, 364

Yates continuity correction, 103
Yule, G. U., 620–621, 628
Yule's Q, 68, 110

Zero cell count
  adding constants, 70–71, 397–398
  effects on estimates, 70–71, 78, 256
  sampling, 392
  structural, 25, 392

Categorical Data Analysis, Second Edition. Alan Agresti

ISBN: 0-471-36093-7

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Peter Bloomfield, Noel A. C. Cressie, Nicholas I. Fisher, Iain M. Johnstone, J. B. Kadane, Louise M. Ryan, David W. Scott, Adrian F. M. Smith, Jozef L. Teugels*
Editors Emeriti: *Vic Barnett, J. Stuart Hunter, David G. Kendall*

A complete list of the titles in this series appears at the end of this volume.

WILEY SERIES IN PROBABILITY AND STATISTICS

ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

The ***Wiley Series in Probability and Statistics*** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data
AGRESTI · An Introduction to Categorical Data Analysis
AGRESTI · Categorical Data Analysis, *Second Edition*
ANDĚL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, *Second Edition*
*ANDERSON · The Statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG · Statistical Methods for Comparative Studies
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and DAVID (editors) · Advances in Biometry
ARNOLD, BALAKRISHNAN, and NAGARAJA · Records
*ARTHANARI and DODGE · Mathematical Programming in Statistics
*BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BALAKRISHNAN and KOUTRAS · Runs and Scans with Applications
BARNETT · Comparative Statistical Inference, *Third Edition*
BARNETT and LEWIS · Outliers in Statistical Data, *Third Edition*
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference
BASILEVSKY · Statistical Factor Analysis and Related Methods: Theory and Applications
BASU and RIGDON · Statistical Methods for the Reliability of Repairable Systems
BATES and WATTS · Nonlinear Regression Analysis and Its Applications
BECHHOFER, SANTNER, and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons
BELSLEY · Conditioning Diagnostics: Collinearity and Weak Data in Regression
BELSLEY, KUH, and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, *Third Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

BERRY, CHALONER, and GEWEKE · Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner
BERNARDO and SMITH · Bayesian Theory
BHAT and MILLER · Elements of Applied Stochastic Processes, *Third Edition*
BHATTACHARYA and JOHNSON · Statistical Concepts and Methods
BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications
BILLINGSLEY · Convergence of Probability Measures, *Second Edition*
BILLINGSLEY · Probability and Measure, *Third Edition*
BIRKES and DODGE · Alternative Methods of Regression
BLISCHKE AND MURTHY (editors) · Case Studies in Reliability and Maintenance
BLISCHKE AND MURTHY · Reliability: Modeling, Prediction, and Optimization
BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, *Second Edition*
BOLLEN · Structural Equations with Latent Variables
BOROVKOV · Ergodicity and Stability of Stochastic Processes
BOULEAU · Numerical Methods for Stochastic Processes
BOX · Bayesian Inference in Statistical Analysis
BOX · R. A. Fisher, the Life of a Scientist
BOX and DRAPER · Empirical Model-Building and Response Surfaces
*BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement
BOX, HUNTER, and HUNTER · Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building
BOX and LUCEÑO · Statistical Control by Monitoring and Feedback Adjustment
BRANDIMARTE · Numerical Methods in Finance: A MATLAB-Based Introduction
BROWN and HOLLANDER · Statistics: A Biomedical Introduction
BRUNNER, DOMHOF, and LANGER · Nonparametric Analysis of Longitudinal Data in Factorial Experiments
BUCKLEW · Large Deviation Techniques in Decision, Simulation, and Estimation
CAIROLI and DALANG · Sequential Stochastic Optimization
CHAN · Time Series: Applications to Finance
CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression
CHATTERJEE and PRICE · Regression Analysis by Example, *Third Edition*
CHERNICK · Bootstrap Methods: A Practitioner's Guide
CHERNICK and FRIIS · Introductory Biostatistics for the Health Sciences
CHILÈS and DELFINER · Geostatistics: Modeling Spatial Uncertainty
CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies
CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, *Second Edition*
*COCHRAN and COX · Experimental Designs, *Second Edition*
CONGDON · Bayesian Statistical Modelling
CONOVER · Practical Nonparametric Statistics, *Second Edition*
COOK · Regression Graphics
COOK and WEISBERG · Applied Regression Including Computing and Graphics
COOK and WEISBERG · An Introduction to Regression Graphics
CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Third Edition*
COVER and THOMAS · Elements of Information Theory
COX · A Handbook of Introductory Statistical Methods
*COX · Planning of Experiments
CRESSIE · Statistics for Spatial Data, *Revised Edition*
CSÖRGŐ and HORVÁTH · Limit Theorems in Change Point Analysis
DANIEL · Applications of Statistics to Industrial Experimentation
DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, *Sixth Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

*DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, *Second Edition*
DAVID · Order Statistics, *Second Edition*
*DEGROOT, FIENBERG, and KADANE · Statistics and the Law
DEL CASTILLO · Statistical Process Adjustment for Quality Control
DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis
DEY and MUKERJEE · Fractional Factorial Plans
DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
DODGE · Alternative Methods of Regression
*DODGE and ROMIG · Sampling Inspection Tables, *Second Edition*
*DOOB · Stochastic Processes
DOWDY and WEARDEN · Statistics for Research, *Second Edition*
DRAPER and SMITH · Applied Regression Analysis, *Third Edition*
DRYDEN and MARDIA · Statistical Shape Analysis
DUDEWICZ and MISHRA · Modern Mathematical Statistics
DUNN and CLARK · Applied Statistics: Analysis of Variance and Regression, *Second Edition*
DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, *Third Edition*
DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations
*ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
ETHIER and KURTZ · Markov Processes: Characterization and Convergence
EVANS, HASTINGS, and PEACOCK · Statistical Distributions, *Third Edition*
FELLER · An Introduction to Probability Theory and Its Applications, Volume I, *Third Edition,* Revised; Volume II, *Second Edition*
FISHER and VAN BELLE · Biostatistics: A Methodology for the Health Sciences
*FLEISS · The Design and Analysis of Clinical Experiments
FLEISS · Statistical Methods for Rates and Proportions, *Second Edition*
FLEMING and HARRINGTON · Counting Processes and Survival Analysis
FULLER · Introduction to Statistical Time Series, *Second Edition*
FULLER · Measurement Error Models
GALLANT · Nonlinear Statistical Models
GHOSH, MUKHOPADHYAY, and SEN · Sequential Estimation
GIFI · Nonlinear Multivariate Analysis
GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems
GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, *Second Edition*
GOLDSTEIN and LEWIS · Assessment: Problems, Development, and Statistical Issues
GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing
GROSS and HARRIS · Fundamentals of Queueing Theory, *Third Edition*
*HAHN and SHAPIRO · Statistical Models in Engineering
HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners
HALD · A History of Probability and Statistics and their Applications Before 1750
HALD · A History of Mathematical Statistics from 1750 to 1930
HAMPEL · Robust Statistics: The Approach Based on Influence Functions
HANNAN and DEISTLER · The Statistical Theory of Linear Systems
HEIBERGER · Computation for the Analysis of Designed Experiments
HEDAYAT and SINHA · Design and Inference in Finite Population Sampling
HELLER · MACSYMA for Statisticians
HINKELMAN and KEMPTHORNE: · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design
HOAGLIN, MOSTELLER, and TUKEY · Exploratory Approach to Analysis of Variance

*Now available in a lower priced paperback edition in the Wiley Classics Library.

HOAGLIN, MOSTELLER, and TUKEY · Exploring Data Tables, Trends and Shapes
*HOAGLIN, MOSTELLER, and TUKEY · Understanding Robust and Exploratory Data Analysis
HOCHBERG and TAMHANE · Multiple Comparison Procedures
HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variance, *Second Edition*
HOEL · Introduction to Mathematical Statistics, *Fifth Edition*
HOGG and KLUGMAN · Loss Distributions
HOLLANDER and WOLFE · Nonparametric Statistical Methods, *Second Edition*
HOSMER and LEMESHOW · Applied Logistic Regression, *Second Edition*
HOSMER and LEMESHOW · Applied Survival Analysis: Regression Modeling of Time to Event Data
HØYLAND and RAUSAND · System Reliability Theory: Models and Statistical Methods
HUBER · Robust Statistics
HUBERTY · Applied Discriminant Analysis
HUNT and KENNEDY · Financial Derivatives in Theory and Practice
HUSKOVA, BERAN, and DUPAC · Collected Works of Jaroslav Hajek—with Commentary
IMAN and CONOVER · A Modern Approach to Statistics
JACKSON · A User's Guide to Principle Components
JOHN · Statistical Methods in Engineering and Quality Assurance
JOHNSON · Multivariate Statistical Simulation
JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz
JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE · The Theory and Practice of Econometrics, *Second Edition*
JOHNSON and KOTZ · Distributions in Statistics
JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present
JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 1, *Second Edition*
JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 2, *Second Edition*
JOHNSON, KOTZ, and BALAKRISHNAN · Discrete Multivariate Distributions
JOHNSON, KOTZ, and KEMP · Univariate Discrete Distributions, *Second Edition*
JUREČKOVÁ and SEN · Robust Statistical Procedures: Aymptotics and Interrelations
JUREK and MASON · Operator-Limit Distributions in Probability Theory
KADANE · Bayesian Methods and Ethics in a Clinical Trial Design
KADANE AND SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence
KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, *Second Edition*
KASS and VOS · Geometrical Foundations of Asymptotic Inference
KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis
KEDEM and FOKIANOS · Regression Models for Time Series Analysis
KENDALL, BARDEN, CARNE, and LE · Shape and Shape Theory
KHURI · Advanced Calculus with Applications in Statistics, *Second Edition*
KHURI, MATHEW, and SINHA · Statistical Tests for Mixed Linear Models
KLUGMAN, PANJER, and WILLMOT · Loss Models: From Data to Decisions
KLUGMAN, PANJER, and WILLMOT · Solutions Manual to Accompany Loss Models: From Data to Decisions
KOTZ, BALAKRISHNAN, and JOHNSON · Continuous Multivariate Distributions, Volume 1, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Supplement Volume
KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 1
KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 2
KOVALENKO, KUZNETZOV, and PEGG · Mathematical Theory of Reliability of Time-Dependent Systems with Practical Applications
LACHIN · Biostatistical Methods: The Assessment of Relative Risks
LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction
LAMPERTI · Probability: A Survey of the Mathematical Theory, *Second Edition*
LANGE, RYAN, BILLARD, BRILLINGER, CONQUEST, and GREENHOUSE · Case Studies in Biometry
LARSON · Introduction to Probability Theory and Statistical Inference, *Third Edition*
LAWLESS · Statistical Models and Methods for Lifetime Data, *Second Edition*
LAWSON · Statistical Methods in Spatial Epidemiology
LE · Applied Categorical Data Analysis
LE · Applied Survival Analysis
LEE and WANG · Statistical Methods for Survival Data Analysis, *Third Edition*
LEPAGE and BILLARD · Exploring the Limits of Bootstrap
LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics
LIAO · Statistical Group Comparison
LINDVALL · Lectures on the Coupling Method
LINHART and ZUCCHINI · Model Selection
LITTLE and RUBIN · Statistical Analysis with Missing Data, *Second Edition*
LLOYD · The Statistical Analysis of Categorical Data
MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, *Revised Edition*
MALLER and ZHOU · Survival Analysis with Long Term Survivors
MALLOWS · Design, Data, and Analysis by Some Friends of Cuthbert Daniel
MANN, SCHAFER, and SINGPURWALLA · Methods for Statistical Analysis of Reliability and Life Data
MANTON, WOODBURY, and TOLLEY · Statistical Applications Using Fuzzy Sets
MARDIA and JUPP · Directional Statistics
MASON, GUNST, and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science, *Second Edition*
McCULLOCH and SEARLE · Generalized, Linear, and Mixed Models
McFADDEN · Management of Data in Clinical Trials
McLACHLAN · Discriminant Analysis and Statistical Pattern Recognition
McLACHLAN and KRISHNAN · The EM Algorithm and Extensions
McLACHLAN and PEEL · Finite Mixture Models
McNEIL · Epidemiological Research Methods
MEEKER and ESCOBAR · Statistical Methods for Reliability Data
MEERSCHAERT and SCHEFFLER · Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice
*MILLER · Survival Analysis, *Second Edition*
MONTGOMERY, PECK, and VINING · Introduction to Linear Regression Analysis, *Third Edition*
MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness
MUIRHEAD · Aspects of Multivariate Statistical Theory

*Now available in a lower priced paperback edition in the Wiley Classics Library.

MURRAY · X-STAT 2.0 Statistical Experimentation, Design Data Analysis, and Nonlinear Optimization
MYERS and MONTGOMERY · Response Surface Methodology: Process and Product Optimization Using Designed Experiments, *Second Edition*
MYERS, MONTGOMERY, and VINING · Generalized Linear Models. With Applications in Engineering and the Sciences
NELSON · Accelerated Testing, Statistical Models, Test Plans, and Data Analyses
NELSON · Applied Life Data Analysis
NEWMAN · Biostatistical Methods in Epidemiology
OCHI · Applied Probability and Stochastic Processes in Engineering and Physical Sciences
OKABE, BOOTS, SUGIHARA, and CHIU · Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, *Second Edition*
OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis
PANKRATZ · Forecasting with Dynamic Regression Models
PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
*PARZEN · Modern Probability Theory and Its Applications
PEÑA, TIAO, and TSAY · A Course in Time Series Analysis
PIANTADOSI · Clinical Trials: A Methodologic Perspective
PORT · Theoretical Probability for Applications
POURAHMADI · Foundations of Time Series Analysis and Prediction Theory
PRESS · Bayesian Statistics: Principles, Models, and Applications
PRESS · Subjective and Objective Bayesian Statistics, *Second Edition*
PRESS and TANUR · The Subjectivity of Scientists and the Bayesian Approach
PUKELSHEIM · Optimal Experimental Design
PURI, VILAPLANA, and WERTZ · New Perspectives in Theoretical and Applied Statistics
PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming
*RAO · Linear Statistical Inference and Its Applications, *Second Edition*
RENCHER · Linear Models in Statistics
RENCHER · Methods of Multivariate Analysis, *Second Edition*
RENCHER · Multivariate Statistical Inference with Applications
RIPLEY · Spatial Statistics
RIPLEY · Stochastic Simulation
ROBINSON · Practical Strategies for Experimenting
ROHATGI and SALEH · An Introduction to Probability and Statistics, *Second Edition*
ROLSKI, SCHMIDLI, SCHMIDT, and TEUGELS · Stochastic Processes for Insurance and Finance
ROSENBERGER and LACHIN · Randomization in Clinical Trials: Theory and Practice
ROSS · Introduction to Probability and Statistics for Engineers and Scientists
ROUSSEEUW and LEROY · Robust Regression and Outlier Detection
RUBIN · Multiple Imputation for Nonresponse in Surveys
RUBINSTEIN · Simulation and the Monte Carlo Method
RUBINSTEIN and MELAMED · Modern Simulation and Modeling
RYAN · Modern Regression Methods
RYAN · Statistical Methods for Quality Improvement, *Second Edition*
SALTELLI, CHAN, and SCOTT (editors) · Sensitivity Analysis
*SCHEFFE · The Analysis of Variance
SCHIMEK · Smoothing and Regression: Approaches, Computation, and Application
SCHOTT · Matrix Analysis for Statistics
SCHUSS · Theory and Applications of Stochastic Differential Equations
SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization
*SEARLE · Linear Models
SEARLE · Linear Models for Unbalanced Data
SEARLE · Matrix Algebra Useful for Statistics

*Now available in a lower priced paperback edition in the Wiley Classics Library.

SEARLE, CASELLA, and McCULLOCH · Variance Components
SEARLE and WILLETT · Matrix Algebra for Applied Economics
SEBER and LEE · Linear Regression Analysis, *Second Edition*
SEBER · Multivariate Observations
SEBER and WILD · Nonlinear Regression
SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems
*SERFLING · Approximation Theorems of Mathematical Statistics
SHAFER and VOVK · Probability and Finance: It's Only a Game!
SMALL and McLEISH · Hilbert Space Methods in Probability and Statistical Inference
SRIVASTAVA · Methods of Multivariate Statistics
STAPLETON · Linear Statistical Models
STAUDTE and SHEATHER · Robust Estimation and Testing
STOYAN, KENDALL, and MECKE · Stochastic Geometry and Its Applications, *Second Edition*
STOYAN and STOYAN · Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics
STYAN · The Collected Papers of T. W. Anderson: 1943–1985
SUTTON, ABRAMS, JONES, SHELDON, and SONG · Methods for Meta-Analysis in Medical Research
TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
THOMPSON · Empirical Model Building
THOMPSON · Sampling, *Second Edition*
THOMPSON · Simulation: A Modeler's Approach
THOMPSON and SEBER · Adaptive Sampling
THOMPSON, WILLIAMS, and FINDLAY · Models for Investors in Real World Markets
TIAO, BISGAARD, HILL, PEÑA, and STIGLER (editors) · Box on Quality and Discovery: with Design, Control, and Robustness
TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
TSAY · Analysis of Financial Time Series
UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
VAN BELLE · Statistical Rules of Thumb
VIDAKOVIC · Statistical Modeling by Wavelets
WEISBERG · Applied Linear Regression, *Second Edition*
WELSH · Aspects of Statistical Inference
WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment
WHITTAKER · Graphical Models in Applied Multivariate Statistics
WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting
WONNACOTT and WONNACOTT · Econometrics, *Second Edition*
WOODING · Planning Pharmaceutical Clinical Trials: Basic Statistical Principles
WOOLSON and CLARKE · Statistical Methods for the Analysis of Biomedical Data, *Second Edition*
WU and HAMADA · Experiments: Planning, Analysis, and Parameter Design Optimization
YANG · The Construction Theory of Denumerable Markov Processes
*ZELLNER · An Introduction to Bayesian Inference in Econometrics
ZHOU, OBUCHOWSKI, and McCLISH · Statistical Methods in Diagnostic Medicine

*Now available in a lower priced paperback edition in the Wiley Classics Library.