Statistics for Industry and Technology

Maria Kateri

Contingency Table Analysis

Methods and Implementation Using R

Birkhäuser

Statistics for Industry and Technology

Series Editor

Editorial Advisory Board

For further volumes:
http://www.springer.com/series/4982

Maria Kateri

Contingency Table Analysis

Methods and Implementation Using R

Birkhäuser

Maria Kateri
Institute of Statistics
RWTH Aachen University
Aachen, Germany

ISBN 978-0-8176-4810-7 ISBN 978-0-8176-4811-4 (eBook)
DOI 10.1007/978-0-8176-4811-4
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2014934697

Mathematics Subject Classification (2010): 62H17, 62J12, 62H25, 62H12, 62F10, 62H15, 62F03

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.birkhauser-science.com)

To my parents Athina and Dimitris

and

to my daughter Zenia

Preface

The focus of this book is on models for contingency table analysis. It deals mainly with log-linear models and special models for ordinal data (log-linear or nonlinear). Special models for ordinal data are dealt with to a greater extent, covered in Chaps. 6, 7, and 9. These models, as for example, association models for two- and multi-way contingency tables or symmetry models for square tables, though very powerful and of great interpreting value, are not very popular in use. This is mainly because they are not readily provided as model options in standard statistical software. Though they can be fitted in R by some special packages, their application usually requires some expertise or experience on these models and R. Existing books on contingency tables or categorical data analysis either do not include association models in their contents or refer only to the simplest types of them. The few exceptions, although treating these models in more detail, remain in the methodological part without providing guidance for applying them in practice. Thus, such models can be used mainly by experts on the topic. This ascertainment was the motivation and the idea behind the concept of this book: to provide a reference that exploits these models, explains their features and interpretation aspects in detail, and simultaneously trains the reader to fit them in practice. Additionally, special issues not covered in other books, such as the adjustment of models to account for structural zeros, are addressed. The goal was to end up with a methodological book that makes all approaches, models, or graphs presented, easy reproducible.

The aim to familiarize readers with methods and models for the analysis of contingency tables and their use in practice is served by discussing the models' features and interconnections and by giving special emphasis on the examples' analysis, their interpretation, and their implementation in R. When needed, special R functions are provided (in the web companion of the book) that automatize the required procedures, simplifying thus their applicability. For example, functions are provided for deriving the midrank scores of a classification variable or computing the local odds ratios of a two-way contingency table (or other types of generalized odds ratios). Hence, all models, measures, and graphs discussed can easily be realized in R by handy functions. The examples are worked out in R, explaining the use of the functions, so that the reader is gradually trained and at the end in the

position to alter the functions and adjust them to special needs. The web companion of the book is to be found under

http://cta.isw.rwth-aachen.de

Framing this book on model-based analysis, the great body of nonparametric methods (especially for ordinal data) and smoothing methods for contingency tables is not considered. Emphasis is given primarily on models that treat the variables symmetrically rather than distinguishing between response and explanatory variables. Additionally, since the book deals only with contingency tables, regression-type models that involve also continuous explanatory variables are not addressed at all. Thus, logistic regression, though very important in categorical data analysis, is partially covered, only for the case of categorical explanatory variables (*logit models*). Furthermore, clustered categorical data and multivariate response models are not considered. Only bivariate response models are considered for data represented in square two-way contingency tables with commensurable classification variables, nominal or ordinal. For more than two occasions, we refer to other special reference sources.

The approach adopted is the asymptotic frequentist approach. A short reference on Bayesian analysis of contingency tables and on small sample inference is provided in the last chapter.

The readership target groups are (a) graduate students or researchers (in statistics or in psychometric, social, biomedical, and pedagogical sciences) and (b) practitioners (e.g., for social or consumers' surveys). The first five chapters (up to Sect. 5.4) address both groups, though group (a) could go quickly through it, for filling gaps and building up gradually the `R` part. Expertise in `R` is not required. Regarding the following material (from Sect. 5.5), group (b) would probably be more interested in the simpler models (as linear-by-linear, row, or column effect association models), easy to present and interpret, while group (a) also in more advanced topics (such as handling structural zeros, the multiplicative row–column association model, models for the marginal distributions, or the generalized odds ratios). An updated rich literature review on a bright aspect of topics is provided at the end of each chapter and is mainly addressed to group (a).

The scope is to simultaneously develop the theoretical and `R` programming skills required for analyzing contingency tables, as well as evaluating and interpreting the results. Parts of the manuscript are based on my notes for the graduate course on categorical data analysis, which I held for about 10 years in the Department of Statistics and Insurance Science of the University of Piraeus in Greece.

I would like to thank those who have been involved in this book project. Special thanks to Alan Agresti who influenced the most my view on categorical data. His *Categorical Data Analysis* book, in its first 1990 edition, is partly responsible for my involvement with categorical data. I also thank him for providing valuable comments on the manuscript, suggesting alterations and additions. I thank Panayiotis Bobotas for proofreading the complete draft, Anna Gottard for her critical comments and suggestions on graphical models, and Anestis Touloumis for his helpful comments on Chaps. 5–10. I appreciate all those who accompanied my scientific journey, in particular my former supervisor Takis Papaioannou.

Thanks also to Narayanaswamy Balakrishnan, editor in chief for the series “Statistics for Industry and Technology,” for his interest in this project, and to Mitch Moulton, editorial assistant at Springer, for his help facilitating it. This book would not have been possible without the generous support of my parents, Athina and Dimitris Kateris, to whom I express my deep gratitude. Finally, I thank my daughter Zenia for her patience during the preparation period and her interest in the development of the project.

Aachen, Germany Maria Kateri
January 2014

Contents

Acronyms

AIC	Akaike's information criterion
ANOAS	Analysis of association
ANOVA	Analysis of variance
BIC	Bayesian information criterion
BT	Bradley–Terry
C	Column effect association model
CA	Correspondence analysis
CI	Confidence interval
GLLM	Generalized log-linear model
GLM	Generalized linear model
LL	Linear-by-linear association model
LR	Likelihood ratio
MCA	Multiple correspondence analysis
MDI	Minimum discrimination information
MH	Marginal homogeneity
ML	Maximum likelihood
MLE	Maximum likelihood estimator
QS	Quasi-symmetry model
R	Row effect association model
RC	Row–column association model
RC(M)	Row–column association model of order M
S	Symmetry model
s.e.	Standard error
T	Conditional (or triangular) symmetry model
U	Uniform association model
WLS	Weighted least squares

Chapter 1
Introduction

Abstract Preliminary material on scales, distributions, and inferential procedures for categorical data is briefly presented. Classes of models, usually applied for categorical data analysis, are introduced and discussed. Finally, the outline of the book is presented.

Keywords Measurement scales • Discrete distributions: binomial, multinomial, poisson, hypergeometric • Asymptotic inference with categorical data

1.1 Categorical Data

Categorical data play an important role in many fields, from biomedicine and social sciences to political sciences, marketing, and quality control. A categorical variable consists of a set of non-overlapping categories and thus categorical data are *counts*, namely the frequencies of occurrence of each category of the variable. When all the involved variables in a problem of interest are categorical, then they are represented in form of a contingency table. This book deals with methods of analysis of contingency tables.

1.1.1 Measurement Scales

The simplest categorical variable is one that has just two categories, usually labeled as "yes-no" or "success-failure". Such a variable is called *binary*. Categorical variables of more than two categories are distinguished, accordingly to their measurement scale, to *nominal* and *ordinal*. Categorical variables, such as nationality and denomination, the categories of which cannot be ordered in any aspect, are nominal. The categories of an ordinal variable exhibit a natural ordering. Characteristic examples of ordinal variables are the social class, the education level,

M. Kateri, *Contingency Table Analysis: Methods and Implementation Using R*,
Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4811-4_1,

or any scale of opinion measurement with categories expanding from "strongly disagree" to "strongly agree" and the middle category being the "neutral." The distances between successive categories of an ordinal variable are not known. *Interval* variables are categorical variables with ordered categories of known in-between distances. These are variables of continuous nature that are categorized, with their categories corresponding to disjoint intervals of values. Variables of this type are, for example, the annual income in euro (from "≤ 5000" to "> 150000") or age (from "less than 18" to "over 75").

The type of the variables of interest influences the choice of the method of analysis to be applied. Methods for nominal variables can also be used for ordinal or interval variables, without utilizing however the additional features of ordered categories of unknown or known distances, respectively. On the other hand, methods for ordinal variables require ordered categories and thus are appropriate for interval variables (ignoring the distances between categories) but inappropriate for nominal ones. Binary variables can be treated as nominal or ordinal. Contingency tables with ordinal classification variables are called *ordinal contingency tables*.

1.1.2 Response and Explanatory Variables

Variables are characterized as *response* or *explanatory*, according to their role in the analysis. In some problems the interest lies on detecting and analyzing possible interactions between variables. In these cases the variables are treated symmetrically. However, very often, from the nature of the problem, the interaction is not symmetric but directed. For example, the educational level of the mother affects the school performance of her child and not the opposite. Thus, in such a context, we are interested on testing whether a set of explanatory variables affects one or more response variables. In other words, the response variable is the "dependent" variable of the problem and the explanatory the "independent" ones. In medical applications, the explanatory variables are also known as *prognostic factors*.

The choice of type of model to be used depends on the existence or not of response variables. Special models apply for response variable analysis that provides more straightforward and sound physical interpretation (see Chap. 8).

1.2 Discrete Distributions and Related Inference Problems

The most common probability distributions for categorical data are the binomial, the multinomial, and Poisson distributions and they are briefly reviewed in this section. Additionally, the hypergeometric distribution is presented, which forms the basis for exact inference in 2×2 contingency tables and the famous *Fisher's exact test* (see Sect. 2.1.7).

1.2.1 Binomial Distribution

A trial with two possible outcomes, usually referred as "success–failure," is repeated n times. We assume that the probability of success π is common for all trials (i.e., the trials are *identical*) and that the outcome of one trial does not affect the outcome of any other (i.e., the trials are *independent*). A trial with these characteristics is called *Bernoulli trial*. The random variable X counting successes out of n such Bernoulli trials takes values in $S = \{0,1,2,\ldots,n\}$ and has the *binomial* distribution, denoted by

$$X \sim \mathscr{B}(n,\pi).$$

The probability that X takes a specific value $x \in S$, $\mathsf{P}(X = x)$, is

$$\mathsf{P}(X = x) = \frac{n!}{x!(n-x)!}\,\pi^x(1-\pi)^{n-x}, \quad x = 0,1,2,\ldots,n. \tag{1.1}$$

The mean (expected value) and variance of $X \sim \mathscr{B}(n,\pi)$ are

$$\propto = \mathrm{E}(X) = n\pi, \quad \sigma^2 = \mathrm{Var}(X) = n\pi(1-\pi)$$

The characterization of an outcome as "success" or "failure" is a convention not necessarily having this meaning. Their role can change, since if $X \sim \mathscr{B}(n;\pi)$, then $n - X \sim \mathscr{B}(n,1-\pi)$.

An important property of the binomial distribution is that the sum of two independent binomial distributions $X_i \sim \mathscr{B}(n_i,\pi)$, $i = 1,2$, of the same success probability π and probably different number of trials n_i $(i = 1,2)$ is also binomial distributed:

$$X_1 + X_2 \sim \mathscr{B}(n_1 + n_2, \pi) \tag{1.2}$$

The property holds also for the sum of more than two independent binomials with common success probability.

For n *sufficiently* large, the standardized version of X is approximated by the standard normal distribution, i.e.,

$$Z = \frac{X - \propto}{\sigma} = \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}} \sim \mathscr{N}(0,1) \tag{1.3}$$

In practice n is considered *sufficiently large* if $n\pi \geq 5$ and $n(1-\pi) \geq 5$.

1.2.1.1 The Binomial Distribution in R

For a random variable X, binomial distributed with $n = 10$ and success probability $\pi = 0.3$, i.e., $X \sim \mathscr{B}(10, 0.3)$, the probability $\mathsf{P}(X = 2)$ is computed in R as

```
> dbinom(2,10,0.3)
[1] 0.2334744
```

while the cumulative probability $\mathsf{P}(X \leq 2)$ by

```
> pbinom(2,10,0.3)
[1] 0.3827828
```

The functions `dbinom(x,n,p)` and `pbinom(x,n,p)`, evaluating the probability mass function and the cumulative distribution function, respectively, for $X = x$, can also be applied on vectors of numbers of successes. Thus, for $X \sim \mathscr{B}(5, 0.1)$ they are derived in R by applying the corresponding functions on the vector of all possible outcomes $x = (0, 1, \ldots, 5)$

```
> x <- 0:5
> dbinom(x,5,0.1)
[1] 0.59049  0.32805  0.07290  0.00810  0.00045  0.00001
> pbinom(x,5,0.1)
[1] 0.59049  0.91854  0.99144  0.99954  0.99999  1.00000
```

Furthermore, the probability mass function can be plotted by the commands

```
> plot(x,dbinom(x,5,0.1), type="h", ylim=c(0,1),lwd=5, lend=3,
+    frame.plot=F, xaxt="n", main = "0.1", ylab="P(X=x)")
> axis(1, at=x, pos=c(0,0))
```

while the plot of the cumulative distribution function is defined analogously.

The normal approximation to the binomial distribution is visualized in Fig. 1.1, where the probability mass functions of binomial distributions with $n = 5$ and $n = 25$ and success probabilities 0.05, 0.25, and 0.5 along with the corresponding normal approximations are to be seen. The R function used to provide this figure is provided in the web appendix (see Sect. A.3.1).

1.2.2 Multinomial Distribution

Consider a trial with K possible outcomes, $K \geq 2$, denoted by $A_1, A_2, \ldots, A_K$. The number of outcomes K is fixed and the probability for each of them to occur is positive and constant across independent trials, equal to π_k, $k = 1, \ldots, K$, with $\sum_{k=1}^{K} \pi_k = 1$. The K outcomes are all possible levels of a categorical variable X, taking values in $\{1, 2, \ldots, K\}$. For a random sample of n independent trials, let $(N_1, \ldots, N_K)$ be the random category frequencies of X. Since the sample size n is *fixed*, $\sum_{k=1}^{K} N_k = n$ and thus $K - 1$ of the category frequencies are random. Under this sampling design, the probability of a sample of observed frequencies $\mathbf{n} = (n_1, \ldots, n_K)$, with $\sum_{k=1}^{K} n_k = n$, to occur is

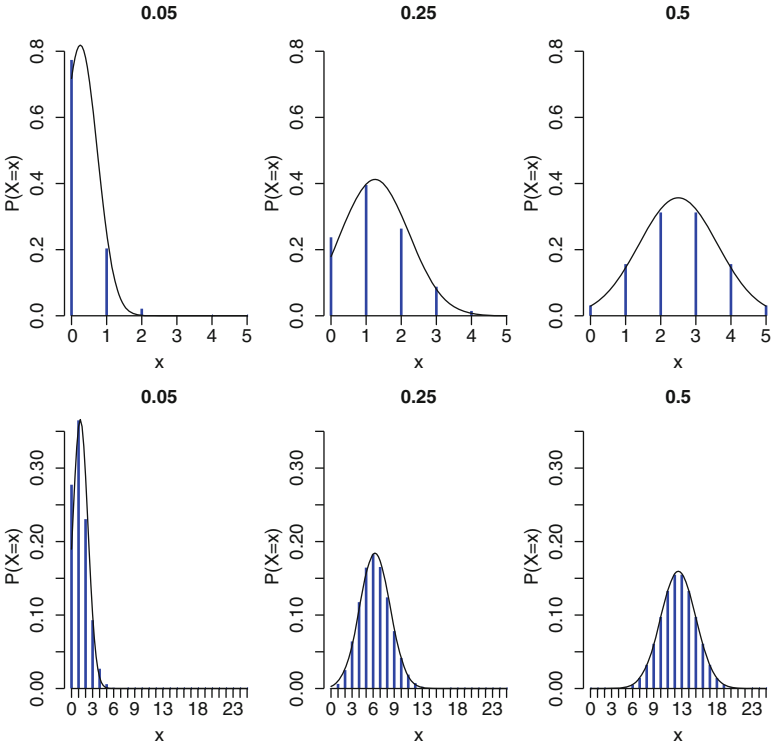


Fig. 1.1 Probability mass functions of binomial distributions and associated normal approximations for $n = 5$ (*first row*) and $n = 25$ (*second row*) and characteristic choices of success probabilities ($\pi = 0.05,\ 0.25,\ 0.5$) in columns

$$p(n_1,\ldots,n_K) = P(N_1 = n_1,\ldots,N_K = n_K) = \left(\frac{n!}{n_1!n_2!\cdots n_K!}\right)\pi_1^{n_1}\pi_2^{n_2}\cdots\pi_K^{n_K} \quad (1.4)$$

We denote, in terms of the random category frequencies,

$$(N_1,\ldots,N_{K-1}) \sim \mathscr{M}\left(n;(\pi_1,\ldots,\pi_{K-1})\right).$$

For $K = 2$, the multinomial distribution reduces to the binomial $\mathscr{B}(n;\pi_1)$.

It is straightforward to obtain probabilities (1.4) in R. For example, for $(N_1,N_2) \sim \mathscr{M}\left(10;(0.35,0.25)\right)$, the probability $P(3,2,5) = 0.0691$ is calculated by

```
> x <- c(3,2,5);  dmultinom(x, prob = c(0.35, 0.25, 0.4))
```

The mean and variance of N_k, $k = 1,\ldots,K$, are

$$\mathrm{E}(N_k) = n\pi_k, \quad \mathrm{Var}(N_k) = n\pi_k(1-\pi_k)$$

and each of them is marginally binomial distributed:

$$N_k \sim \mathscr{B}(n;\pi_k), \quad k = 1,2,\ldots K$$

Since the probabilities for outcomes $A_1, A_2, \ldots, A_K$ sum to 1, it is expected that the variables $N_1, \ldots, N_K$ of their corresponding frequencies are negatively correlated. Indeed, the covariance of any N_i, N_j, for $i \neq j$, is $\mathrm{Cov}(N_i, N_j) = -n\pi_i\pi_j$ and thus

$$\mathrm{Corr}(N_i, N_j) = -\frac{n\pi_i\pi_j}{\sqrt{[n\pi_i(1-\pi_i)][n\pi_j(1-\pi_j)]}} = -\sqrt{\frac{\pi_i}{1-\pi_i} \cdot \frac{\pi_j}{1-\pi_j}}$$

An important property of the multinomial distribution is that any collapsing between categories leads to a multinomial distribution with fewer categories. The probabilities for the new categories are derived by summing the probabilities of the combined categories. For example, if the initial categories $A_1, \ldots, A_6$ are combined to $B_1 = A_1, A_2$, $B_2 = A_3$, $B_3 = A_4, A_5, A_6$, then the initial distribution

$$(N_1, \ldots, N_5) \sim \mathscr{M}\,(n; (\pi_1, \ldots, \pi_5))$$

with corresponding probability vector $\pi^{\mathrm{T}} = (\pi_1,\ \pi_2,\ \ldots,\ \pi_6)$ leads to

$$(Y_1, Y_2) \sim \mathscr{M}\,(n; (\pi_1^*,\ \pi_2^*))$$

with $(Y_1,\ Y_2) = (N_1 + N_2,\ N_3)$ and $(\pi_1^*,\ \pi_2^*) = (\pi_1 + \pi_2,\ \pi_3)$. Obviously, $Y_3 = n - Y_1 - Y_2 = N_4 + N_5 + N_6$ and the associated probability vector is $\pi^{*\mathrm{T}} = (\pi_1^*,\ \pi_2^*,\ \pi_3^*)$ with $\pi_3^* = \pi_4 + \pi_5 + \pi_6$. This property makes data reduction in contingency tables feasible. Data reduction is meant as either collapsing categories of a classification variable or collapsing classification variables themselves in multi-way tables.

Another property of the multinomial distribution refers to the distribution of a subset of outcomes, conditional on their total number of observations. If, without loss of generality, we are interested in the first q ($q < K$) outcomes, then

$$(N_1, \ldots, N_{q-1}) \sim \mathscr{M}\left(n; (\pi_{1|q}, \ldots, \pi_{q-1|q})\right) \tag{1.5}$$

with components of the probability vector equal to $\pi_{k|q} = \frac{\pi_k}{\pi_1 + \ldots + \pi_q}$, $1 \leq k \leq q$. In contingency tables' framework, this property lies behind the equivalence of the "multinomial" and "product multinomial" sampling schemes (see Sect. 2.2.1).

1.2.3 Poisson Distribution

It can be that frequency data do not arise from a fixed number of trials. A characteristic example is the number of car accidents that happened in a region during

the weekend. Thus, if the number of events is X, there exists no upper limit for it. What is fixed in such experiments is an interval, within which we count the event occurrences. The interval is usually of time but can be of any other form as well, such as space. The simplest distribution adequate for setups of this type is the Poisson. If X is Poisson distributed, $X \sim \mathscr{P}(\lambda)$, then

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots, \tag{1.6}$$

where the parameter $\lambda > 0$ is the expected number of arrivals in the specified interval, i.e., $\mathrm{E}(X) = \lambda$. Note that for the Poisson distribution the variance equals the mean, that is, $\mathrm{Var}(X) = \mathrm{E}(X) = \lambda$.

In `R`, Poisson probabilities and cumulative probabilities are computed by `dpois()` and `ppois()`, respectively. Thus, if $X \sim \mathscr{P}(2)$, then

```
> dpois(3,2)
```

calculates $P(X = 3) = 0.18045$, while $P(X > 4) = 0.05265$ is computed as

```
> 1-ppois(4,2)
```

As λ increases, the Poisson distribution is approximated by a normal. For large λ

$$Z = \frac{X - \propto}{\sigma} = \frac{X - \lambda}{\sqrt{\lambda}} \sim \mathscr{N}(0, 1)$$

The binomial distribution with large n and small p is approximated by a Poisson with $\lambda = np$.

A handy property of Poisson distributions is that if $X_1, \ldots, X_K$ are *independent* Poisson random variables with parameters $\lambda_1, \ldots, \lambda_K$, respectively, then their sum $\sum_{k=1}^{K} X_k$ is also Poisson distributed with parameter $\lambda = \sum_{k=1}^{K} \lambda_k$.

Based on this property, the following can be proved that connects the Poisson to the multinomial distribution.

The conditional distribution of $X_1, \ldots, X_K$, given their sum $\sum_{k=1}^{K} X_k = n$, is

$$P[(X_1{=}n_1, \ldots, X_K{=}n_K) \mid \sum_{k=1}^{K} X_k{=}n] = \frac{P(X_1 = n_1, \ldots, X_K = n_K)}{P(\sum_{k=1}^{K} X_k = n)} = \frac{\prod_k \left(\frac{e^{-\lambda_k} \lambda_k^{n_k}}{n_k!} \right)}{\frac{e^{-\lambda} \lambda^n}{n!}},$$

which, since $\sum_{k=1}^{K} \lambda_k = \lambda$ and $\sum_{k=1}^{K} n_k = n$, yields

$$P[(X_1 = n_1, \ldots, X_K = n_K) \mid \sum_{k=1}^{K} X_k = n] = \frac{n!}{n_1! n_2! \cdots n_K!} \prod_k \left(\frac{\lambda_k}{\lambda} \right)^{n_k}, \tag{1.7}$$

which is the multinomial distribution $\mathscr{M}(n; \pi)$, with $\pi_k = \lambda_k / \lambda$, $1 \le k \le K$, being the components of π.

This last result is practically important because it states that if we have K possible response categories and count the number of their occurrences $n_1, \ldots, n_K$, without any restriction on their total, then conditioning on their sum $\sum_{k=1}^{K} n_k = n$ afterward, we can consider that the underlying variable is multinomial distributed. In contingency table analysis, this proves the inferential equivalence of the *multinomial* and *independent Poisson* sampling schemes (see Sect. 2.2.1).

We have seen that the Poisson distribution has variance equal to its mean. In order to model event occurrences of higher variance (*overdispersion*), an alternative distribution is the negative binomial. This distribution is not used in the sequel and thus not presented.

1.2.4 Hypergeometric Distribution

In a binary data setup, the success probability p may not be constant from trial to trial (for example, when sampling from finite populations without replacement). In such cases, the binomial distribution is no longer adequate and other distributions are required. Consider a population of N items, with M of them being of "type A." If a sample of size q is selected from this population, then the number X of "type A" items in the sample is modeled by the hypergeometric distribution, $X \sim \mathscr{H}(N, M, q)$, according to which

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{q-x}}{\binom{N}{q}}, \quad \max(0,\ q + M - N) \le x \le \min(q, M) \tag{1.8}$$

The mean and variance of the hypergeometric X are

$$\mathrm{E}(X) = \frac{qM}{N}, \qquad \mathrm{Var}(X) = \frac{qM(N-q)(N-M)}{N^2(N-1)}$$

When M is small compared to N, then the hypergeometric distribution resembles the binomial.

The hypergeometric distribution is the basis for the *Fisher's exact test* of independence for 2×2 contingency tables (Sect. 2.1.7).

Hypergeometric probabilities in `R` are computed by `dhyper()`. If, for example, $X \sim \mathscr{H}(10, 6, 4)$, then $P(X = 2) = 0.3709$ is computed as

```
> dhyper(2, 10, 6, 4)
```

Furthermore $P(X \le 2) = 0.4890$ is obtained by

```
> phyper(2, 10, 6, 4)
```

1.3 Statistical Inference with Categorical Data

The *likelihood function* is a key quantity in statistical inference as a basic tool for estimation and hypothesis testing. The likelihood $L(\mathbf{x};\ \theta)$ is a function of the sample and the parameter and is the probability of observing this sample $\mathbf{x}$ as a function of the parameter θ.

The most classical method of parameter estimation is the method of maximum likelihood, according to which the *maximum likelihood* (ML) estimate of a parameter θ is the value that maximizes the likelihood with respect to θ, denoted as $\hat{\theta}$. At $\hat{\theta}$, the observed sample has the highest occurrence probability. MLEs have many attractive properties that justify their dominance. In practice it is often easier to maximize the $\log L$ instead of the likelihood L.

For example, for the binomial distribution $X \sim \mathscr{B}(n;\pi)$, the parameter is $\theta = \pi$ and, upon observing $X = x$, the likelihood is

$$L(x;\ \pi) = \frac{n!}{x!(n-x)!}\ \pi^x(1-\pi)^{n-x}, \quad x = 0,1,2,\ldots,n.$$

It can easily be derived (solving $\frac{\partial \log L(x;\ \pi)}{\partial \pi} = 0$ and verifying that $\frac{\partial^2 \log L(x;\ \pi)}{\partial \pi^2} < 0$) that it is maximized with respect to π at

$$p = \frac{x}{n},$$

the observed sample proportion of successes. Thus the ML estimate of π is p.

For the random number of successes $X \sim \mathscr{B}(n;\pi)$, $\hat{\pi} = \frac{X}{n}$ is a random variable with $\mathrm{E}(\hat{\pi}) = \pi$ and $\mathrm{Var}(\hat{\pi}) = \frac{\pi(1-\pi)}{n}$, i.e., the MLE $\hat{\pi}$ is unbiased and its standard error (s.e.) is estimated by

$$\mathrm{SE}(p) = \sqrt{\frac{p(1-p)}{n}}.$$

Analogously, the ML estimates of the category probabilities of the multinomial distribution $\mathscr{M}(n;\pi)$ are proved to be the observed sample proportions, i.e., $p_k = \frac{n_k}{n}$, $k = 1,\ldots,K$.

In general, for a scalar parameter θ, if $\hat{\theta}$ is its MLE and $\mathrm{SE}(\hat{\theta})$ the estimated s.e. of $\hat{\theta}$, then standard methods are applied for related asymptotic hypothesis testing and confidence intervals derivation, based on the asymptotic normality of the MLE. Thus, the $(1-\alpha)100\%$ two-sided Wald confidence interval (CI) is

$$\hat{\theta} \pm z_{\alpha/2}\mathrm{SE}(\hat{\theta}), \tag{1.9}$$

where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ is the $\alpha/2$ upper quantile of $\mathscr{N}(0,\ 1)$. A discussion on its performance and comparison to alternative confidence intervals is provided,

among others, by Brown et al. (2001) and Cai (2005). In R this is easily obtained by the function

```
CI <- function(mle, se, conf.level=0.95)
        {mle+c(-1,1)*se*qnorm(0.5*(1+conf.level))}
```

The Wald CI is the most well-known and easiest to derive asymptotic CI. It can be constructed by inverting the asymptotic *Wald test* for testing the null hypothesis

$$H_0: \ \theta = \theta_0 \tag{1.10}$$

vs. the two-sided alternative $H_1: \ \theta \neq \theta_0$, for known θ_0. The Wald test statistic for a scalar parameter θ, like in (1.10), is given by

$$W = \left(\frac{\hat{\theta} - \theta_0}{\mathrm{SE}(\hat{\theta})} \right)^2. \tag{1.11}$$

Under (1.10), W is asymptotically $\mathscr{X}^2_1$ distributed, or equivalently, $Z = \frac{\hat{\theta}-\theta_0}{\mathrm{SE}(\hat{\theta})}$ is $\mathscr{N}(0,\ 1)$ distributed. In particular, (1.9) is the set of all θ_0, for which (1.10) is not rejected at significance level $\alpha = 0.05$.

Asymptotically equivalent alternative tests to Wald for testing (1.10) are the score test and the likelihood ratio (LR) test. The first is based on the score test statistic

$$S = \left(\frac{u(\theta_0)}{\mathrm{SE}(u(\theta_0))} \right)^2 = \frac{\left(\partial \log L(\theta)/\partial\theta \,\big|_{\theta=\theta_0} \right)^2}{-\mathrm{E}\left(\partial^2 \log L(\theta)/\partial\theta^2 \,\big|_{\theta=\theta_0} \right)},$$

where $u(\theta_0) = \partial \log L(\theta)/\partial\theta \,\big|_{\theta_0}$ is the *score function* (vector of partial derivatives of the log-likelihood with respect to θ, evaluated at θ_0), while the second on the LR test statistic

$$G^2 = 2\left(\log L(\hat{\theta}) - \log L(\theta_0) \right).$$

Under (1.10), S and G^2 are asymptotically $\mathscr{X}^2_1$ distributed. The $(1-\alpha)100\%$ score and LR CIs can be defined analogously to the Wald CI.

The Wald, score, and LR statistics are provided above for testing a scalar parameter. They all extend to vector parameters as well. For example, if the parameter of interest is $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in \Theta$, $K \geq 1$, for testing the hypothesis

$$H_0: \ \boldsymbol{\theta} \in \Theta_0, \tag{1.12}$$

with Θ_0 the parameter subspace under H_0, the LR test statistic is defined as

$$G^2 = 2\log\{L_1/L_0\} = 2\left(\log(L_1) - \log(L_0) \right), \tag{1.13}$$

where $L_0 = \max_{\boldsymbol{\theta} \in H_0} (L(\boldsymbol{\theta}))$ and $L_1 = \max_{\boldsymbol{\theta} \in H_0 \cup H_1} (L(\boldsymbol{\theta}))$. Under (1.12), $G^2 \sim \mathscr{X}^2_{df}$ with $df = \dim(\Theta) - \dim(\Theta_0)$.

Let $(N_1, N_2, \ldots, N_{K-1}) \sim \mathscr{M}(n, (\pi_1, \ldots, \pi_{K-1})$ be a multinomial distributed random sample and consider for its probability vector the following null hypothesis:

$$H_0 : \ \boldsymbol{\pi} = \boldsymbol{\pi}_0. \tag{1.14}$$

The null value $\boldsymbol{\pi}_0 = (\pi_{01}, \ldots, \pi_{0K})$ can be fixed or depending on a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)$ of size $s < K-1$, i.e., $\boldsymbol{\pi}_0 = \boldsymbol{\pi}_0(\boldsymbol{\theta})$. For given n, the expected categories' frequencies under H_0 are $\mathbf{m} = (m_1, \ldots, m_K)$, with $m_k = E(N_k) = n\pi_{0k}$, $k = 1, \ldots, K$, and $\sum_{k=1}^K m_k = n$. In case of parametric $\boldsymbol{\pi}_0$, which is more realistic and common in applications, $\mathbf{m}$ cannot be specified exactly and has to be estimated by the sample. For a sample of observed frequencies $\mathbf{n} = (n_1, \ldots, n_K)$ with $\sum_{k=1}^K n_k = n$, the ML estimate of m_k, $k = 1, \ldots K$, is $\hat{m}_k = n\hat{\pi}_{0k} = n\pi_{0k}(\hat{\theta})$.

In general, a null hypothesis of the type (1.14) imposes a specific structure on the category probabilities of a multinomial distribution. To decide whether an observed data vector $\mathbf{n}$ supports this structure or not, the closeness of the expected under H_0 frequencies m_k to the corresponding observed n_k, $k = 1, \ldots, K$, needs to be evaluated. The further apart the vectors $\mathbf{n}$ and $\hat{\mathbf{m}}$ are, the stronger is the evidence against H_0. The most well-known test statistic for testing H_0 is it Pearson's chi-squared statistic, proposed by Pearson (1900a). This is

$$X^2 = \sum_{k=1}^K \frac{(n_k - \hat{m}_k)^2}{\hat{m}_k}, \tag{1.15}$$

for parametric $\boldsymbol{\pi}_0$. If $\hat{\mathbf{m}} = \mathbf{n}$, then $X^2 = 0$. In all other cases $X^2 > 0$, with larger values indicating stronger deviation from H_0, for fixed sample size n.

For large random samples, under the null hypothesis, X^2 is chi-squared distributed with *degrees of freedom*

$$df = K - 1 - s, \tag{1.16}$$

where s is the number of parameters estimated under the null hypothesis.

For fixed $\boldsymbol{\pi}_0$, no parameter is estimated; thus $s = 0$ and the $\hat{m}_k$'s in (1.15) are replaced by the corresponding m_k's.

An alternative statistic for testing H_0 is the *LR statistic* (1.13), which in this context takes the form

$$G^2 = 2 \sum_{k=1}^K n_k \log(\frac{n_k}{\hat{m}_k}), \tag{1.17}$$

for parametric $\boldsymbol{\pi}_0$. Also $G^2 \geq 0$, with larger values indicating stronger departure from H_0, for fixed n, and $G^2 = 0$ for $\hat{\mathbf{m}} = \mathbf{n}$. The test statistic G^2 is asymptotically

equivalent to X^2, that is, under H_0 and for large n, G^2 is distributed. The $\hat{m}_k$'s in (1.17) are replaced by the corresponding m_k's when $\boldsymbol{\pi}_0$ is fixed.

In practice, H_0 is rejected at significance level α, if the observed value of the test statistic lies in the associated critical region, i.e., if

$$X^2 > \mathscr{X}^2_{df;\alpha} \quad \text{or} \quad G^2 > \mathscr{X}^2_{df;\alpha}$$

for X^2 or G^2, respectively, where $\mathscr{X}^2_{df;\alpha}$ is the α^{th} upper quantile of the $\mathscr{X}^2_{df}$ distribution.

The p-values for these two tests are

$$P(\mathscr{X}^2_{df} > X^2_{\text{obs}}) \quad \text{and} \quad P(\mathscr{X}^2_{df} > G^2_{\text{obs}}),$$

respectively, where X^2_{obs} and G^2_{obs} are the observed values of X^2 and G^2. In `R` the p-value corresponding to a test statistic T with null distribution $\mathscr{X}^2_{df}$ and observed value *Tobs* is easily computed by

```
> p.value <- 1-pchisq(Tobs, df)
```

where `pchisq(Tobs, df)` is the cumulative density function of the $\mathscr{X}^2_{df}$ distribution evaluated at *Tobs*.

X^2 converges faster to the $\mathscr{X}^2$ distribution than G^2 while the approximation is poor for G^2 if $n/K < 5$. In general, for fixed n and K, the approximation by the $\mathscr{X}^2$ distribution is better for lower df. Theoretical results regarding the derivation of asymptotic distributions of parameter estimators or test statistics and their properties are out of the scopes of this book. Such issues are addressed in Bishop et al. (1975, Chap. 14) and in Agresti (2013, Chap. 16).

1.4 Classes of Models for Discrete Data

It is noticeable that methods for the analysis of categorical data have been developed with a certain delay, compared to methods for continuous data. When all d observed attributes in a study are categorical, then the most common way to represent the data is a d-dimensional *contingency table*, produced by cross-classifying the attributes. The information of a contingency table is traditionally summarized through appropriate measures (*measures of association*), which differentiate according to the nature of the underlying classification variables (nominal or ordinal). Association measures, though handy in computation and interpretation, lead to a major loss of information. Models provide a more sensitive analysis.

The most characteristic models for contingency tables are the *log-linear models*. In case some or all of the classification variables are ordinal, special models (log-linear or log-multiplicative) have been developed that utilize the additional information of categories' ordering, assigning scores to them. These are the *association* models, introduced in the 1970s and developed mainly by Leo Goodman. Methods

for analysis of nominal variables are appropriate also for ordinal ones but the reverse is not true. Ordinal data analysis is an area of special interest (Agresti 2010; Liu and Agresti 2005).

Log-linear models treat all the classification variables in a symmetric way in terms of their interactions. Whenever we are interested in the effect of a set of continuous and/or categorical data on a categorical *response variable*, the logistic regression is applied. Logistic regression, initially developed for a *binary* response variable (success–failure), has a long history, as is extensively explained in Cramer (2002). Its origins lie back to the definition of the logistic function by Verhulst in 1840s, rediscovered by Pearl and Reed in 1920. However, logistic regression received attention just in the 1960s. One of its early supporters was Cox, whose book (Cox 1970a) helped in the establishment of the logistic regression model. Logistic regression has been extended to polytomous responses, for nominal or ordinal response variables. One of the most fundamental references is McCullagh (1980), who introduced the *proportional odds model*.

A new boost was given to the analysis of categorical data through the development of the *generalized linear model* (GLM), introduced by Nelder and Wedderburn (1972). Via the GLM, various models for categorical data were unified; their options were naturally extended while some new did arise. Beyond log-linear models, the *Poisson regression* is a classical GLM model for categorical data. Response is the expected number of an event (failures or successes), and it is modeled by a regression function upon a set of explanatory variables. In the presence of *overdispersion*, the *negative binomial model* is applied instead.

An area on categorical data analysis that is of special interest, research- and application-wise, is the analysis of *clustered data*, i.e., data that are correlated. The most common framework is that of *repeated measurements*. Categorical and ordinal repeated measurements are treated either through *marginal* or *conditional models*, depending on whether the effects are population-averaged or subject-specific, respectively. A comprehensive reference book on this field is Molenberghs and Verbeke (2005) while they are extensively treated also in Agresti (2013).

Contemporary problems in categorical data analysis often refer to extreme high-dimensional data that can be clustered. These require the development of complex models and computational demanding procedures. For a more detailed overview on categorical data analysis, see in Kateri (2008).

Overall, the classical reference book of the 1970s on categorical data analysis is that of Bishop et al. (1975). Other fundamental books of the same period are those of Plackett (1974), Upton (1978), Fienberg (first published in 1977; reprint of its revised 2nd edition in 2007), and Haberman (1979). A classical book adjusted to models and applications for social sciences is by Andersen (1980). In our days, the most comprehensive book on categorical data analysis is that of Agresti, in its recent revised 3rd edition (Agresti 2013) and its introductory counterpart (Agresti 2007). Further on, noticeable recent reference books include Zelterman (2006), Simonoff (2003), and Andersen (2001) while Johnson and Albert (2000) is Bayes orientated and restricted to ordinal models. Tutz (2012) focuses on regression models for

categorical data. For an information theoretic approach in the analysis of categorical data, see Gokhale and Kullback (1978a), Read and Cressie (1988), and Pardo (2006).

1.5 Outline of the Book

The focus of this book is on model-based analysis of contingency tables with emphasis to special models for ordinal classification variables. Hence, logistic regression is partially covered, only for contingency tables applications, i.e., when all explanatory variables are categorical as well. Due to space limitations, correlated categorical data are treated only for the paired case.

Basic concepts of two-way contingency tables are introduced in Chap. 2. Traditional descriptive and inferential results on estimation and hypothesis testing are presented and applied on characteristic examples in `R`. The notions of independence and association are extended to multi-way contingency tables in Chap. 3. A model-based approach is adopted in Chap. 4, where the log-linear models are introduced for two- and multi-way tables. The analysis of log-linear models is more flexible in the framework of the GLM. For this, the log-linear models are viewed as members of the GLM family in Chap. 5, making thus easier the consideration of special issues, such as treating structural zeros. Furthermore, the generalized log-linear models (GLLM) are considered, which allow modeling of functions of the cell probabilities. For example, the local odds ratios of a contingency table can be modeled by a GLLM.

Chapters 6–9 are devoted to models for *ordinal contingency tables*. Thus, the association models are discussed in Chap. 6 for two- and multi-way contingency tables. Some more specialized features of association models, like the role of scores in merging categories or association models for odds ratios, are discussed in Chap. 7, along with a short reference to correspondence analysis and its connection to association models. So far, it was considered that the classification variables of a contingency table interact in a symmetric way. In case one of them is a response and the remaining are explanatory variables, special models can be applied, the *logit models*. Logit models for ordinal and nominal response and/or explanatory variables are the subject of Chap. 8. Chapter 9 deals with special models for matched pairs. The models of symmetry, quasi symmetry, marginal homogeneity, and conditional and diagonal symmetry are introduced and discussed. Furthermore, models for rater agreement are presented. Finally, in the epilogue Chap. 10, we briefly refer to alternative models and approaches, not covered in this book, and to recent trends in the contingency table analysis.

The approach adopted in this book is the frequentist approach and the inferential results are asymptotic. Small sample inference is discussed in Sect. 10.4. An area of interest is the Bayesian analysis of contingency tables. The first fundamental results lie back to the classical Bayesian analysis in the 1960s and early 1970s while the interest on the Bayesian approach is renewed in the last two decades, after the

development of the computer-intensive Bayesian computational methods. Bayesian inference overcomes limitations of the classical approach related to asymptotics. They are not asymptotic, so samples need not to be large and they apply easier to sparse tables. Section 10.5 is devoted to the Bayesian analysis of contingency tables, bringing up main issues in the Bayesian analysis of contingency tables and providing an extended bibliography review.

The book is applications orientated. For this, though the theory is presented detailed and accurate, theoretical results are not derived but rich relevant citations are provided. The aim is to familiarize readers with these models and their application in practice although some of them cannot be directly fitted in statistical packages. To achieve this, all examples are worked out in `R` and presented in a way that can be followed in `R` by the reader. This procedure is simplified through handy `R` functions that are used throughout the book and provided in the book's web companion, available at

http://cta.isw.rwth-aachen.de.

It consists mainly of an `R` Appendix, where the `R` options for contingency tables are discussed, helpful `R` packages are listed, and all the functions and examples used in the book are available to download. Additionally, SPSS syntax code scripts are provided for fitting the association and symmetry models.

Chapter 2
Analysis of Two-way Tables

Abstract Basic concepts of two-way contingency table analysis are introduced. Descriptive and inferential results on estimation and testing of basic hypotheses are discussed and illustrated in R. In particular the comparison of two independent proportions, the test of independence for 2×2 and $I \times J$ contingency tables, the linear trend test, and the Fisher's exact test are presented. Special emphasis is given to the odds ratio for 2×2 tables, while the generalized odds ratios for $I \times J$ tables are treated in detail. Finally, graphical displays of categorical data (barplot, fourfold plot, sieve diagram, and mosaic plot) are derived using R for examples of this chapter and discussed.

Keywords Binary variables • Odds ratio • Fisher's exact test • Independence for $I \times J$ tables • Residuals • Generalized odds ratios • Linear trend test • Fourfold plots • Sieve diagrams • Mosaic plots

2.1 Analyzing 2×2 Tables

2×2 contingency tables are very common in biomedical and social sciences applications, where binary variables (yes–no) play an important role, in the context of survival, success of a treatment, or presence of a characteristic or prognostic factor. The extent of related literature is impressive and this very simple table keeps the continuous interest of researchers since the early 1900s. Indicatively we mention that Upton (1982) compared twenty-two alternative tests of the literature for the 2×2 comparative trial commenting that the range of different possible sampling schemes for 2×2 tables is responsible for this amount of literature.

Different sampling schemes correspond to different experimental scenarios and to different hypotheses of interest. A 2×2 table can arise by cross-classifying two binary variables on a sample. If X and Y are the row and column classification variables, respectively, then the hypothesis of interest is the independence of X and Y. Alternatively, a binary response for two independent samples can be reported

M. Kateri, *Contingency Table Analysis: Methods and Implementation Using R*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4811-4_2,

by a 2×2 table, setting, for example, the response in the column variable Y and letting X define the two underlying populations. In this case, the hypothesis to be tested is the equality of the success probabilities in Y for the two independent binomial populations. Notation is unified for both cases. Thus, for a data set, let n_{ij} denote the observed cell frequency at cell (i, j), $i, j = 1, 2$, i.e., the number of cases for which the combination $X = i$ and $Y = j$ is observed. Notation-wise, the first index (i) stands for the row and the second (j) for the column category. Then, $n_{i+} = n_{i1} + n_{i2}$ is the marginal frequency of the ith row, $i = 1, 2$, while the marginal column frequencies are defined analogously, $n_{+j} = n_{1j} + n_{2j}$, $j = 1, 2$. In general a "+" in place of an index denotes summation over this index. Finally, $n = n_{++} = n_{1+} + n_{2+} = n_{+1} + n_{+2}$ is the total number of observations of the data set. In table form this is stated as follows:

n_{11}	n_{12}	n_{1+}
n_{21}	n_{22}	n_{2+}
n_{+1}	n_{+2}	n

The sampling scheme underlying the first scenario is either a multinomial distribution of four categories, corresponding to the cells of the table, or four independent Poisson distributions, one for each cell. In the first case the total sample size n is fixed (known) while in the second random. In the second scenario, we observe two independent binomial samples, one for each row, of sizes n_{1+} and n_{2+}, respectively. Thus, one set of margins is fixed, here the row marginals (n_{1+}, n_{2+}). Obviously, the case of fixed column marginals is analogous. These two scenarios will be treated in Sects. 2.1.1 and 2.1.3, respectively.

To illustrate, consider the indicative examples of Table 2.1. Data in Table 2.1(a) present a sample of size $n = 3213$, collected in the period 1980–1983 in the St. Louis Epidemiologic Catchment Area Survey (Glassman et al. 1990) and cross-classified according to regular smoking habit (rows) and major depressive disorder (columns). Interest lies on testing for possible relation between cigarette smoking and major depressive disorder. Table 2.1(b) reports the binary response (success–failure) of two treatments (high–low dose) received by two independent samples of patients (hypothetical data). The goal is to compare the success probabilities for the high and low dose treatment, based on two independent samples of patients. In Table 2.1(a) the total sample size is fixed while in Table 2.1(b) the row marginals are fixed, not necessarily equal. Data in Table 2.1(c) seem similar to Table 2.1(b) and serve the same goal, but the experiment is designed differently; they correspond to a crossover study. Just one sample of patients is considered and they receive both treatments in sequence, after a follow-up period (hypothetical data). A pair of responses is available for each patient and Table 2.1(c) cross-classifies these responses, reporting the number of patients for which both treatments were successful, both failed, or only the high or low dose was successful. This last example is a longitudinal study. As in Table 2.1(b), the success probabilities for high and low dosages have to be compared. However, it is different from the second scenario setup, since the proportions to be compared are dependent. At this point, we will deal with the first two problems while we will return to the dependent proportions comparison in Sect. 9.3.

Table 2.1 (a) Survey respondents cross-classified by smoking habit and major depressive disorder (Glassman et al. 1990). (b) Response for two independent samples of low and high dose treatments (hypothetical data). (c) Crossover trial comparing low and high dose treatments on a sample of 100 patients (hypothetical data)

(a)

Ever smoked	Major depression	
	Yes	No
Yes	144	1729
No	50	1290

(b)

	Response	
Dose	Success	Failure
High	41	9
Low	37	13

(c)

High dose	Low dose	
	Success	Failure
Success	62	18
Failure	8	12

2.1.1 Independence of Two Binary Variables

For the 2×2 contingency table $\mathbf{n} = (n_{ij})$ that cross-classifies two binary variables X and Y on a sample of fixed size n, let N_{ij} be the random number of observations in cell (i, j) and $\pi_{ij} = P(X = i,\ Y = j)$ the associated cell probability, $i, j = 1, 2$. Since the total sample size is fixed, $\sum_{i,j} N_{ij} = n$ and thus only three cell frequencies of the table $\mathbf{N} = (N_{ij})$ are random. Thus $\sum_{i,j} \pi_{ij} = 1$ and the underlying distribution is the multinomial:

$$(N_{11}, N_{12}, N_{21}) \sim \mathcal{M}\left(n, (\pi_{11}, \pi_{12}, \pi_{21})\right) .$$

The probability vector $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ is the *joint distribution* of X and Y. The probability of the i^{th} row category is $P(X = i) = \pi_{i1} + \pi_{i2} = \pi_{i+}$, $i = 1, 2$ and of the j^{th} column category $P(Y = j) = \pi_{1j} + \pi_{2j} = \pi_{+j}$, $j = 1, 2$. The probabilities vectors (π_{1+}, π_{2+}) and (π_{+1}, π_{+2}) are the row and column *marginal distributions*, respectively. In matrix notation, we have

π_{11}	π_{12}	π_{1+}
π_{21}	π_{22}	π_{2+}
π_{+1}	π_{+2}	1

It is well known that variables X and Y are independent if $P(X = i,\ Y = j) = P(X = i)P(Y = j)$ for all possible values of i and j. Thus, in our context the null hypothesis of independence is

$$H_0 \ : \ \pi_{ij} = \pi_{i+}\pi_{+j} \ , \quad i, j = 1, 2 \ . \tag{2.1}$$

For multinomial distribution, the expected cell frequencies are $m_{ij} = n\pi_{ij}$ (adjusting the vector notation of Sect. 1.2.2 to two-way arrays) and under (2.1), $m_{ij} = n\pi_{i+}\pi_{+j}$, $i, j = 1, 2$. The corresponding MLEs are

$$\hat{m}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} \ . \tag{2.2}$$

It can be easily verified that $\hat{\pi}_{i+}(\mathbf{n}) = p_{i+}$, where p_{i+} is the ith row marginal sampling proportion. Analogously, $\hat{\pi}_{+j}(\mathbf{n}) = p_{+j}$ for the column marginal probabilities. Thus, the ML estimates of the expected cell frequencies under H_0 of independence are

$$\hat{m}_{ij} = np_{i+}p_{+j} = \frac{n_{i+}n_{+j}}{n} , \quad i,j = 1,2.$$

Note that the ML estimates of the row and column marginals satisfy $\hat{m}_{i+} = n_{i+}$ and $\hat{m}_{+j} = n_{+j}$, $i,j = 1,2$, respectively.

The within rows probabilities are the *conditional row probabilities*

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} , \quad i,j = 1,2 ,$$

while the *conditional column probabilities* are defined analogously. The independence hypothesis (2.1) could equivalently be expressed in terms of the conditional row probabilities as

$$\pi_{1|i} = \pi_{+1} , \; i = 1,2 , \tag{2.3}$$

which means that under independence the within rows success probability is the same for both rows (obviously $\pi_{2|i} = 1 - \pi_{1|i}$, $i = 1,2$).

Actually only one of the row marginals and one of the column marginals probabilities, say π_{1+} and π_{+1}, respectively, need to be estimated in (2.2), since $\sum_{i=1}^{2} \pi_{i+} = \sum_{j=1}^{2} \pi_{+j} = 1$. Thus, the number of parameters to be estimated under H_0 is $s = 2$ and Pearson's X^2 statistic (1.15) becomes

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \tag{2.4}$$

The asymptotic distribution for (2.4) under H_0 is $\mathscr{X}_1^2$. Alternatively, the asymptotic equivalent LR statistic (1.17) can be applied, here expressed as

$$G^2 = 2\sum_{i,j} n_{ij} \log(\frac{n_{ij}}{\hat{m}_{ij}}) \tag{2.5}$$

Yates (1934) suggested to correct the Pearson's X^2 test (2.4) in order to reduce the approximation error encountered by approximating the binomial distribution by the continuous chi-square distribution; therefore, the correction is known as *continuity correction*. The formula of the Yates' corrected X^2 is

$$X^2 = \sum_{i,j} \frac{(|n_{ij} - \hat{m}_{ij}| - 0.5)^2}{\hat{m}_{ij}} .$$

This correction reduces the Pearson's X^2 statistic value and consequently increases the corresponding p-value.

2.1.2 Example 2.1(a)

Applying the procedure described above on the smoking vs. depression data, we get $X^2 = 21.557$, highly significant for $df = 1$ (p-value < 0.00005). Thus, we conclude that indeed, as expected, the smoking habit is strongly related to depression. The ML estimates of the expected cell frequencies under the H_0 of independence ($\hat{m}_{ij}$) are

	Depression	
Ever_Smoker	Yes	No
Yes	113.091	1759.909
No	80.909	1259.091

Observing that the observed frequency of people smoking and with a depression ($n_{11} = 144$) is higher than the corresponding expected ($\hat{m}_{11} = 113.09$), we can conclude about the direction of the association. In particular, the probability of smoking is higher for people who have experienced a major depressive disorder. Identification of the cells that are responsible for the deviation from H_0 and evaluation of their contribution, in strength and direction, are achieved by the inspection of the *residuals*, presented for the general $I \times J$ table in Sect. 2.2.4.

The X^2 test of independence is very easily implemented in any statistical package. In R, the appropriate function is `chisq.test()` that reads the data in a matrix form. For this example, we enter the data and the labels for the variables' names and their values as

```
> depsmok <- matrix(c(144,1729,50,1290),byrow=T,ncol=2);
> dimnames(depsmok) <- list(Ever_Smoker=c("Yes","No"),
+        Depression=c("Yes","No"));
```

The created frequency table can be viewed by typing `depsmok`. The table can be enriched with the row and column marginals as follows:

```
> addmargins(depsmok)
```

```
                Depression
Ever_Smoker   Yes     No    Sum
        Yes   144   1729   1873
         No    50   1290   1340
        Sum   194   3019   3213
```

Command `prop.table(depsmok)` computes the sampling proportions while the proportions table along with the marginal proportions will be printed by

```
> addmargins(prop.table(depsmok))
```

```
                Depression
Ever_Smoker       Yes       No      Sum
        Yes    0.0448    0.5381   0.5829
         No    0.0156    0.4015   0.4171
        Sum    0.0604    0.9396   1.0000
```

The row conditional proportions are derived by `prop.table(depsmok,1)`. Analogously, the column conditional proportions are

```
> prop.table(depsmok,2)
```

```
                  Depression
Ever_Smoker          Yes      No
        Yes       0.7423  0.5727
         No       0.2577  0.4273
```

Command

```
> chisq.test(depsmok)
```

computes the X^2 test of independence providing the following output:

```
Pearson's Chi-squared test with Yates' continuity correction
data: depsmok
X-squared = 20.8652,   df = 1,   p-value = 4.928e-06
```

For 2×2 tables, the standard expression of `chisq.test()` engages the continuity correction of Yates (see Sect. 1.3). The test without the continuity correction is fitted by

```
> chisq.test(depsmok, correct = FALSE)
```

```
            Pearson's Chi-squared test
data: depsmok
X-squared = 21.557,   df = 1,   p-value = 3.435e-06
```

The ML estimates of the expected cell frequencies under H_0 are derived by

```
> chisq.test(depsmok)$expected
```

`chisq.test()` does not provide the G^2 statistic (2.5). This can be computed, along with the associated p-value, by the function `G2()`, which is based on `chisq.test()` and is provided in the web appendix (see Sect. A.3.2). For our example we apply

```
> G2(depsmok)
```

```
$G2
[1] 22.75493
$df
1
$p.value
[1] 1.840319e-06
```

The options and features of `chisq.test()` will be further discussed in the context of the general $I \times J$ contingency tables later in Sects. 2.2.3 and 2.2.4.

2.1.3 Comparison of Two Independent Proportions

Consider data of the type of Example 2.1(b) and let n_{11} and n_{21} be the frequencies of successes for two independent samples of sizes n_1 and n_2, respectively. Then, for a sample of fixed sample size n_i from the ith population ($i = 1, 2$), the random number of successes N_{i1} for population i is binomial distributed

$$N_{i1} \sim \mathcal{B}(n_i, \pi_i)$$

and the two distributions are independent. The underlying probability pattern of the 2×2 contingency table formed by two independent binomials is

π_1	1-π_1	1
π_2	1-π_2	1

The basic associated hypothesis testing problem is

$$H_0 \ : \ \pi_1 = \pi_2 \ (= \pi) \tag{2.6}$$

and can be faced by a number of alternative approaches. The most direct is the well-known asymptotic Z test with test statistic

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1} + \frac{1}{n_2})}} \overset{H_0}{\sim} \mathcal{N}(0,1) \ , \tag{2.7}$$

where $\hat{\pi}_1 = N_{11}/n_1$ and $\hat{\pi}_2 = N_{21}/n_2$ are the random sample success proportions for the 1st and 2nd sample, respectively, while $\hat{\pi} = \frac{N_{11}+N_{21}}{n_1+n_2}$ is the MLE of the common success probability under H_0. This test is based on the normal approximation of a binomial distribution (1.3) and the fact that under H_0, $N_{i1} + N_{i2} \sim \mathscr{B}(n_1 + n_2, \pi)$ (see property (1.2)).

Possible alternatives to (2.6) are

$$H_{1a} \ : \ \pi_1 > \pi_2 \quad \text{or} \quad H_{1b} \ : \ \pi_1 < \pi_2 \quad \text{or} \quad H_1 \ : \ \pi_1 \neq \pi_2 \ .$$

The null hypothesis (2.6) is then rejected at significance level α in favor of the one-sided alternatives H_{1a}, H_{1b} or the two-sided H_1, if $Z \geq z_\alpha$, $Z \leq -z_\alpha$, or $|Z| \geq z_{\alpha/2}$, respectively.

The asymptotic $(1-\alpha)100\%$ Wald CI for the difference $\pi_1 - \pi_2$ is

$$\left(p_1 - p_2 - z_{a/2}\sqrt{\widehat{\text{Var}}(\hat{\pi}_1 - \hat{\pi}_2)}, \quad p_1 - p_2 + z_{a/2}\sqrt{\widehat{\text{Var}}(\hat{\pi}_1 - \hat{\pi}_2)} \right) \ , \tag{2.8}$$

where $\text{Var}(\hat{\pi}_1 - \hat{\pi}_2)$ is equal to

$$\text{Var}(\hat{\pi}_1 - \hat{\pi}_2) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2} \tag{2.9}$$

and is estimated by substituting in (2.9) the probabilities with the corresponding sample proportions $p_i = n_{i1}/n_i, \ i = 1,2$. For alternative methods of constructing confidence intervals for the difference of independent binomial proportions and simulation based comparisons among them, we refer to Newcombe (1998) and Brown and Li (2005).

Such a data setup could also be viewed in a 2×2 contingency table form, in the context of Sect. 2.1.1, produced by cross-classifying variables X for the sample (1st and 2nd) and Y for the response (success–failure). Then

$$\pi_i = P(Y = 1 \,|X = i) = \pi_{1|i}\,, i = 1,\ 2, \tag{2.10}$$

and by (2.3) and the equivalence between independence and equality (homogeneity) of conditional row probabilities (see Sect. 2.1.1), we conclude that hypothesis (2.6) can equivalently be viewed as a hypothesis of independence (success is independent of population) and tested by the X^2 test (2.4). In this setup, the X^2 test is known as *test of homogeneity*.

2.1.4 Example 2.1(b)

For the data on successes for the high–low dose treatments [Table 2.1(b)], we have $n_1 = n_2 = 50$ and the Z-test (2.7) gives

$$Z = \frac{0.82 - 0.74}{\sqrt{0.78(1-0.78)(\frac{1}{50} + \frac{1}{50})}} = 0.9656\,,$$

which is nonsignificant. The corresponding X^2 statistic (2.4) is equal to $X^2 = 0.9324$, with p-value = 0.3342 for $df = 1$. (Note that $Z^2 = 0.9656^2 = 0.9324$, as expected.) Thus, though the sample success proportion is higher for the high dose treatment, the difference in success proportion of 8% between high and low doses is not statistically significant for the sample size under consideration.

For the X^2 test of independence, this example can be worked out in `R` by `chisq.test()`, exactly as Example 2.1(a). The Z-test above can be applied by `prop.test()` that has the additional feature of providing the $(1-\alpha)100\%$ confidence interval (2.8) for the difference $\pi_1 - \pi_2$. The following script of commands reads the data, creates labels, and applies the Z-test

```
> dosesuc<- matrix(c(41,9,37,13),byrow=TRUE,ncol=2);
> dimnames(dosesuc) <- list(Dose=c("high","low"),
+        Response=c("success","failure"));
> prop.test(dosesuc, correct=FALSE)
```

The derived output is

```
        2-sample test for equality of proportions
              without continuity correction
data: dosesuc
X-squared = 0.9324,    df = 1,    p-value = 0.3342
alternative hypothesis: two.sided
95 percent confidence interval:
-0.08162277  0.24162277
sample estimates:
prop 1   prop 2
0.82     0.74
```

Since the data support H_0 for $\alpha = 0.05$, the 0.95% CI for the difference $\pi_1 - \pi_2$ includes the value 0.

The $(1-\alpha)100\%$ CI provided by `prop.test()` for the difference of two proportions is the Wald CI, though the CI provided by `prop.test()` for one proportion is the score CI. The score CI for the difference of proportions, along with Wald CI and further types of CIs, can be derived in the `PropCIs` package.

For significance level $\alpha = 0.01$ and for the one-sided alternative $H_1 : \pi_1 > \pi_2$,

```
> prop.test(dosesuc, alternative="greater",
+       conf.level = 0.99, correct=F)
```

leads to

```
      2-sample test for equality of proportions
            without continuity correction
data: dosesuc
X-squared = 0.9324,   df = 1,   p-value = 0.1671
alternative hypothesis: greater
99 percent confidence interval:
-0.1118356  1.000000
sample estimates:
prop 1   prop 2
0.82   0.74
```

2.1.5 The Odds Ratio

For a binary response, results are often presented and interpreted not directly on the success probability π but regarding success's relative importance to failure. Hence, the ratio of success vs. failure probabilities for a response, known as *odds* of success

$$odds = \frac{\pi}{1-\pi},$$

is a key quantity. An odds of 2 means that success is twice as possible as failure for the population under study while of 0.25 that failure is four times more possible than success. When comparing the response of two independent populations, for example, cases/controls, with/without a prognostic factor, or comparing two treatments, as in Example 2.1(b), their odds are compared. If π_1 and π_2 are the success probabilities of the two populations, then their *odds ratio* is defined as

$$\theta = \frac{odds_1}{odds_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \tag{2.11}$$

and is more informative for the comparison of π_1 and π_2 than their difference. For example, the cases $\pi_1 = 0.9$, $\pi_2 = 0.8$ and $\pi_1 = 0.6$, $\pi_2 = 0.5$ have both $\pi_1 - \pi_2 = 0.1$ while their odds ratios are 2.25 and 1.5, respectively, incorporating the relative importance of success probabilities in terms of their level of magnitude.

In terms of the joint distribution of a 2×2 contingency table and due to (2.10), it is easy to verify that θ is equivalently defined as

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \tag{2.12}$$

A value of $\theta = 1$ is equivalent to $\pi_1 = \pi_2$, i.e., to independence of the binary classification variables of the table. A value of $\theta > 1$ or < 1 corresponds to positive or negative dependence, respectively, while dependence becomes stronger as θ moves away from 1.

The odds ratio is a fundamental *association measure* for a 2×2 contingency table and, as we shall see in the sequel, the odds ratio as a concept plays an important role in model formulation and interpretation in contingency table analysis. It does not depend on the marginal distributions of the classification variables and is therefore a good measure of their association. The marginal invariance of θ can easily be verified as follows. When multiplying row i ($i = 1,2$) and/or column j ($j = 1,2$) of the table by a fixed positive number α_i and/or β_j, respectively, the cell probabilities for the derived table are $\pi^*_{ij} = \frac{\alpha_i\beta_j\pi_{ij}}{\sum_{i,j}\alpha_i\beta_j\pi_{ij}}$, $i, j = 1,2$, and θ^* is the corresponding odds ratio. Then, it holds

$$\theta^* = \frac{\pi^*_{11}\pi^*_{22}}{\pi^*_{12}\pi^*_{21}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \theta\ . \tag{2.13}$$

The sample odds ratio is

$$\hat{\theta}(\mathbf{n}) = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}\ . \tag{2.14}$$

The computation of $\hat{\theta}$ is straightforward by (2.12) while definition (2.11) is more convenient for meaningful interpretation. $\hat{\theta}$ takes values in the interval $[0, \infty)$, with $\hat{\theta} = 0$ or $\hat{\theta} = \infty$ when a sampling zero occurs in nominator or denominator of (2.14), respectively, while it is undefined when sampling zeros occur in both cells of a row or column. A classical way to treat such cases is, in presence of sampling zeros, to add 0.5 to the cell frequencies. This procedure has however been criticized, especially in cases of small sample sizes (see discussion in Sect. 2.5.2).

It has been proved that for random sample, $\log\hat{\theta}$ is better normally approximated than $\hat{\theta}$. Thus, inference is drawn in terms of $\log\theta$. In particular, it can be proved that asymptotically

$$\log\hat{\theta} \sim \mathcal{N}(\log\theta,\ \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}) \tag{2.15}$$

Furthermore, in log-scale, interpretation is more straightforward, since independence corresponds to $\log\theta = 0$, positive (negative) dependence to positive (negative) values of $\log\theta$ and the strength of association is increasing in $|\theta|$.

Based on (2.15), the asymptotic $(1-\alpha)100\%$ confidence interval for θ can be derived

$$\left(e^{L(\hat{\theta},0)},\ e^{L(\hat{\theta},2)}\right)$$

where

$$L(\hat{\theta},c) = \log\hat{\theta} - (1-c)z_{\alpha/2}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}$$

Also, hypotheses about θ, like

$$H_0 : \theta = \theta_0 \quad \Leftrightarrow \quad \log\theta = \log\theta_0 \tag{2.16}$$

for θ_0 known, can be asymptotically tested by the associated Z test

$$Z = \frac{\log\hat{\theta} - \log\theta_0}{\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}} \overset{H_0}{\sim} \mathcal{N}(0,1) \tag{2.17}$$

Since $\theta = 1 \Leftrightarrow \pi_1 = \pi_2$, hypothesis (2.16) for $\theta_0 = 1$ is equivalent to the hypothesis of equality of two independent proportions (2.6) or to independence (2.1). However, (2.17) is a Wald test and is not equivalent to the X^2 or G^2 tests (2.4) and (2.5), which are score and LR tests, respectively, and are preferable.

In medical applications, the "success" probabilities π_1 and π_2 refer often to the occurrence of a disease and are therefore called *risk*. The risks of two independent populations are then compared through their ratio, which, as the odds ratio, is more informative than their difference. Thus, the *relative risk* is defined by

$$r = \frac{\pi_1}{\pi_2}\ . \tag{2.18}$$

Substituting in (2.18) the probabilities with the corresponding sampling proportions, the corresponding sampling relative risk $\hat{r}$ is obtained. The odds ratio and relative risk are related through

$$\theta = r \cdot \frac{1-\pi_2}{1-\pi_1}\ . \tag{2.19}$$

The relative risk is easier to interpret than the odds ratio but with the cost that it cannot be defined for all types of studies. Risks can be defined directly only for cohort studies while odds ratios also for case-controls or cross-sectional studies. Also, covariate adjustment, required by some designs, is easier for odds ratios, through logistic regression models, than relative risks (see, e.g., Simon 2001). Therefore, for rare diseases, it is common to compute the odds ratio and interpret it as relative risk, since $\theta \approx r$ for small π_1, π_2, due to (2.19). Furthermore, r does

not exhibit nice mathematical properties, in contrast to θ. From definition (2.12) it can easily be verified that θ is invariant under table rotation while it becomes θ^{-1} when the rows (or columns) are interchanged. These properties do not hold for the relative risk r. Practically speaking, this means that when changing the reference response category, the new θ is simply the reciprocal of the initial one while r has to be recomputed.

2.1.6 *Example 2.1 (Continued)*

For a 2×2 data table, function `odds.ratio()`, to be found in web appendix (see Sect. A.3.2), computes the ML estimate $\hat{\theta}$, its asymptotic $(1-\alpha)100\%$ confidence interval as well as the Z test for testing (2.16) against the two-sided alternative. In case of sampling zeros, `odds.ratio()` adds 0.5 in every cell frequency.

In order to derive the 95% confidence interval for θ and to test the hypothesis of independence ($\theta_0 = 1$, set as default in the function) at $\alpha = 0.05$ (default), function `odds.ratio()` is applied on Table 2.1(a) as

```
> odds.ratio(depsmok)
```

The derived output is

```
$estimator
[1] 2.148757

$asympt.SE
[1] 0.1682201

$conf.interval
[1] 1.545247  2.987971

$conf.level
[1] 0.95

$Ztest
[1] 4.546955

$p.value
[1] 5.442761e-06
```

and $\hat{\theta} = 2.149$ implies that the odds of smoking is 2.15 times higher for people with a major depression disorder than for people without.

An alternative convenient way to apply functions of `R` is to save the output of the function and then extract the parts of the results needed. For example, the test of hypothesis (2.16) for $\theta_0 = 1.7$ at significance level 5% can be saved in `theta1.7` by

```
> theta1.7<- odds.ratio(depsmok, 0.95, 1.7)
```

Then,

```
> theta1.7$Ztest
```

provides just the value of the test statistic (2.17) for $\theta_0 = 1.7$, $Z = 1.3926$, and

```
> theta1.7$p.value
```

the corresponding p-value=0.1637.

For Table 2.1(b), we find by `odds.ratio(dosesuc)` that $\hat{\theta} = 1.6$ and the null hypothesis (2.16) cannot be rejected (p-value= 0.3364). The 95% confidence interval for θ is (0.613420, 4.176457). Thus, the odds of success does not differ significantly for high and low dose treatments, conclusion equivalent to that drawn by the procedure of Sect. 2.1.3 in terms of the difference in success probabilities for high and low dose treatments.

2.1.7 Fisher's Exact Test

We have seen that for 2×2 contingency tables, independence (2.1) can be tested in terms of the odds ratio. The corresponding test discussed in the section above is asymptotic and thus inappropriate for small samples. Fisher introduced an exact test for testing

$$H_0 \ : \ \theta = 1 \ \text{ vs. } \ H_1 \ : \ \theta > 1 \ , \tag{2.20}$$

which is a conditional test and is based on the *hypergeometric* distribution (Fisher 1934). In particular, it can be verified that, under independence, the *conditional distribution* of N_{11}, given n_{1+}, n_{+1}, and $n = n_{++}$, is $N_{11} \sim \mathscr{H}g(n, n_{1+}, n_{+1})$, i.e., hypergeometric with probability function (under independence)

$$p(t) = P(N_{11} = t) = \frac{\binom{n_{1+}}{t}\binom{n - n_{1+}}{n_{+1} - t}}{\binom{n}{n_{+1}}} \ , \tag{2.21}$$

$$\max(0, \ n_{1+} + n_{+1} - n) \le N_{11} \le \min(n_{1+}, n_{+1})$$

The p-value for testing (2.20) equals the sum of the "extreme" probabilities, where "extreme" is meant toward the direction of the alternative. Hence, if t_{obs} denotes the observed value of N_{11}, then

$$P^{+} = \mathrm{P}(N_{11} \ge t_{\text{obs}}) \ . \tag{2.22}$$

For the alternative hypothesis of the opposite direction $\theta < 1$, the p-value is defined analogously as

$$P^{-} = \mathrm{P}(N_{11} \le t_{\text{obs}}) \ . \tag{2.23}$$

Due to the high degree of discreteness of the hypergeometric distribution, when n is small, only a few values can be attained for these p-values. The conservatism of such discrete tests can be attenuated by using the mid-p-values. For the alternative $\theta > 1$, the mid-p-value is defined by

$$\text{mid}-P^+ = \text{P}(N_{11} > t_{\text{obs}}) + \frac{1}{2}\text{P}(N_{11} = t_{\text{obs}}) \ ,$$

while for $H_1 : \theta < 1$, it is defined analogously as

$$\text{mid}-P^- = \text{P}(N_{11} < t_{\text{obs}}) + \frac{1}{2}\text{P}(N_{11} = t_{\text{obs}}) \ .$$

For $H_1 : \theta \neq 1$, the definition of the two-sided p-value is not that obvious. The classical choice for Fisher's exact test is

$$P_\ell = \sum_{t:p(t)\leq p(t_{\text{obs}})} \text{P}(N_{11} = t) \ , \tag{2.24}$$

called by Hirji (2006) "the probability based method," which is the sum of probabilities of outcomes that are at most as probable as the observed outcome t_{obs}. An easy to compute alternative p-value is derived by taking twice the minimum one-tail probability, bounded by 1, i.e.,

$$P_{tw} = \min\{1, 2\min[P^+, P^-]\} \ , \tag{2.25}$$

where P^+ and P^- are given in (2.22) and (2.23), respectively. This is the direct analogue of the definition of two-sided p-values for continuous distributions of test statistics. Another option of p-value that is based on both tail probabilities is

$$P_{CH} = \min[P^+, P^-] + p^* \ , \tag{2.26}$$

where p^* is the one-sided p-value from the other tail of the distribution, nearest to but not exceeding $\min[P^+, P^-]$ (see Cox and Hinkley 1974, p. 79). The computation of exact p-values will be clarified in the example that follows in Sect. 2.1.8.

Based on two-sided Fisher's exact test, an exact $(1-\alpha)100\%$ CI for the odds ratio θ can be constructed by inversion of the exact test that tests the null hypothesis $H_0 : \theta = \theta_0$ vs. the alternative $H_1 : \theta \neq \theta_0$, for $\theta_0 \neq 1$. This test is based on the distribution of N_{11}, given n_{1+}, n_{+1}, and n, when the odds ratio equals θ. This is the *noncentral hypergeometric* with probabilities

$$p(t, \theta) = P(N_{11} = t, \theta) = \frac{\binom{n_{1+}}{t}\binom{n-n_{1+}}{n_{+1}-t}\theta^t}{\sum_{k=t_{\min}}^{t_{\max}}\binom{n_{1+}}{k}\binom{n-n_{1+}}{n_{+1}-k}\theta^k} \ ,$$

$$t_{\min} = \max(0, \ n_{1\cdot} + n_{\cdot 1} - n) \leq N_{11} \leq \min(n_{1\cdot}, n_{\cdot 1}) = t_{\max} \ .$$

For $\theta = 1$, the hypergeometric probability is derived. The associated exact $(1-\alpha)100\%$ CI will consist of the set of θ_0 values for which the corresponding

test fails to reject H_0 at significance level α. The classical CI based on Fisher's exact test is based on the test with two-sided p-value (2.24). Using the p-values defined by (2.25) or (2.26), alternative confidence intervals are derived for θ. The CI based on (2.26) is less conservative than the classical one and was proposed by Blaker (2000). Exact confidence interval can also be derived by the inversion of two one-sided tests. However, the CIs based on the inversion of a single two-sided test are shorter and their coverage probabilities tend to be closer to the nominal level (Agresti 2003). For alternative options for deriving an exact confidence interval for the odds ratio, see Sect. 2.5.2.

2.1.8 *Example 2.2*

Consider the following hypothetical data set, where 20 patients are cross-classified according to treatment and therapy outcome.

Group	Success	Failure	Total
A	10	3	13
B	2	5	7
Total	12	8	20

For given $n_{1+} = 13$, $n_{+1} = 12$ and $n = 20$, $N_{11} \sim \mathscr{H}(20, 13, 12)$. All possible values for N_{11} along with the corresponding probabilities $p(t) = \mathrm{P}(N_{11} = t)$ are given below.

t	5	6	7	8	9	10	11	12
$p(t)$	0.0102	0.0954	0.2861	0.3576	0.1987	0.0477	0.0043	0.0001

In this case, $t_{\text{obs}} = 10$ and $p(10) = 0.0477$. Testing $H_0: \theta = 1$, we get the following p-values:

H_1	p-value
$\theta > 1$	$P^+ = \mathrm{P}(N_{11} \geq t_0) = p(10) + p(11) + p(12) = 0.0521$ $\text{mid}{-}P^+ = \frac{1}{2}p(10) + p(11) + p(12) = 0.0283$
$\theta < 1$	$P^- = \mathrm{P}(N_{11} \leq t_0) = p(5) + \ldots + p(10) = 0.9956$ $\text{mid}{-}P^- = p(5) + \ldots + p(9) + \frac{1}{2}p(10) = 0.9717$
$\theta \neq 1$	$P_\ell = \sum_{t:p(t)\leq p(10)} = p(5) + p(10) + p(11) + p(12) = 0.0623$ $P_{tw} = 2P^+ = 0.1042$ $P_{CH} = P^+ + p^* = P^+ + p(5) = 0.0623$

Note that $p^* = p(5)$, since the next left tail probability would be $p(5) + p(6) = 0.1057 > p(10)$. Thus, for this data set we have $P_{CH} = P_\ell$.

2.1.8.1 Example 2.2 in R

In R, the Fisher's exact test along with the exact $(1-\alpha)100\%$ confidence interval is computed by `fisher.test()`. For our example,

```
> example <- matrix(c(10,2,3,5), 2, 2)
> fisher.test(example)
```

leads to the output

```
Fisher's Exact Test for Count Data
data: example
p-value = 0.06233
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.7406562   117.2637532
sample estimates:
odds ratio
7.320765
```

Command

```
> fisher.test(example, alternative = "greater")
```

would provide the Fisher's exact test for the one-sided alternative $H_1: \theta > 1$. The two-sided p-value adopted in `fisher.test()` is (2.24) and the provided confidence interval, based on the acceptance region and this p-value, can be inconsistent with the test (Fay 2010a). To observe this, replace in the data set above the first column by quite larger frequencies, setting, for example,

```
> exampl2 <- matrix(c(127,45,3,5),2,2)
```

Then, function `fisher.test()` gives the following output

```
> fisher.test(exampl2)
```

```
Fisher's Exact Test for Count Data
data: exampl2
p-value = 0.03876
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.8661222   31.1888976
sample estimates:
odds ratio
4.655061
```

Note, that although the null hypothesis of $\theta = 1$ is rejected at $\alpha = 0.05$, value 1 belongs to the 95% CI for θ. Fay (2010b) constructed algorithms that match the p-values of testing and CI, implemented in R's package `exact2x2`. Thus, the CI based on the inversion of the two-sided test Fisher's exact test but with the p-value defined by (2.25) is derived as

```
> exact2x2(exampl2, tsmethod = "central")
```

```
Central Fisher's Exact Test
data: exampl2
p-value = 0.07752
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.8661222   31.1888976
sample estimates:
odds ratio
4.655061
```

An alternative option of `exact2x2` is to construct Blaker's confidence interval, using the p-value (2.26). For this example

```
> exact2x2(exampl2, tsmethod = "blaker")
```

```
Blaker's Exact Test
data: exampl2
p-value = 0.03876
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
1.0919   23.1823
sample estimates:
odds ratio
4.655061
```

Alternatively, exact tests for the odds ratio and associated confidence intervals can be computed in `R` by packages `propCIs` and `pairwise.CI`.

2.2 Analyzing $I \times J$ Tables

2.2.1 Possible Sampling Schemes

Let X and Y be two categorical variables of $I \geq 2$ and $J \geq 2$ levels, respectively, that are cross-classified in a $I \times J$ contingency table and n_{ij} be the observed frequency for cell (i, j), $i = 1 \ldots, I$, $j = 1, \ldots, J$. The table will be of the following form.

n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	n_{1+}
n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	n_{2+}
$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdots$	$\cdot$	
n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	n_{i+}
$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdots$	$\cdot$	
n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	n_{I+}
n_{+1}	n_{+2}	$\cdots$	n_{+j}	$\cdots$	n_{+J}	n

Regarding the sample size and according to the study design, there are three options: (a) the total sample size n is fixed, (b) one set of marginals is fixed, without

loss of generality assume the row marginals $(n_{1+}, n_{2+}, \ldots n_{I+})$ are fixed, or (c) no restriction is imposed on the sample size. The associated sampling proportions are denoted by $p_{ij} = \frac{n_{ij}}{n}$.

Case (a) corresponds to the situation where a sample of prespecified sample size n is collected and its items are cross-classified with respect to the categorical characteristics X and Y. The underlying sampling scheme is *multinomial* and interest lies on testing *independence* of these characteristics. If N_{ij} is the random number of observations in cell (i, j) with $\sum_{i,j} N_{ij} = n$, then

$$(N_{11}, N_{12}, \ldots, N_{I.J-1}) \sim \mathscr{M}\left(n, (\pi_{11}, \pi_{12}, \ldots, \pi_{I.J-1})\right) \tag{2.27}$$

where $(\pi_{11}, \pi_{12}, \ldots, \pi_{I.J-1})^{\mathrm{T}}$ is the $(IJ-1) \times 1$ vector of cell probabilities, expanded by rows. The probabilities matrix $\boldsymbol{\pi} = (\pi_{ij})_{I \times J}$, with $\sum_{i,j} \pi_{ij} = 1$, is the *joint distribution* of (X, Y). The likelihood function under (2.27) is

$$L(n_{11}, \ldots, n_{IJ}) = \frac{n!}{\prod_{i,j}(n_{ij}!)} \prod_{i,j} \pi_{ij}^{n_{ij}} \tag{2.28}$$

Situation (b) arises when samples from I independent populations and of prespecified sizes $n_{1+}, \ldots n_{I+}$ are available. That is, a categorical characteristic (in Y) is recorded for I independent samples aiming to test the *homogeneity* of the characteristic's distribution across the samples. Thus, an independent multinomial distribution is considered for each row i

$$(N_{i1}, N_{12}, \ldots, N_{i.J-1}) \sim \mathscr{M}\left(n_{i+},\ (\pi^*_{i1}, \pi^*_{i2}, \ldots, \pi^*_{i.J-1})\right)\ ,\quad i = 1, \ldots, I\ , \tag{2.29}$$

with $\boldsymbol{\pi}_i^{*\mathrm{T}} = (\pi^*_{i1}, \pi^*_{i2}, \ldots, \pi^*_{iJ})$ the probability vector for the ith population and $\sum_j \pi^*_{ij} = 1$, for $i = 1, \ldots, I$. This sampling scheme is the *product multinomial* and the corresponding likelihood function is

$$L(n_{11}, \ldots, n_{IJ}) = \prod_{i=1}^{I} L(n_{i1}, n_{12}, \ldots, n_{i.J}) = \prod_{i=1}^{I} \left(\frac{n_{i+}!}{\prod_{j=1}^{J} n_{ij}!} \prod_{j=1}^{J} (\pi^*_{ij})^{n_{ij}} \right)$$

Since the row marginals $(n_{1+}, n_{2+}, \ldots n_{I+})$ are fixed, in the light of property (1.5), the I independent multinomials can be derived from a multinomial of the type (2.27) with $n = \sum_i n_{i+}$, fixed row marginal probabilities $\pi_{i+} = \frac{n_{i+}}{n}$ $(i = 1, \ldots, I)$, and $\pi^*_{ij} = \pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}}$. Thus, the above likelihood function equals

$$L(n_{11}, \ldots, n_{IJ}) = \frac{n^n \prod_{i=1}^{I} \left(n_{i+}^{-n_{i+}} n_{i+}! \right)}{\prod_{i,j}(n_{ij}!)} \prod_{i,j} \pi_{ij}^{n_{ij}} \tag{2.30}$$

Note that for both (2.28) and (2.30), it holds

$$L(n_{11},\ldots,n_{IJ}) \propto \prod_{i,j} \pi_{ij}^{n_{ij}} \tag{2.31}$$

and thus they are inferentially equivalent.

Finally, under (c) the concept is as in (a) with the difference that the total sample size is random. Randomness of n arises because by design a different aspect is constrained than sample size. Usually the design is time constrained. For example, we record the monthly arrivals in a clinic and cross-classify them according to two categorical characteristics X and Y. Then, if m_{ij} is the expected frequencies for the combination $(X = i,\ Y = j)$,

$$(N_{ij}) \sim \mathscr{P}(m_{ij})\ , \quad i = 1,\ldots,I,\ \ j = 1,\ldots,J\ , \tag{2.32}$$

and this sampling scheme is known as *independent Poisson*. The likelihood function for case (c) is

$$L(n_{11},\ldots,n_{IJ}) = \prod_{i,j} \frac{e^{-m_{ij}} m_{ij}^{n_{ij}}}{n_{ij}!}$$

Upon observing the sample, we can condition on the total sample size n. Then, applying property (1.7), the likelihood function conditional on $\sum_{ij} m_{ij} = n$ becomes

$$L(n_{11},\ldots,n_{IJ}\ |n) = \frac{n!}{\prod_{i,j}(n_{ij}!)} \prod_{i,j} \left(\frac{m_{ij}}{n}\right)^{n_{ij}} \tag{2.33}$$

and by setting $\pi_{ij} = \frac{m_{ij}}{n}$, this is equivalent to (2.28).

Overall, testing independence is not influenced by the underlying sampling scheme. Furthermore, testing homogeneity of independent samples in terms of a characteristic is equivalent to testing independence between the variable of the characteristic and the variable defining the samples. Thus, all hypothesis testing problems related to the setups discussed here are treated unified under the test of independence, presented in the next subsection.

2.2.2 *Test of Independence*

The hypothesis of independence introduced and discussed for 2×2 tables in Sect. 2.1.1 extends directly to the general $I \times J$ contingency table. The variables X and Y are independent if

$$P(X = i,\ Y = j) = P(X = i)P(Y = j)\ , \quad i = 1,\ldots,I,\ \ j = 1,\ldots,J\ .$$

The distribution of X, ignoring the level of Y, is defined by the vector of the row marginal probabilities $\pi_r = (\pi_{1+}, \pi_{2+}, \ldots, \pi_{I+})$ and is known as the *row marginal distribution*. Analogously, the *column marginal distribution* is defined for Y by $\pi_c = (\pi_{+1}, \pi_{+2}, \ldots, \pi_{+J})$. Thus, variables X and Y are independent if the following hypothesis holds

$$H_0 \; : \; \pi_{ij} = \pi_{i+}\pi_{+j} \; , \quad i = 1, \ldots, I, \quad j = 1, \ldots, J \tag{2.34}$$

For the multinomial sampling scheme (2.27), the expected under (2.34) frequencies are $m_{ij} = n\pi_{i+}\pi_{+j}$ and their MLEs $\hat{m}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$. Further, by property (1.5), the distribution for the row marginals is

$$(N_{1+}, N_{2+}, \ldots, N_{I-1+}) \sim \mathscr{M}(n, \pi_r)$$

and the ML estimates of the row marginal probabilities are $\hat{\pi}_{i+} = p_{i+}$, $i = 1, \ldots, I$, with the analogous result holding also for the column marginals ($\hat{\pi}_{+j} = p_{+j}$, $j = 1, \ldots, J$). The ML estimates of the expected cell frequencies under H_0 are thus

$$\hat{m}_{ij} = np_{i+}p_{+j} = \frac{n_{i+}n_{+j}}{n} \; , \quad i = 1, \ldots, I, \quad j = 1, \ldots, J \; . \tag{2.35}$$

Hypothesis H_0 will be tested asymptotically by Pearson's X^2. Since the row (column) marginal probabilities sum to one, only $I-1$ ($J-1$) of them are unknown, and the number of parameters to be estimated under H_0 is $(I-1)+(J-1)$. The associated df are by (1.16) equal to $df = IJ-(I-1)-(J-1)-1 = (I-1)(J-1)$. Thus, Pearson's X^2 statistic (1.15) for testing (2.34) becomes

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \tag{2.36}$$

The asymptotic distribution for (2.36) under H_0 is $\mathscr{X}^2_{(I-1)(J-1)}$. Alternatively, the asymptotic equivalent LR statistic (1.17) can be applied, here expressed as

$$G^2 = 2\sum_{i,j} n_{ij} \log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right) \; . \tag{2.37}$$

2.2.3 Example 2.3

The test of independence will be illustrated with a 2×3 contingency table, formed from the General Social Survey basis for year 2008 (GSS2008), cross-classifying responders by gender and confidence in banks and financial institutions. The data are given in Table 2.2. The ML estimates of the expected cell frequencies under the hypothesis of independence (2.34) are provided in brackets.

Table 2.2 Respondents' cross-classification by gender and their confidence in banks and financial institutions (*GSS* 2008)

	Confidence in banks			
Gender	Great deal	Only some	Hardly any	Total
Male	98 (119.62)	363 (366.58)	153 (127.80)	614
Female	165 (143.38)	443 (439.42)	128 (153.20)	736
Total	263	806	281	1,350

In parentheses are give the maximum likelihood estimates under the hypothesis of independence

Test statistics (2.36) and (2.37) are asymptotically equivalent and $\mathcal{X}_2^2$ distributed. For this example, their observed values are $X^2 = 16.34$ and $G^2 = 16.40$, respectively, that are highly significant with both corresponding p-values < 0.0003. Hence, H_0 of independence is rejected and we conclude that the level of confidence in banks and financial institutions depends on the gender of the responder. With respect to the conditional row distributions, we could say that the distribution of the confidence level is nonhomogeneous for men and women. However, just the confirmation of the speculation that confidence in banks and gender are dependent is not enough. We would like to describe this dependence and investigate its direction. For this, we need to compare the estimates of the expected under independence cell frequencies to the observed frequencies. We can observe that men feel lower confidence for banks than expected under independence while women higher. The cells that are farther apart from independence are (1,3) and (2,3), in opposite directions, with $n_{13} - \hat{m}_{13} = -(n_{23} - \hat{m}_{23}) = 25.2$ followed by the set (1,1) and (2,1) with $-(n_{11} - \hat{m}_{11}) = n_{21} - \hat{m}_{21} = 21.2$. How can we evaluate the contribution of each cell to the deviance from independence? Is the simple difference $n_{ij} - \hat{m}_{ij}$ appropriate for such type of conclusions? These questions will be addressed in the next subsection.

In `R`, the analysis above is carried out by `chisq.test()`. As explained in Sect. 2.1.2, the data are read by `chisq.test()` in a matrix form. Thus, data in Table 2.2 is entered in matrix `confinan` as

```
> confinan <- matrix(c(98,363,153,165,443,128),byrow=T,ncol=3)
```

while labels can be added to the classification categories of the table

```
> dimnames(confinan) <- list(Gender=c("males","females"),
+        Conf=c("great deal","only some","hardly any"))
```

The X^2 test of independence is then applied by

```
> chisq.test(confinan)
```

and the ML estimates of the expected cell frequencies under independence are derived by

```
> chisq.test(confinan)$expected
```

To see more about `chisq.test()` and its possibilities for analysis and output, one can consult `R`'s help command, `help(chisq.test)`. The G^2 test of independence is achieved by the `> G2()` function of the web appendix (see Sect. A.3.2) as follows:

```
> G2(depsmok)
```

2.2.4 Analysis of Residuals

Upon rejecting the H_0 of independence, or more general any H_0, interest lies on detecting parts of the contingency table (single cells or whole regions) that contribute more in the value of the goodness-of-fit statistic, i.e., parts of the table that are mainly responsible for the rejection of H_0. The natural quantities to observe for this are the differences between the observed and the estimates of the expected under H_0 cell frequencies, called *residuals*

$$e_{ij} = n_{ij} - \hat{m}_{ij} \ , \quad i = 1,\ldots,I, \quad j = 1,\ldots,J \ . \tag{2.38}$$

The residuals are examined in terms of sign and magnitude. The detection of a systematic structure of their signs is of special interpretational interest. However, the evaluation of the importance of the contribution of a particular cell to the deviation from independence, when based on these residuals, can be misleading. More appropriate are the residuals that standardize (2.38) by dividing them by their s.e.

$$e^*_{ij} = \frac{e_{ij}}{\sqrt{\text{Var}(e_{ij})}} = \frac{n_{ij} - \hat{m}_{ij}}{\sqrt{\text{Var}(\hat{m}_{ij})}} \ , \quad i = 1,\ldots,I, \quad j = 1,\ldots,J \ , \tag{2.39}$$

and are under the H_0 of independence, $e^s_{ij} \sim \mathcal{N}(0,1)$, asymptotically. For Poisson sampling, $\text{Var}(\hat{m}_{ij}) = m_{ij}$ and estimating $\text{Var}(\hat{m}_{ij})$ by $\hat{m}_{ij}$, the estimates of (2.39) are

$$e^P_{ij} = \hat{e}^*_{ij} = \frac{n_{ij} - \hat{m}_{ij}}{\sqrt{\hat{m}_{ij}}} \ , \quad i = 1,\ldots,I, \quad j = 1,\ldots,J \ , \tag{2.40}$$

and are called *Pearsonian residuals*, since

$$X^2 = \sum_{i,j} \left(e^P_{ij}\right)^2 \ . \tag{2.41}$$

Thus, e^P_{ij} are adequate quantities to evaluate the merit of each cell to the deviation from independence.

Under multinomial sampling, $\text{Var}(\hat{m}_{ij})$ is different than under Poisson and consequently the Pearsonian residuals (2.40) are no more asymptotic standard normal distributed. Desired properties for a residual type would be that it is invariant of the sampling scheme and asymptotic standard normal distributed. The Pearsonian residuals are asymptotically normal distributed $e^P_{ij} \sim \mathcal{N}(0, v_{ij})$ but $v_{ij} \neq 1$, due to the approximation of the variance $\text{Var}(\hat{m}_{ij})$ under H_0 by estimating it. Haberman (1973b) proved that under independence and for multinomial sampling, the asymptotic variances of the expected cell frequencies are $v_{ij} = v_{ij}(\pi) = (1-\pi_{i+})(1-\pi_{+j})$, as $n \to \infty$. He suggested to estimate asymptotic variances by their ML estimates

$$\hat{v}_{ij} = (1 - \frac{n_{i+}}{n})(1 - \frac{n_{+j}}{n}) \ , \quad i = 1,\ldots,I, \ \ j = 1,\ldots,J \ ,$$

introduced the *standardized residuals*

$$e_{ij}^s = \frac{e_{ij}^P}{\sqrt{\hat{v}_{ij}}} = \frac{e_{ij}}{\sqrt{\hat{m}_{ij}\hat{v}_{ij}}} \ , \quad i = 1,\ldots,I, \ \ j = 1,\ldots,J \ , \tag{2.42}$$

and proved that they are asymptotically standard normal distributed. Standardized residuals (Haberman (1973b) called them *adjusted* residuals) are also common for both sampling schemes, multinomial and independent Poisson. The standardized residuals e_{ij}^s are thus more informative and preferable for reporting and analyzing. Cells can be characterized as significantly influential against H_0 at level $a = 0.05$, for example, if $|e_{ij}^s| > z_{0.025} = 1.96$. Since $v_{ij} < 1$, for all $i, \ j$, it is always $|e_{ij}^P| < |e_{ij}^s|$.

For Example 2.3, the Pearsonian residuals (2.40) are obtained in R by

```
> chisq.test(confinan)$residuals
```

	Conf		
Gender	great deal	only some	hardly any
males	-1.976451	-0.1870200	2.228841
females	1.805225	0.1708179	-2.035749

while the standardized residuals (2.42) by

```
> chisq.test(confinan)$stdres
```

	Conf		
Gender	great deal	only some	hardly any
males	-2.983089	-0.3990097	3.392228
females	2.983089	0.3990097	-3.392228

By the Pearsonian residuals we conclude that the deviation from independence is no more symmetric for the set of cells in column 3 neither in column 1. The cells in decreasing significance order of deviation from H_0 are (1,3), (2,3), (1,1), and (2,1). Thus, the level of confidence in banks is significantly different for men and women. The major contribution to deviation from independence is due to the "nonconfidence" category with the men being highly non-confident while the women are less non-confident than under independence. The next significant category is that of confidence, for which women show higher confidence than under independence while men lower. Finally the partial confidence category does not differ significantly for men and women.

Similar to the Pearsonian residuals, the *deviance residuals* are defined by the cell components of the G^2-statistic. They are equal to

$$e_{ij}^d = \text{sign}(n_{ij} - \hat{m}_{ij}) \cdot \left[2n_{ij} \log(\frac{n_{ij}}{\hat{m}_{ij}}) \right]^{1/2} \ , \quad i = 1,\ldots,I, \ \ j = 1,\ldots,J \ . \tag{2.43}$$

with $G^2 = \sum_{i,j} \left(e_{ij}^d \right)^2$.

The residuals discussed above in the context of the hypothesis of independence are defined and analyzed in the same manner for any other hypothesis H_0, provided the $\hat{m}_{ij}$'s involved are the estimates of the expected cell frequencies under the assumed H_0. Furthermore, they are defined analogously for testing hypothesis on multi-way contingency tables.

The only residual options of `chisq.test()` are the Pearsonian and the standardized. The deviance residuals are provided in the log-linear models framework and we shall revisit the example for this in Sect. 4.2.2.

The analysis of a contingency table is completed by visualizing the residuals graphically. For large n, the normal probability plots for the ordered standardized residuals are a standard companion while Santner and Duffy (1989) suggest also plots of the residuals vs. the row or column category indexes. Informative are also graphical displays presented in Sect. 2.4 and illustrated for Example 2.2 in Fig. 5.1 (right).

2.2.5 *Odds Ratios for $I \times J$ Tables*

The odds ratio θ is a powerful measure of association for a 2×2 table of high interpretational importance. It is the basis for detecting association structures also in $I \times J$ tables. For this, a decomposition of the $I \times J$ table to a set of 2×2 tables is needed. In general, for an $I \times J$ table, a set of $(I-1)(J-1)$ basic 2×2 tables is formed and the corresponding odds ratios describe the underlying associations. However this decomposition is not unique. Depending upon the type of the classification variables but also on the inference problem under consideration, there are alternative options, leading to different types of odds ratios.

For nominal classification variables this set of basic 2×2 tables is defined in terms of a reference category, usually the cell (I,J). Then the 2×2 tables formed have in their upper diagonal cell the (i,j) cell of the initial table, for $i = 1,\ldots,I-1$, $j = 1,\ldots,J-1$, and in the lower diagonal cell always the reference cell (I,J). The non-diagonal cells are the cells of the initial table that share one classification variable index with each diagonal cell, i.e., they are the cells (i,J) and (I,j). Thus, the *nominal odds ratios* are defined as

$$\theta_{ij}^{IJ} = \frac{\pi_{ij}\pi_{IJ}}{\pi_{Ij}\pi_{iJ}}\,, \quad i = 1,\ldots,I-1, \quad j = 1,\ldots,J-1\,. \tag{2.44}$$

The diagonal cells are indicated in the sub- and superscript of the notation. Of course, any cell (r,c) of the table could serve as reference category and the nominal odds ratios are then defined analogously for all $i \neq r$, $j \neq c$.

Different types of odds ratios are adequate for ordinal variables. A fixed reference cell is not meaningful and a more natural choice is either to compare each level of the ordinal classification variable to the immediate next or for each level, to oppose the events of being up to it or above it. The first option refers locally to

just two successive categories while the second engages cumulatively all categories. Adoption of the same type (local or cumulative) for both classification variables or different for each of them leads to the three more characteristic odds ratios for ordinal variables. Consideration of the same option for both classification variables treats them symmetrically while otherwise not. The nonsymmetric case is adequate for problems with a response variable, for which the cumulative option is adopted. Of course, the odds ratios treating both variables symmetrically do also apply for response variables.

When both classification variables are treated locally, the 2×2 tables are formed by two successive rows i and $i+1$, for $i = 1, \ldots, I-1$, and two successive columns j and $j+1$, for $j = 1, \ldots, J-1$. This way there are formed $(I-1)(J-1)$ local tables and the corresponding odds ratios are the *local odds ratios*

$$\theta_{ij}^{L} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i+1,j}\pi_{i,j+1}}\ , \quad i = 1, \ldots, I-1, \ \ j = 1, \ldots, J-1\ . \tag{2.45}$$

This minimal set is sufficient to describe association and derive odds ratios for any other 2×2 table formed by non-successive rows or columns. A 2×2 subtable is determined by its diagonal cells. Once they are chosen, the non-diagonal cells are specified by combining the levels of the classification variables of the diagonal ones. Thus, assuming that both classification variables are in increasing order, the odds ratio for comparing cell (i, j) to the cell that is k levels higher for the row and ℓ levels higher for the column classification variable, i.e., the $(i+k, j+\ell)$ cell, refers to the subtable

	j	$j+\ell$
i		
$i+k$		

and is derived by the local odds ratios as

$$\theta_{ij}^{i+k,j+\ell} = \frac{\pi_{ij}\pi_{i+k,j+\ell}}{\pi_{i+k,j}\pi_{i,j+\ell}} = \prod_{\rho=0}^{k-1}\prod_{\xi=0}^{\ell-1}\theta_{i+\rho,j+\xi}^{L}\ , \ 1 \le k \le I-i, \ 1 \le \ell \le J-j\ . \tag{2.46}$$

For $k = \ell = 1$, (2.46) is the local odds ratio, i.e., $\theta_{ij}^{L} = \theta_{ij}^{i+1,j+1}$, while for $k = I-i$ and $\ell = J-j$, (2.46) becomes the nominal odds ratio (2.44).

For nominal and local odds ratios, the minimal set of 2×2 tables is a set of subtables of the initial table. If the cumulative option is adopted for at least one of the classification variables for defining the odds ratios, then the associated 2×2 tables are no more subtables. When both classification variables are treated cumulatively, then the 2×2 tables are collapsed versions of the $I \times J$ table, produced by transforming the classification variables to binary with cut points i ($i = 1, \ldots, I-1$) and j ($j = 1, \ldots, J-1$) for rows and columns, respectively. This way, all cells of the initial table participate in the formulation of each 2×2 table and association is faced globally. The associated odds ratios are the *global odds ratios*, defined by

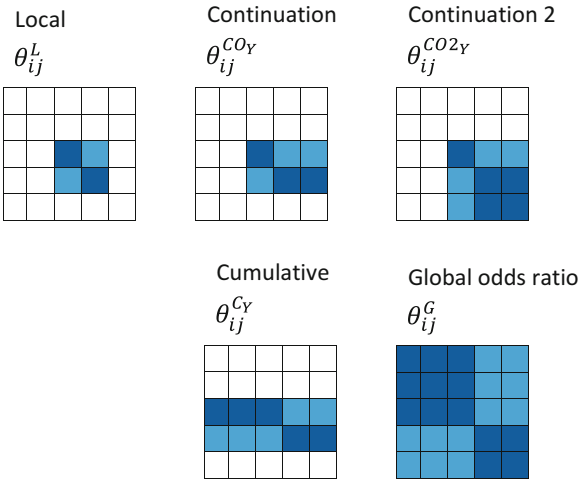


Fig. 2.1 Formulation of the generalized odds ratios for $I \times J$ contingency tables. With respect to ordinality of the row and column classification variables X and Y, odds ratios in the first, second, and third column require ordinality of none, only Y, or both X and Y, respectively

$$\theta_{ij}^{G} = \frac{\left(\sum_{l\le i}\sum_{k\le j}\pi_{lk}\right)\left(\sum_{l>i}\sum_{k>j}\pi_{lk}\right)}{\left(\sum_{l\le i}\sum_{k>j}\pi_{lk}\right)\left(\sum_{l>i}\sum_{k\le j}\pi_{lk}\right)}\,, \quad i=1,\ldots,I-1,\;\; j=1,\ldots,J-1\,, \tag{2.47}$$

and illustrated in Fig. 2.1. In the numerator is the product of the sums of cells in the dark shadowed rectangles while in the denominator the product of the sums in the light shadowed rectangles.

Odds ratios θ_{ij}^{L} and θ_{ij}^{G} refer to different types of associations and the choice between them relies on the needs of our analysis and the nature of the underlying classification variables.

Both types of odds ratios, θ_{ij}^{L} and θ_{ij}^{G}, treat both classification variables in a symmetric way (the θ_{ij}^{L}'s locally and the θ_{ij}^{G}'s cumulatively). If only one classification variable is treated cumulatively, say the columns' variable Y and the other locally, then for the formulation of the 2×2 tables only the columns of the initial table are collapsed and each of them is based on all cells of two successive rows of the table. Hence, for given i ($i = 1,\ldots,I-1$) and j ($j = 1,\ldots,J-1$), the tables constructed are of the form presented in Fig. 2.1.

The odds ratios applied on these tables are the *cumulative odds ratios*, defined by

$$\theta_{ij}^{C_Y} = \frac{\left(\sum_{k\le j}\pi_{ik}\right)\left(\sum_{k>j}\pi_{i+1,k}\right)}{\left(\sum_{k>j}\pi_{ik}\right)\left(\sum_{k\le j}\pi_{i+1,k}\right)}\,, \quad i=1,\ldots,I-1,\;\; j=1,\ldots,J-1\,. \tag{2.48}$$

The cumulative odds ratio $\theta_{ij}^{C_X}$ is cumulative with respect to the rows, applies on successive columns j and $j+1$, and is defined analogously.

Cumulative and global odds ratios make sense for ordinal classification variables. They are also meaningful for tables with one ordinal classification variable and one

binary. For $2 \times J$ tables, the global and cumulative odds ratios, (2.47) and (2.48), coincide.

Less popular are the continuation odds ratios

$$\theta_{ij}^{CO_Y} = \frac{\pi_{j|i}/(\sum_{k>j} \pi_{k|i})}{\pi_{j|i+1}/(\sum_{k>j} \pi_{k|i+1})} \tag{2.49}$$

and the continuation type 2 odds ratios

$$\theta_{ij}^{CO2_Y} = \frac{\pi_{j|i}/(\sum_{k>j} \pi_{k|i})}{\sum_{\ell>i} \pi_{j|\ell}/(\sum_{k>j,\ell>i} \pi_{k|\ell})} \,. \tag{2.50}$$

Odds ratios (2.49) and (2.50) consider Y to be the response variable. Analogously are defined the $\theta_{ij}^{CO_X}$ and $\theta_{ij}^{CO2_X}$, when X is the response.

For the generalized odds ratios presented above, the ordinality of the classification variables is required only whenever a classification variable is treated cumulatively. Thus, the local odds ratios are also appropriate for nominal variables. In Fig. 2.1 is illustrated the formulation of the generalized odds ratios. They are organized in columns according to requirements on ordinality of the classification variables. In the first column is only the θ_{ij}^L that can be applied also when both X and Y are nominal. In the second column ordinality is required only for the column classification variable Y while in the third for both X and Y.

We have seen that an $I \times J$ probability table $\boldsymbol{\pi} = (\pi_{ij})$ with positive entries determines uniquely the corresponding $(I-1) \times (J-1)$ table of local odds ratios or any other type of generalized odds ratios. On the other hand, an $(I-1) \times (J-1)$ table of positive and finite local odds ratios corresponds to more than one probability tables, since property (2.13) for θ of the 2×2 table generalizes also to the local odds ratios of the $I \times J$ table. Hence, given an $(I-1) \times (J-1)$ table of positive and finite local odds ratios $\boldsymbol{\theta}^L = (\theta_{ij}^L)$, a corresponding $I \times J$ probability table $\boldsymbol{\pi} = (\pi_{ij})$ is derived by

$$\pi_{ij} = \frac{\alpha_i \beta_j \theta_{11}^{ij}}{\sum_{i=1}^{I} \sum_{j=1}^{J} \alpha_i \beta_j \theta_{11}^{ij}} \,, \quad i = 1, \ldots, I, \;\; j = 1, \ldots, J \,, \tag{2.51}$$

where θ_{11}^{ij}, for $i, j > 1$, are defined by (2.46), $\theta_{11}^{ij} = 1$ for $i = 1$ or $j = 1$, and α_i, β_j positive parameters. It can be proved that the probability table $\boldsymbol{\pi}$ becomes unique once its row and column marginals, $\boldsymbol{\pi}_r^{\mathrm{T}} = (\pi_{1+}, \ldots, \pi_{I+})$ and $\boldsymbol{\pi}_c^{\mathrm{T}} = (\pi_{+1}, \ldots, \pi_{+J})$, are fixed, which uniquely specify the parameters α_i, $i = 1, \ldots, I$, and β_j, $j = 1, \ldots, J$, respectively. In other words, θ, $\boldsymbol{\pi}_r$, and $\boldsymbol{\pi}_c$ determine uniquely the table of joined probabilities $\boldsymbol{\pi}$, a result that holds also when θ is replaced by any other minimal set of odds ratios.

In analogy to the simple 2×2 table, where independence was equivalent to $\theta = 1$, it can be verified that for an $I \times J$ contingency table, the independence hypothesis (2.34) is equivalent to the hypothesis that all odds ratios in a minimal set are equal to 1. Thus, in terms of local odds ratios, (2.34) is equivalent to

$$\theta_{ij}^L = 1 \,, \quad i = 1, \ldots, I-1, \;\; j = 1, \ldots, J-1 \,. \tag{2.52}$$

Independence could equivalently be expressed by (2.52) for any other type of minimal set of odds ratios. The hypothesis formulation for independence is simpler for the odds ratio than for the expected probabilities parameterization, since (2.52) assigns fixed values to the parameters while under (2.34) parameters have to be estimated. Also the df of independence are directly understood by (2.52).

In general, beyond independence, any hypothesis considered for the structure of an $I \times J$ probability table $\pi = (\pi_{ij})$ can equivalently be expressed in terms of the corresponding $(I-1) \times (J-1)$ table of local odds ratios, as we shall see in Chaps. 6 and 8. In view of the discussion above, when a hypothesis H_0 is defined in terms of odds ratios, the row and column marginal probabilities are required for the expected under H_0 cell probabilities to be fully determined.

The odds ratios presented above refer to the population under consideration and are unknown. Upon observing a sample, the sample local odds ratio is

$$\hat{\theta}_{ij}^{L}(\mathbf{n}) = \frac{n_{ij} n_{i+1,j+1}}{n_{i+1,j} n_{i,j+1}} , \quad i = 1, \ldots, I-1, \quad j = 1, \ldots, J-1 . \tag{2.53}$$

The ML estimate of θ_{ij}^{L} of θ_{ij} under a hypothesis H_0 is provided by (2.53) with the observed frequencies (n_{ij}) being replaced by the ML estimates of the expected under H_0 frequencies $(\hat{m}_{ij})$. The sample odds ratios $\hat{\theta}_{ij}^{IJ}$, $\hat{\theta}_{ij}^{G}$, $\hat{\theta}_{ij}^{C_Y}$, $\hat{\theta}_{ij}^{CO_Y}$, and $\hat{\theta}_{ij}^{CO2_Y}$ are defined analogously.

In `R`, the various sets of generalized odds ratios are easier computed in log-scale and working with matrices. It can easily be proved that the set of the sample log local odds ratios is derived in a $(I-1)(J-1) \times 1$ vector $\log \mathbf{L}$ (expanded by rows) as

$$\log \mathbf{L} = \mathbf{C}_L \cdot \log \mathbf{n} , \tag{2.54}$$

where $\mathbf{n}$ is the $IJ \times 1$ vector of the observed frequencies (given by rows) and $\mathbf{C}_L$ is an appropriate design matrix of size $(I-1)(J-1) \times IJ$. Analogously, the global, cumulative, continuation, and continuation of type 2 odds ratios, in log-scale, are provided in vector form by

$$\log \mathbf{O}_i = \mathbf{C}_i \cdot \log(\mathbf{M}_i \cdot \mathbf{n}) , \quad i = 1, \ldots, 4 , \tag{2.55}$$

where $\mathbf{C}_i$ and $\mathbf{M}_i$ are appropriate matrices. The `R` functions `local.odds.DM()`, `global.odds.DM()`, `cum.odds.DM()`, and `cont.odds.DM()`, provided in the web appendix (see Sect. A.3.2), produce the design matrices used in (2.54) and (2.55), for deriving the various sets of generalized odds ratios for any choice of I and J. The use of these functions is illustrated in the example below.

Table 2.3 Respondents' cross-classification by educational level and their opinion about national spending for welfare (*GSS* 2008)

	Highest degree obtained					
Welfare spending	LT high school	High school	Junior college	Bachelor	Graduate	Total
Too little	45	116	19	48	23	251
About right	40	167	33	68	41	349
Too much	47	185	34	63	26	355
Total	132	468	86	179	90	955

Table 2.4 The sample ordinal odds ratios (a) $\hat{\theta}_{ij}^{L}$, (b) $\hat{\theta}_{ij}^{G}$, and (c) $\hat{\theta}_{ij}^{X_Y}$ for the data in Table 2.3

		Highest degree obtained				
	Welfare spending	LT high school	High school	Junior college	Bachelor	Graduate
(a)	Too little	1.62	1.21	0.82	1.26	
	About right	0.94	0.93	0.90	0.68	
	Too much					
(b)	Too little	1.55	1.08	0.99	1.04	
	About right	1.08	0.84	0.78	0.66	
	Too much					
(c)	Too little	1.57	1.16	0.77	1.07	
	About right	1.18	1.00	0.83	0.75	
	Too much					

2.2.6 *Example 2.4*

Data in Table 2.3 are from the General Social Survey basis for year 2008 (GSS2008). Responders are cross-classified by their opinion on the sufficiency of the amount of national spending for welfare and their educational level, measured by the highest degree they obtained. Both classification variables are ordinal. The national spending can be considered as a response variable, thus the cumulative odds ratio is applicable. Since the response variable is in rows (X), the appropriate cumulative odds ratio is $\hat{\theta}_{ij}^{C_X}$, the cumulative on X.

For this example, the ML estimates of the local odds ratios, global odds ratios, and cumulative odds ratios are presented in Table 2.4. Indicatively, we calculate

$$\hat{\theta}_{12}^{L} = \frac{116 \cdot 33}{167 \cdot 19} = 1.21$$

$$\hat{\theta}_{12}^{G} = \frac{(45+116)(33+68+41+34+63+26)}{(40+167+47+185)(19+48+23)} = 1.08$$

$$\hat{\theta}_{12}^{X_Y} = \frac{116(33+34)}{(167+185)19} = 1.16$$

This means that the odds of believing that the welfare spending is about right than too little is 1.21 times higher for junior college than high school graduates. Similarly, the odds of spending being about right or above than too little is 1.08 times higher for responders with education higher than high school than up to high school. Finally, the odds of spending being about right or above than too little is 1.16 times higher for junior college than high school graduates.

For this data set, the R function for producing the 2×3 tables of the local odds ratios is implemented as follows:

```
freq<-c(45,116,19,48,23,40,167,33,68,41,47,185,34,63,26)
NI <- 3;  NJ <- 5;  C <- local.odds.DM(NI,NJ)
L.OR <- exp(t(matrix(as.vector(C%*%log(freq)), NJ-1)))
```

Analogously, the global odds ratios are derived by

```
C1 <- global.odds.DM(NI,NJ)$C ;   M1 <- global.odds.DM(NI,NJ)$M
GL.OR <- exp(t(matrix(as.vector(C1%*%log(M1%*%freq)), NJ-1)))
```

By (2.55), the cumulative, the continuation, and the continuation type 2 odds ratios are produced as the global odds ratios by replacing the set of matrices (C1, M1) by (C2, M2), (C3, M3), and (C4, M4), respectively, where

```
C2 <- cum.odds.DM(NI,NJ)$C;       M2 <- cum.odds.DM(NI,NJ)$M
C3 <- cont.odds.DM(NI,NJ,1)$C;    M3 <- cont.odds.DM(NI,NJ,1)$M
C4 <- cont.odds.DM(NI,NJ,2)$C;    M4 <- cont.odds.DM(NI,NJ,2)$M
```

Functions `cum.odds.DM()` and `cont.odds.DM()` derive the matrices required for the calculation of the odds ratios $\hat{\theta}_{ij}^{C_Y}$, $\hat{\theta}_{ij}^{CO_Y}$, and $\hat{\theta}_{ij}^{CO2_Y}$, i.e., with Y being the response variable. In case the response is in rows variable X, we only need to apply the procedure described above on the transpose of the data table.

The fact that for this example all sample odds ratios are close to 1 indicates that whatever association there is between the belief about welfare spending and the responder's educational level is very weak. Indeed, Pearson's statistic (2.36) for testing independence equals $X^2 = 10.52$ and is nonsignificant ($df = 8$, p-value = 0.2304). This example will be revisited in Sects. 2.4 and 4.2.1.

2.3 Test of Independence for Ordinal Variables

When both classification variables of a contingency table are ordinal, we are interested in the direction of the underlying association (positive or negative). The ordering information of a classification variable is captured in scores, assigned to its categories. Thus, for an $I \times J$ table let $x_1 \leq x_2 \leq \ldots \leq x_I$ and $y_1 \leq y_2 \leq \ldots \leq y_J$ be the scores assigned to the categories of the row and column classification variables, X and Y, respectively, with $x_1 < x_I$ and $y_1 < y_J$.

The structure of the underlying association is then expressed through relations among the scores. A first sensible assumption is that association exhibits a linear trend. The linear trend is measured by Pearson's correlation ρ between X and Y, defined through their categories' scores. It is easy to verify that for

marginally weighted scores, i.e., scores satisfying $\sum_{i=1}^{I} \pi_{i+} x_i = \sum_{j=1}^{J} \pi_{+j} y_j = 0$ and $\sum_{i=1}^{I} \pi_{i+} x_i^2 = \sum_{j=1}^{J} \pi_{+j} y_j^2 = 1$, the sample correlation is $r = \frac{1}{n} \sum_{i,j} x_i y_j n_{ij}$. The linear trend test (Mantel 1963) restricts interest to linearly associated classification variables and tests the significance of ρ. Thus, the testing problem is

$$H_0 : \rho = 0 \qquad \text{vs} \qquad H_1 : \rho \neq 0 \tag{2.56}$$

and the corresponding test statistic

$$M^2 = (n-1)r^2 \tag{2.57}$$

The linear trend is a strong assumption that concentrates all association information of the table in just one parameter, r, regardless of the size $I \times J$ of the table. Thus, not surprisingly, the linear trend test is a 1 *df* test. Under H_0 in (2.56) and for a random sample of large n, $M^2 \overset{H_0}{\sim} \mathcal{X}_1^2$. Consequently,

$$R = \text{sign}(r)\sqrt{M^2} \overset{H_0}{\sim} \mathcal{N}(0,1) \ ,$$

and the test statistic R can be used for testing one-sided alternatives. The values of M^2 range from 0 (independence) to $n-1$ (perfect linear association), with evidence against independence increasing in M^2.

The test remains invariant under linear transformation of the scores. Thus, important are not the scores' values themselves but the *distances* between scores of successive categories. Therefore, for a classification variable of only two categories ($I = 2$ or $J = 2$), M^2 remains invariant under any choice of two (different) scores, since there is just one distance between categories. Since for a binary variable the scores serve just as labels, the linear trend test can be applied also to $2 \times J$ tables with the binary variable nominal. In general, methods and models appropriate for ordinal contingency tables can still be applied in presence of binary nominal classification variables.

2.3.1 The Choice of Scores

Scores is a powerful tool in the analysis of ordinal contingency tables and the development of special, very informative models, as we shall see in the sequel (Chaps. 6–9). Often, it is not clear how scores should be chosen. Typically, different choices of monotone scores lead to the same results, but different scores' systems can lead to different results (Graubard and Korn 1987). There is no direct way to measure the sensitivity of an analysis on the scores used. Test results may be sensitive in the choice of scores when the margins of the table are highly unbalanced or even if some cells have considerably larger frequencies than the others. Hence, scores' assignment can be crucial.

Table 2.5 Cross-classification of response on presence of varicella complications vs. age for 170 children in Germany (Boulesteix and Strobl 2007)

	Age category (in years)			
Varicella complications	0–1	1–2	2–3	3–18
No	10	7	9	59
Yes	6	19	12	48

The most common scores used are (a) the *equally spaced* scores, appropriate for ordinal classification variables, usually set equal to the category order $(1,2,\ldots)$, (b) the *category midpoints* for interval classification variables, and (c) the *midranks*. Midranks assign to each category the mean of the ranks of its cases, when all items of the sample are ranked from 1 to n. When midrank scores are applied to both classification variables, then r is *Spearman's* coefficient. For an interval scaled classification variable with an open category (the first or the last), the midpoint score of the open category is not uniquely determined, since we have to arbitrarily assume a lower or upper limit for the scale. Of course, any other choice is possible, provided it can be justified from the knowledge about the data. When inference differs significantly for alternative scoring sets, it is important to choose scores based on nature of the data and not guided by the desired result.

Often, scores are standardized. Standardization does not affect the test, since it is a linear transformation of the initially considered set of scores. Scores and their influence in trend analysis will be clarified in the example that follows.

2.3.2 Example 2.5

We shall consider a data set on varicella disease (Boulesteix and Strobl 2007) that cross-classifies 170 children according to their age (in four categories) and their binary response about complications. Data are provided in Table 2.5. The hypothesis of independence is rejected at $\alpha = 0.05$, since $X^2 = 8.098$ and $G^2 = 8.328$ with $df = 3$ and associate p-values equal to 0.044 and 0.040, respectively.

In order to perform the linear rank test, scores need to be assigned to the row and column categories of the table. Since $I = 2$, the choice for the row scores does not influence the outcome of the test and it will be $x_1 = 1$ and $x_2 = 2$, the simplest and natural choice. The column classification variable is interval scaled, hence the adequate choice is the category midpoints. However, we shall consider the raw and the midrank scores as well, to reveal the differences and innovate the discussion. Hence, for the column scores (y_1, y_2, y_3, y_4) we consider (a) the raw scores $(1,2,3,4)$, (b) the category midpoint (0.5, 1.5, 2.5, 10.5), and (c) the midranks (8.5, 29.5, 53, 117). For the computation of the midranks the column marginals and their cumulative distribution are required. In our example, they are $(16,26,21,107)$ and $(16,42,63,170)$, respectively. Hence, there are 16

Table 2.6 Linear trend tests for Example 2.4 and for the indicated choices of scores

	r	M^2	df	p-value
(a) Raw scores	−0.085	1.223	1	0.269
(b) Category midpoints	−0.125	2.636	1	0.104
(c) Midranks	−0.112	2.130	1	0.144

p-values are based on the $\mathscr{X}_1^2$ approximation

children with ranks $\{1,2,\ldots,16\}$ in the 1st age category , 26 children with ranks $\{17,18,\ldots,42\}$ in the 2nd, etc. Thus, $y_1 = \frac{1+\ldots+16}{16}$, $y_2 = \frac{17+\ldots+42}{26}$, $y_3 = \frac{43+\ldots+63}{21}$, and $y_4 = \frac{64+\ldots+170}{107}$.

The linear trend tests for the above discussed different choices of scores are provided in Table 2.6. For this example all choices of scores do not give much evidence against H_0 but with different significance. We have argued that the appropriate choice is (b); thus, the associated p-value is 0.104. Interpretation conclusions should be drawn with caution. Acceptance of $\rho = 0$ does not imply independence. In our case, we conclude that provided there is a linear trend in the probability of varicella complications across age categories, this trend seems negative but is nonsignificant (p-value = 0.104). We do not conclude that complications are independent of age. As we shall see later on in Sect. 6.6.3, complications are age dependent but not linearly. Thus, the linear trend test is a powerful 1 df test but of restricted origin.

2.3.3 *The Linear Trend Test in* `R`

In `R` the linear trend test can be fitted by function `linear.trend()`, provided in the web appendix (see Sect. A.3.2). It requires the data in vector form (by rows), the number of rows and columns (here, $I = 2$, $J = 4$), and the row and column scores to be used in vectors. The implementation for Example 2.5 and midpoint scores follows.

```
> varicella <- c(10,7,9,59,6,19,12,48)
> x <- c(1,2) ;  y <- c(0.5, 1.5, 2.5, 10.5)
> linear.trend(varicella, 2, 4, x, y)
```

The derived output is

```
$r
[1] -0.1248894
$M2
[1] 2.635954
$p.value
[1] 0.1044693
```

Raw and midpoint scores are easily typed, midrank scores can be computed through the `midrank()` function, provided in the web appendix (see Sect. A.3.2).

This function requires the data in a vector form (by rows), the number of rows and columns and a logical parameter (*row*), controlling whether the row (*row* = T) or column scores (*row* = F) are to be computed. For our example, command

```
> y <- midrank(varicella, 2, 4, F)
```

saves in vector `y` the midrank scores for the columns of Example 2.5 in Sect. 2.3.2.

For $2 \times J$ tables formed by a binary response and an explanatory variable, such as Example 2.5 above, independence can equivalently be expressed as equality of success proportions across the levels of the explanatory variable. In this context, (2.57) tests the significance of linear trend in the success probabilities and can be fitted by `prop.trend.test()` of `R`.

2.4 Graphs for Two-way Tables

The first type of plot one thinks of to describe a two-way contingency table is a stacked barplot of the observed frequencies or proportions of the table. Furthermore, special graphs have been developed to visualize graphically the sizes of the cells of a table (observed or expected under an assumed model) and the structure of underlying associations (illustrating the residuals for the assumed model). Characteristic such special graphs are the *sieve diagram* and the more popular *mosaic plot*. For a 2×2 table, the odds ratio is visualized by the *fourfold display*.

The barplots (simple or stacked) can be constructed in the basic `graphics` package of `R`. Fourfold displays and mosaic plots can be obtained in `graphics` as well, but for the construction of graphs for categorical data, the special package `vcd` (Visualizing Categorical Data) has been developed, offering more options.

In the following subsections stacked barplots, sieve diagrams, and mosaic plots are illustrated for Examples 2.2 and 2.3. The fourfold display is derived for Example 2.1(a).

We do not present the features of packages `graphics` and `vcd` for controlling the appearance of a graph. For more on `R` graphics we refer to Murrell (2006) and for applying and programming in `vcd` to Meyer et al. (2006).

2.4.1 Barplots

In `R`, barplots are produced by `barplot()`. The input can be a vector or a matrix, resulting to a simple or stacked barplot, respectively. In case of matrix input, the column categories define the bars while the row categories form the stacked levels.

We shall illustrate the barplots for Example 2.2 on the gender by confidence to banks and financial institutions cross-classification. The corresponding data table is in `R` matrix `confinan`, defined in Sect. 2.2.3 while labels have also been assigned to the row and column classification categories. The barplot in terms of proportions is then derived by applying `barplot()` on the table of proportions

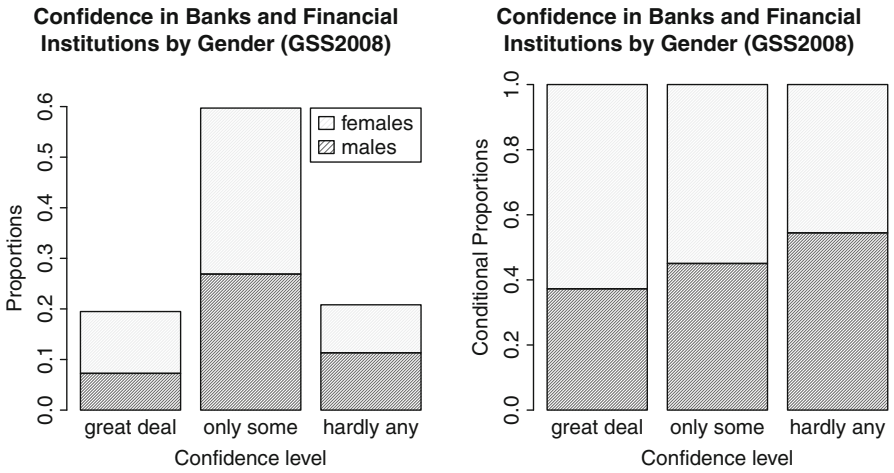


Fig. 2.2 Barplot of the observed proportions (*left*) and the conditional proportions (*right*) for data of Table 2.2

`prop.table(confinan)` as

```
> barplot(prop.table(confinan), density=30, legend.text=T, main=
+     "Confidence in Banks and Financial Institutions by Gender
+     (GSS2008)", xlab="Confidence level", ylab="Proportions",
+     ylim=c(0,0.65))
```

and is to be seen in Fig. 2.2 (left).

The role of gender in confidence to banks is better visualized by the barplot of the conditional column proportions of the table. This barplot is obtained by the command above, replacing the matrix of proportions by `prop.table(confinan,2)`, the matrix of conditional column proportions, and changing the label for axis `y` accordingly. The conditional barplot is provided in Fig. 2.2 (right). Observing Fig. 2.2 (right), we see that the gender analogy is not fixed within confidence categories, with the proportion of men growing as we move to categories of less confidence. This visualizes the dependence of confidence to banks on gender with women being less suspicious.

The required argument by `barplot()` is only the table to be plotted. The remaining arguments process the appearance of the barplot, like defining the shading of the sub-bars (`density`), adding labels to the row categories that are stacked in the bars (`legend.text=TRUE`), adding labels to the figure (`main`) and its axes (`xlab`, `ylab`), or specifying the limits of the axes (`xlim`, `ylim`).

Analogously, for Example 2.4, the barplot and the conditional barplot of the GSS 2008 respondents' opinion on national welfare pending are presented in Fig. 2.3.

In Sect. 2.2.6, the data (Table 2.3) were entered in a vector form (expanded by rows)

```
> freq <- c(45,116,19,48,23,40,167,33,68,41,47,185,34,63,26)
```

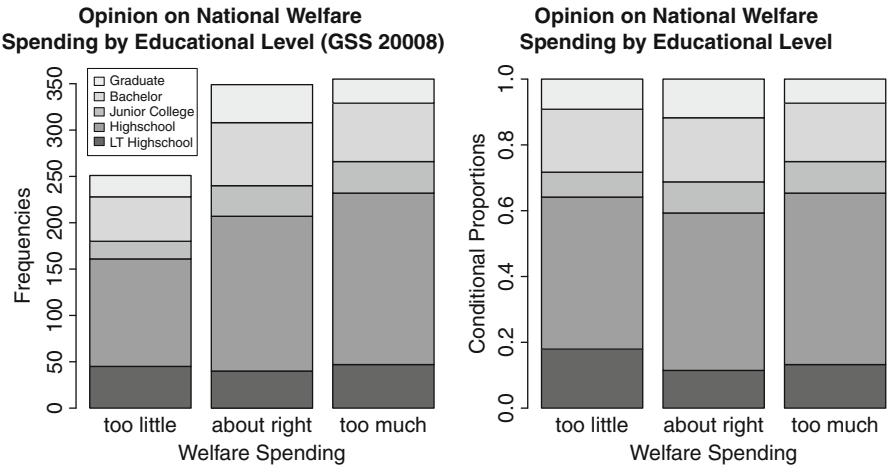


Fig. 2.3 Barplot of the observed proportions (*left*) and the conditional proportions (*right*) for data of Table 2.3

In order to produce a stacked barplot, they need to be in their table form. For this, we construct the matrix `natfare` as follows:

```
> natfare <- matrix(freq, byrow=TRUE,ncol=5)
> dimnames(natfare) <- list(WELFARE=c("too little","about right",
+      "too much"), DEGREE=c("LT HS","HS", "JColg","BA", "Grad"))
```

In this case, we want to define the bars by the rows of the table, thus `barplot()` is applied on the transpose of the data matrix `t(natfare)`. Hence the barplot in Fig. 2.3 (left) is obtained by

```
> barplot(t(natfare), legend.text=T, args.legend=list(x=1,y=350,
+    cex=.8), main="Opinion on National Welfare Spending by
+    Educational Level (GSS 2008)", xlab="Welfare Spending",
+    ylab="Frequencies")
```

The labels of the categories stacked are printed in the upper right corner of the plot, by default. In this case, it is convenient to move the legend box to the upper left corner. This is achieved by the argument `args.legend=list(x=, y=, cex=)`, where `x` and `y` define the `(x,y)` coordinates of the legend's location and `cex` rescales the font size of the legend. For the barplot of the conditional proportions (Fig. 2.3 right), the input matrix `t(natfare)` in the command above is replaced by the matrix `prop.table(t(natfare),2)` and the label of the y-axis is changed accordingly. Observing the conditional barplot, we realize that the distributions of educational levels within each category of opinion about welfare spending are similar, in agreement with the independence model that is not rejected for this data set.

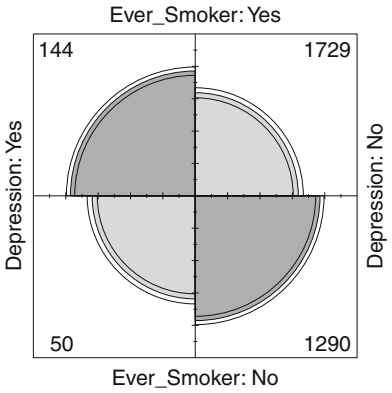


Fig. 2.4 Fourfold plot for the odds ratios of Example 2.1(a) [Table 2.1(a)]

2.4.2 Fourfold Plots

A fourfold plot provides a graphical expression of the association in a 2×2 table, visualizing the odds ratio. Each cell entry n_{ij}, $i, j = 1, 2$, is represented as a quarter-circle with radius proportional to $\sqrt{n_{ij}}$. Thus, the area of each of the quarter-circles is proportional to the corresponding cell frequency.

If the diagonal areas are greater (less) than the off-diagonal areas, then the association between the two binary classification variables is positive (negative), i.e., the odds ratio is $\theta > 1$ (< 1). The direction of the association is visually strengthened by the use of color. In case of no association ($\theta = 1$), the quarter-circles should form a circle. The test of the null hypothesis of no association is also visualized on the fourfold plot by the confidence rings provided for each quarter-circle. The observed frequencies support the null hypothesis if the rings for adjacent quarters overlap.

The fourfold plot of Example 2.1(a) is displayed in Fig. 2.4 and is obtained in package `graphics` by the function

```
> fourfoldplot(depsmok, color = c("#CCCCCC", "#999999"))
```

where `depsmok` is the data matrix, constructed in Sect. 2.1.2.

It is thus verified that the association between smoking and depression is significant (the confidence rings do not overlap) and positive (the diagonal quarters—dark colored—are of greater area).

The standard confidence level is set to 95% but can be controlled through the argument `conf.level =`. Also the colors are set by default to red–blue, i.e., `color = c("#99CCFF", "#6699CC")`. Hence a red–blue fourfold display with 99% confidence rings would be derived by

```
> fourfoldplot(depsmok, conf.level = 0.99)
```

Fourfold plots can also be drawn for the generalized odds ratios of $I \times J$ tables. For example, the fourfold plots for the local odds ratios of any $I \times J$ table can be produced in a $(I-1) \times (J-1)$ matrix form by the function `ffold.local()`, provided in the web appendix (see Sect. A.3.2). Thus, the local odds ratios of Table 3.5 can be visualized in Fig. 2.5, which is produced by

```
> ffold.local (natfare)
```

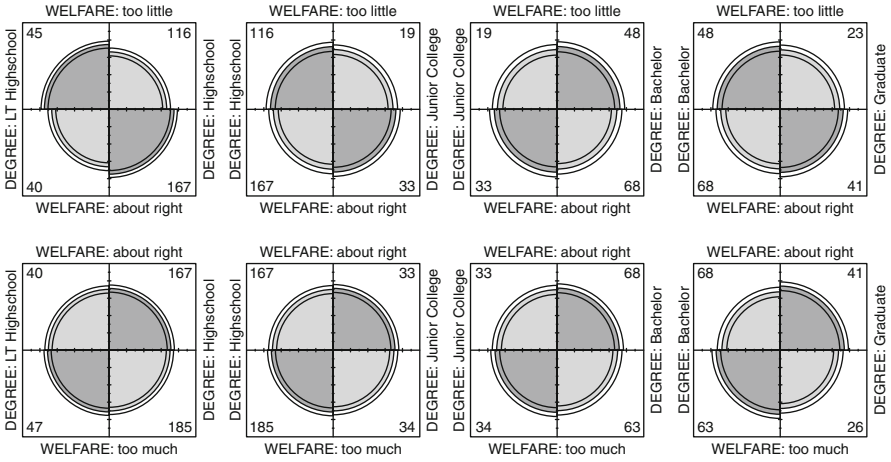


Fig. 2.5 Fourfold plots for the local odds ratios of Example 2.4 (Table 2.3)

2.4.3 Sieve Diagrams

The sieve diagram (or *parquet diagram*) represents for a $I \times J$ table the expected cell frequencies under independence, as a rectangular formed by a collection of IJ rectangles, each of them having height and width proportional to the corresponding row and column marginal frequencies, respectively. This way the area of each rectangular is proportional to the expected under independence frequency for the corresponding cell. The number of squares in each rectangular equals the observed frequency for this cell. The sieve diagram can also be constructed for the observed cell frequencies of the table. In this case the rectangles are colored and their frame is dashed according the sign of the corresponding residuals. Blue-dashed squares indicate positive while red non-dashed negative residuals.

Figure 2.6 provides the sieve diagrams for expected under independence and observed cell frequencies of Examples 2.3 and 2.4. The command `sieve()` of the `vcd` package, applied on data matrix `confinan` as

```
> sieve(confinan, sievetype="expected", shade=T)
```

and

```
> sieve(confinan, shade=T)
```

produces the sieve diagrams for Example 2.3, to be seen in Fig. 2.6 upper left and Fig. 2.6 upper right plots, respectively. The sieve diagrams for Example 2.4 are derived analogously.

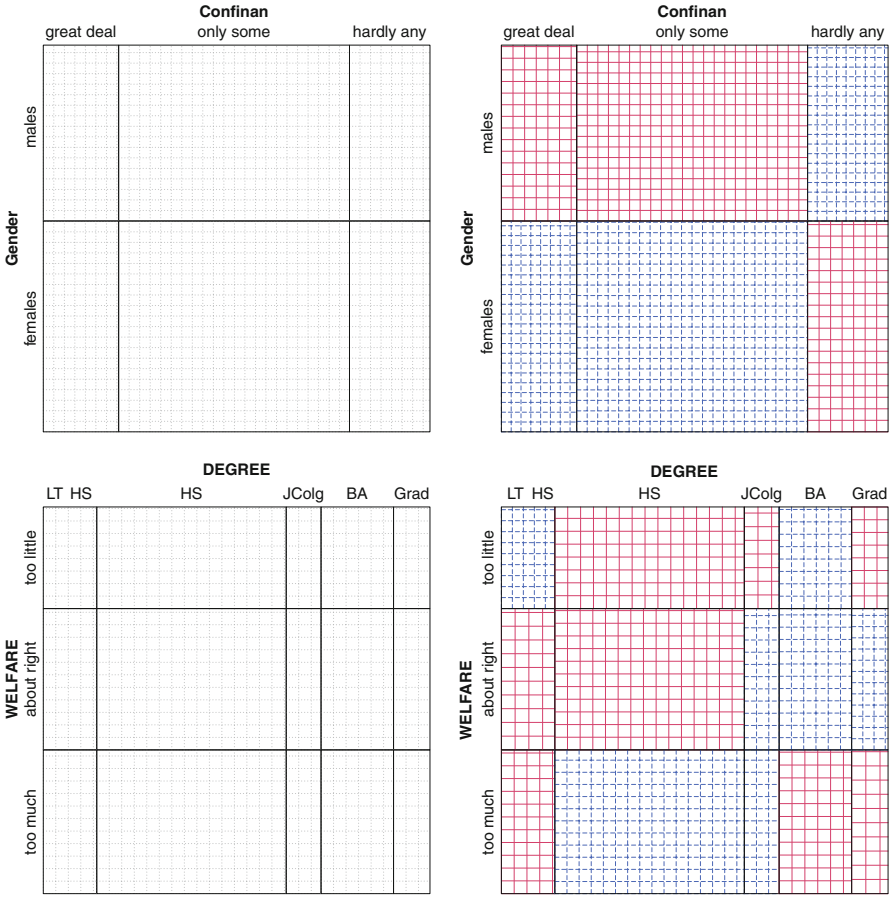


Fig. 2.6 Sieve Diagrams of the expected under independence (*left*) and the observed (*right*) cell frequencies for data of Table 2.2 (*upper*) and Table 2.3 (*lower*)

2.4.4 Mosaic Plots

Mosaic plots for two-way tables display graphically the cells of a contingency table as rectangular areas of size proportional to the corresponding observed frequencies. Were the classification variables independent, the areas would be perfectly aligned in rows and columns. The worse the alignment is, the stronger is the lack of fit for independence. Furthermore, specific locations of the table that deviate from independence the most can be identified and thus the pattern of underlying association can be explained. The strength of individual cells' contribution to divergence from independence as well as the direction of the divergence are reflected in the magnitude and sign of the corresponding independence model's residuals that can be incorporated in a mosaic plot.

Fig. 2.7 Mosaic plots based on the independence model applied on Table 2.2: plain (*left*) and incorporating the significant ($\alpha = 5\%$) deviance residuals (*right*)

Mosaic plots can be obtained in `graphics` by `mosaicplot()`. In `vcd`, the corresponding function is `mosaic()`. The simplest version of mosaic plot constructed requires only the specification of the matrix on which it is applied. Thus, `mosaic(natfare)` produces the mosaic plot for Example 2.3 (Fig. 2.7, left). The boxes corresponding to each cell have area proportional to the observed cell frequency.

The residuals for independence are incorporated in a mosaic plot through the option `residuals_type`, with default type the Pearsonian, and the option `gp` for controlling color and shading. Hence,

```
> mosaic(confinan, gp=shading_hcl)
```

shades the boxes of the nonsignificant at the 5% level Pearsonian residuals gray, colors the significant ones (blue the positive and red the negative), and reports the p-value of the independence model fit. Alternative options for `gp` are, for example, `gp=shading_max` or `gp=shading_Friendly`. The last replaces the gray-shaded boxes for nonsignificant residuals by non-shaded boxes color framed (dashed red for negative and solid blue for positive). Furthermore, `gp` can be controlled by the user. Adding in `mosaic()` the argument `labeling = labeling_residuals` would cause the printing of the residual values only for the cells of significant at 5% level residuals.

For the deviance residuals, the corresponding command would be

```
> mosaic(confinan, gp=shading_hcl, residuals_type="deviance",
+        labeling = labeling_residuals)
```

leading to the mosaic plot in Fig. 2.7 (right).

To use the standardized residuals on the mosaic plot, they have to be computed ahead and provided then in `mosaic()` through the option

```
residuals_type="Std\nresiduals"
```

This will be illustrated for Examples 2.2 and 2.3 in Sect. 5.4.1.

2.5 Overview and Further Reading

2.5.1 The Continuity Correction

On continuity correction, characteristic sources are Plackett (1964), Grizzle (1967), Cox (1970b), Pirie and Hamdan (1972), Conover (1974), and Haber (1980), while a comprehensive review can be found in Haber (1982). As also mentioned in Coull (2005), this traditional continuity correction, applied for inference on a single binomial proportion, a 2×2 contingency table and stratified 2×2 tables (Sect. 3.3), yields to conservative inference in small samples. Martín Andrés et al. (2005) and references therein explore conditions for the asymptotic X^2 test to be valid in 2×2 tables and provide validity conditions in case continuity corrections are used. In our days continuity correction is usually not preferred due to its conservatism. Alternative contemporary methods for treating discreteness exist (for a short discussion see Sect. 10.4).

2.5.2 2×2 Tables and the Odds Ratio

The extent of the literature for the analysis of the very basic 2×2 table is impressive. This lies primary on the range of different sampling schemes that generate 2×2 tables and the variability of methods that exist for analyzing such tables, as noted by Upton (1982). He mentions that Barnard (1947) was the first to report that there were at least three distinct sampling schemes leading to a 2×2 table. These three schemes were discussed in detail by Pearson (1947).

A significant part of the discussion on analyzing 2×2 contingency tables by different approaches deals with the small sample case and associated exact tests, with most famous the exact test of Fisher for testing independence (see Sect. 2.1.7), which is a conditional test (conditioning on the marginals). Its major competitor is the unconditional test of Barnard (1945, 1947), comparing two independent binomial proportions for small samples. Unconditional tests are generally preferable with small samples, since conditioning increases the discreteness and thus the conservatism of an approach. McDonald et al. (1977) provided a simpler version of Barnard's test while Silva Mato and Martín Andrés (1997) proposed a procedure that reduces the computation time of the traditional Barnard's test. An overview of the dispute conditional vs. unconditional tests can be found in the sound discussion paper by Yates (1984) while comparisons of Fisher's exact test to an unconditional test are provided by Suissa and Shuster (1985). On the same dispute, through an information theoretic approach, Cheng et al. (2008) establish information identities for testing independence in 2×2 tables, yielding a unified power analysis for Fisher's exact test, Pearson's X^2, and the LR test G^2. Barnard's exact test is nonparametric and can be more powerful for 2×2 tables (Mehta and Senchaudhuri 2003), with the cost of being computational more demanding.

With regard to p-values for small samples, the mid-p-value was first proposed by Lancaster (1961) for testing independence in a contingency table. The mid-p-value is less conservative than the p-value derived by the Fisher's exact test and has been further supported by Hirji et al. (1991), Agresti (1992), and Upton (1992), among others. Hwang and Yang (2001) provided the theoretical justification of the mid-p-value. They derived the *expected p-value* and showed that in the one-sided case it coincides to the mid-p-value while in a contingency table of two independent binomials with balanced sample sizes, it becomes the two-sided mid-p-value.

Focusing on exact confidence intervals for the odds ratio θ, the classical conditional exact $(1-\alpha)100\%$ confidence interval is derived by inverting two separate one-sided tests, each having size $\leq \alpha/2$, and is based on the noncentral hypergeometric distribution (Cornfield 1956). Agresti and Min (2001) showed that confidence intervals derived by inverting a single two-sided test are less conservative than those based on inverting two independent one-sided tests of half nominal size. Baptista and Pike (1977) were the first to propose a confidence interval based on the inversion of a single two-sided test. The concept of mid-p-value is extended to the confidence intervals as well. For a review on mid-p confidence intervals, see Berry and Armitage (1995). The mid-p confidence interval behaves better in terms of length (is shorter than Cornfield's exact and tends to be shorter than that based on two-sided p-value) but does not guarantee that the coverage probability will be at least equal to the nominal level (Agresti 2003). Exact conditional confidence intervals for the odds ratio are treated in Agresti and Min (2001) and unconditional in Agresti and Min (2002). Agresti (2003) provides an enlightening discussion on the discreteness problem related to exact confidence intervals for proportions and odds ratios, comparing confidence intervals derived by diverse methods and based on alternative p-values. A detailed discussion on exact p-values and further options in exact analysis of a 2×2 table can be found in Hirji (2006) and Agresti (2013).

Beyond small sample inference, a variety of point estimators and confidence intervals have been proposed for the odds ratio. One of the cons of odds ratio is the problem in estimating it by maximum likelihood in presence of zeros that lead either to null or infinite estimates. To overcome this, many researchers suggested the addition of a small constant $\varepsilon = 0.5$ to all cells (Haldane 1956; Gart and Zweifel 1967) or only to the zero cells (Walter and Cook 1991). This approach has been criticized because it adds "fake data," the effect of which is stronger for smaller sample sizes (Bishop et al. 1975; Agresti and Yang 1987). For the more general case of a $2 \times k$ table, corresponding to a binary response and an explanatory or factor variable of k levels, Gart et al. (1985) and Davis (1985) have shown that the optimal ε correction depends on k. Alternative estimators have been proposed by Berkson (1953) and Birch (1964) and for small samples by Jewell (1984, 1986) and Walter (1985). Gart and Zweifel (1967) and Walter and Cook (1991) compared different estimators. Parzen et al. (2002) suggest an alternative estimator that always lies in $(0, \infty)$ and compare it to the standard obtained by adding 0.5 to all cells of the table. They also provide bootstrap confidence intervals for the odds ratio. Confidence intervals have been considered also by Gart and Thomas (1982) while the small sample behavior of various confidence intervals for the odds ratio has been studied by Agresti (1999).

In biomedical and behavioral sciences, the odds ratio is connected to the relative risk and often (not always correctly) interpreted as relative risk. Furthermore, its role is fundamental in meta analysis studies (cf. Kulinskaya et al. 2008). Newcombe (2006) demonstrates a deficiency of the odds ratio as a measure of effect size and argues for the relative risk. The role of the odds ratio in case-control design in connection to the way the controls have been selected is discussed by Pearce (1993). Limitations of the odds ratio in evaluating the performance of markers are exposed by Pepe et al. (2004), who suggest as an alternative the use of ROC curves and logistic regression. Kraemer (2004) criticizes the use of the odds ratio as a measure of association and uses ROC methods to point out when it produces misleading results. The major issue is for cases of *perfect association*, i.e., $\theta = \infty$. Rudas and Bergsma (2004) however commented Kraemer's attitude and stated that it is a matter of definition of the perfect association.

2.5.3 *Inference for Two-way Tables*

The X^2 test of independence in contingency tables, one of the most widely used statistical tests, was introduced by Pearson (1900a) while the term *contingency table* appeared first in Pearson (1904). Pearson however assigned to the test wrongly the degrees of freedom, which were later corrected by Fisher (1922). For an early literature review and a discussion on the impact of Pearson's work on X^2, we refer to Plackett (1983) and Stigler (2008). Wilks (1935) proposed the LR test for testing independence in contingency tables.

Pearson's X^2 (and G^2 as well) for testing independence tends to be highly significant when the sample size n is large, without necessarily the corresponding table being that far from independence. This was first pointed out by Berkson (1938). For this, Diaconis and Efron (1985) in a stimulating discussion paper introduce the *volume test*, by considering the uniform alternative, under which all tables of a given dimension and sample size are equal probable. However, this problem is not restricted only to two-way tables and the hypothesis of independence.

The traditional type of estimation associated with contingency tables and log-linear models is the maximum likelihood (ML). The method, developed by Fisher in 1912–1922, was named as maximum likelihood in 1922 (for related history and the development of related concepts such as sufficiency, efficiency, and information, see Aldrich (1997) and references cited there). A discussion on the major contributions in the development of ML estimation of log-linear models will be provided in Sect. 4.9, after introducing log-linear models for multi-way contingency tables.

2.5.4 *Partitioning of the X^2 Statistic*

A popular approach for explaining the lack of fit of the independence model was the *partitioning* of the X^2 statistic. The first partition of the total X^2 statistic in a $I \times J$ table is due to Irwin (1949) and Lancaster (1949, 1950). By such a partitioning, $(I-1)(J-1)$ statistics of one degree of freedom are obtained and they can be used to test orthogonal contrasts. Kastenbaum (1960) managed to handle testing for a broader class of orthogonal contrasts, with one or more degrees of freedom. Early related contributions are also these of Yates (1948) and Cochran (1954), directing the departure from the null hypothesis toward alternatives of particular type. Partitioning of the X^2 statistic for multi-way tables has been considered by Goodman (1969, 1971c).

Johnson (1975) and Gokhale and Johnson (1978) proposed a class of alternative hypotheses to independence in two-way contingency tables by removing from a set of cells the probability mass under independence and redistributing it over the remaining cells, preserving the marginal totals. The alternatives are expressed in a log-linear form and can be analyzed by minimum discrimination information, maximum likelihood, or weighted least squares.

2.5.5 *Ordinal Odds Ratios and Positive Dependencies*

It is clear by now the crucial role odds ratios play in the analysis of contingency tables. Ordinal contingency tables are connected to the ordinal odds ratios, presented in Sect. 2.2.5. Although the analysis of ordinal contingency tables will be discussed in detail in Chaps. 6–9, we refer here briefly to their connection to concepts of *positive dependence*, in order to highlight the role of the type of ordinal odds ratio used.

By constraining specific log odds ratios of a table to be nonnegative, different notions of positive dependence are ensured (see Douglas et al. 1990 or Silvapulle and Sen 2005). Thus the positivity of all log local odds ratios ($\log \theta_{ij}^L > 0$, $i = 1,\ldots,I-1$, $j = 1,\ldots,J-1$) is equivalent to the strongest notion of positive dependence, the total positivity of order 2 (TP_2). TP_2 is equivalent to the positive likelihood ratio dependence (Dykstra et al. 1995). Analogously, the positivity of all the log cumulative odds ratios for all ways of collapsing the response (here the column classification variable) to binary

$$\log \theta_{ij}^C = \log \left(\frac{\left(\sum_{k \le j} \pi_{ik}\right)\left(\sum_{k > j} \pi_{i+1,k}\right)}{\left(\sum_{k > j} \pi_{ik}\right)\left(\sum_{k \le j} \pi_{i+1,k}\right)} \right) > 0, \quad i = 1,\ldots,I-1, \quad j = 1,\ldots,J-1,$$

is equivalent to the positive regression dependence while that of the log global odds ratios

$$\log \theta_{ij}^{G} = \log \left(\frac{\left(\sum_{l \le i} \sum_{k \le j} \pi_{lk}\right) \left(\sum_{l > i} \sum_{k > j} \pi_{lk}\right)}{\left(\sum_{l \le i} \sum_{k > j} \pi_{lk}\right) \left(\sum_{l > i} \sum_{k \le j} \pi_{lk}\right)} \right) > 0, \;\; i = 1, \ldots, I-1, \;\; j = 1, \ldots, J-1,$$

to that of the positive quadrant dependence, introduced by Lehmann (1966). The likelihood ratio ordering is the strongest ordering and implies the other weaker types. For a detailed insight on the various concepts of orderings, please refer to Shaked and Shanthikumar (2007). Dardanoni and Forcina (1998) considered various hypotheses of stochastic orders among the conditional row distributions of two-way contingency tables with ordered margins.

The focus in this book lies on local odds ratios, since they are appropriate for nominal and ordinal classification variables. However, cumulative or global odds ratios can be modeled in a similar manner, as illustrated in Sects. 5.6.1 and 7.1.1. The choice of the type of the ordinal odds ratios used in an analysis lies mainly on the specific application under consideration and the researcher's decision about whether description will refer to individual categories or to groupings (e.g., above vs. below) of categories. The problem of local vs. global odds ratios choice will be further discussed in the context of association models in Sect. 7.1.

Chapter 3
Analysis of Multi-way Tables

Abstract Issues discussed in Chap. 2 for two-way tables are extended to multi-way contingency tables. Emphasis is given to clarifying the concepts of partial and marginal association. Further on, stratified 2×2 tables are analyzed by the Mantel–Haenszel and the Breslow–Day–Tarone tests. Types of independence for three-way tables are introduced. Graphs are presented for multi-way contingency tables while fourfold plots are used to visualize stratified 2×2 tables. All examples are implemented in R.

Keywords Partial and marginal tables • Stratified 2×2 tables • Homogeneous association tests • Independence for three-way tables • Graphs for multi-way tables

3.1 Describing Multi-way Contingency Tables

Multi-way contingency tables are very common in practice, derived by the presence of more than two cross-classification variables. For a three-way contingency table (n_{ijk}), with $i = 1,\ldots,I$, $j = 1,\ldots,J$, and $k = 1,\ldots,K$ indexing the row, column, and layer level of the classification variables X, Y, and Z, respectively, there are associated *partial* and *marginal* tables. A partial table is the cross-classification of two of these variables for fixed level of the remaining third. Thus, there are three possible sets of partial tables, according to which variable is kept at a fixed level. By keeping a variable at a fixed level, we control over this variable. In particular, the set of the XY-partial tables consists of the K corresponding two-way layers, denoted as $(n_{ij(k)})$ for $k = 1,\ldots,K$. In terms of notation, the index of the control variable is given in parenthesis. Analogously, the XZ- and YZ-partial tables are denoted as $(n_{i(j)k})$ and $(n_{(i)jk})$, respectively. Summing all possible layers of a set of partial tables leads to the corresponding marginal table. Thus the XY-, XZ-, and YZ-marginal tables are the (n_{ij+}), (n_{i+k}), and (n_{+jk}), respectively. While a partial table controls over the third variable, a marginal table ignores it. Furthermore,

M. Kateri, *Contingency Table Analysis: Methods and Implementation Using R*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4811-4_3,

Table 3.1 Smoking habit vs. major depressive disorder by gender for a sample collected in the St. Louis epidemiologic catchment area survey (Glassman et al. 1990)

Ever smoked	Major depression		Ever smoked	Major depression	
	Yes	No		Yes	No
Yes	40	889	Yes	104	840
No	10	417	No	40	873

information on the single classification variables is summarized in the marginal vectors $(n_{1++}, \ldots, n_{I++})$, $(n_{+1+}, \ldots, n_{+J+})$, and $(n_{++1}, \ldots, n_{++K})$, respectively.

A multi-way $I_1 \times I_2 \ldots \times I_q$ contingency table will analogously be denoted as $(n_{i_1 i_2 \ldots i_q})$, $i_\ell = 1, \ldots, I_\ell$, $\ell = 1, \ldots, q$, while the definition of partial and marginal tables follows straightforward. In general, for a q-way table, there can be defined $(q-1)$-way down to two-way partial tables, when controlling 1 up to $q-2$ variables, respectively. Analogously, the marginal tables formed are from one way (i.e., vectors) up to $(q-1)$ way. Thus, for an $I_1 \times \ldots \times I_5$ table $(n_{i_1 i_2 i_3 i_4 i_5})$, $(n_{i_1 (i_2) i_3 i_4 (i_5)})$ is a three-way partial table controlling over the second and fifth classification variables while $(n_{i_1 i_2 (i_3 i_4 i_5)})$ is a two-way partial table controlling the third up to the fifth classification variables.

A characteristic $2 \times 2 \times 2$ data example follows.

3.1.1 Example 3.1

Consider the 2×2 data set in Example 2.1 (a), where a sample was cross-classified according to smoking and depression. This data set is actually a marginal table from the data in Glassman et al. (1990), ignoring gender. The data set is a $2 \times 2 \times 2$ table, providing the smoking (S) vs. depression (D) vs. gender (G) cross-classification. The data are provided in Table 3.1, where the layers are defined by the gender ($k = 1, 2$ for males and females, respectively). Thus, the SD-partial tables $(n_{ij(1)})$ and $(n_{ij(2)})$ are the left and right two-way tables in Table 3.1, respectively, while Table 2.1(a) is the SD-marginal table (n_{ij+}).

Multi-way contingency tables can be defined in R through the `array()` command (see Sect. A.2.2 of the Appendix). Thus, given the data entries in a vector form, Table 3.1 can be produced in R and saved under `depsmok3` as follows:

```
> freq <- c(40, 10, 889, 417, 104, 40, 840, 873)
> names <- list(Ever_Smoker=c('Yes', 'No'), Depression=
+           c('Yes', 'No'), Gender=c('male', 'female'))
> depsmok3 <- array(freq, c(2,2,2), dimnames=names)
```

Partial and marginal tables are easily produced in R. For our example, the $(n_{ij(1)})$ and $(n_{ij(2)})$ partial tables are simply the `depsmok3[,,1]` and `depsmok3[,,2]`,

respectively. The depression vs. gender partial table for smokers would then be `depsmok3[1,,]`. The marginal table (n_{ij+}), i.e., the 2×2 table of Example 2.1 (a), is derived by

```
> margin.table(depsmok3, c(1,2))
```

while the marginal gender vector $(n_{++1}, n_{++2}) = (1356, 1857)$ would be

```
> margin.table(depsmok3, c(3))
```

3.2 On Partial and Marginal Tables

3.2.1 Joint, Conditional, and Marginal Probabilities

Recall that for a two-way $I \times J$ table of observed frequencies (n_{ij}), associated were the *joint* probabilities (π_{ij}), the *conditional within rows* $(\pi_{j|i})$ with $\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}}$, $i = 1, \ldots, I$ (or analogously within columns), and the *row* and *column marginal* probabilities, $(\pi_{1+}, \ldots, \pi_{I+})$ and $(\pi_{+1}, \ldots, \pi_{+J})$, respectively.

Analogously, for three-way tables, the table of joint probabilities is (π_{ijk}), with

$$\pi_{ijk} = \mathsf{P}(X = i, Y = j, Z = k), \;\; i = 1, \ldots, I, \;\; j = 1, \ldots, J, \;\; k = 1, \ldots, K \,. \qquad (3.1)$$

The marginal row (X) probability vector is defined as $(\pi_{1++}, \ldots, \pi_{I++})$, with $\pi_{i++} = \mathsf{P}(X = i)$, while the marginal column (Y) and layer (Z) probability vectors, $(\pi_{+1+}, \ldots, \pi_{+J+})$ and $(\pi_{++1}, \ldots, \pi_{++K})$, are defined analogously. Furthermore, there are also defined the XY, XZ, and YZ marginal probability tables (π_{ij+}), (π_{i+k}), and (π_{+jk}), with $\pi_{ij+} = \mathsf{P}(X = i, Y = j)$, $\pi_{i+k} = \mathsf{P}(X = i, Z = k)$, and $\pi_{+jk} = \mathsf{P}(Y = j, Z = k)$. The correspondence between marginal frequencies and marginal probabilities tables is obvious.

The partial frequency tables are related to corresponding conditional probabilities for the three-way table. For example, the conditional within layers probabilities table $(\pi_{ij|k})$, where $\pi_{ij|k} = \pi_{ijk}/\pi_{++k}$, $k = 1, \ldots, K$, corresponds to the partial frequency table $(n_{ij(k)})$, and its sample estimate is $(p_{ij|k}) = (n_{ijk}/n_{++k})$.

Joint, conditional, and marginal probabilities for multi-way tables are defined analogously.

3.2.2 Conditional and Marginal Odds Ratios for $2 \times 2 \times K$ Tables

Consider a $2 \times 2 \times K$ contingency table cross-classifying two binary characteristics X and Y across the K levels of an explanatory variable Z. Then, for the partial frequency table $(n_{ij(k)})$, $i, j = 1, 2$, at each level k of Z, $k = 1, \ldots, K$, the associated

conditional probabilities table is $(\pi_{ij|k})$. The odds ratio can be defined for each of these conditional probabilities tables, straightforward, as

$$\theta_{(k)}^{XY} = \frac{\pi_{11|k}\pi_{22|k}}{\pi_{12|k}\pi_{21|k}} = \frac{\frac{\pi_{11k}}{\pi_{++k}}\frac{\pi_{22k}}{\pi_{++k}}}{\frac{\pi_{12k}}{\pi_{++k}}\frac{\pi_{21k}}{\pi_{++k}}} = \frac{\pi_{11k}\pi_{22k}}{\pi_{12k}\pi_{21k}}, \quad k = 1,\ldots,K. \tag{3.2}$$

These odds ratios are known as *conditional odds ratios*.

The odds ratio for the corresponding marginal probabilities table (π_{ij+})

$$\theta^{XY} = \frac{\pi_{11+}\pi_{22+}}{\pi_{12+}\pi_{21+}}, \tag{3.3}$$

is called a *marginal odds ratio*.

Marginal and conditional odds ratios express the association between the variables denoted in their superscripts and are estimated by the corresponding sample odds ratios. Thus, the estimates of (3.2) and (3.3) are $\hat{\theta}_{(k)}^{XY} = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}$, $k = 1,\ldots,K$, and $\hat{\theta}^{XY} = \frac{n_{11+}n_{22+}}{n_{12+}n_{21+}}$. However, the conditional odds ratios express the *XY partial* association controlling over the level of Z, while the marginal odds ratio expresses the *XY marginal* association, ignoring Z. Conditional odds ratios can differ substantially over k, even in direction of association. This phenomenon is known as *Simpson's paradox* and will be discussed in more extent in Sect. 4.8. If this is the case, the marginal odds ratio will be misleading for describing the XY association, which has to be captured by the conditional odds ratios, taking into consideration the explanatory variable Z. Marginal and conditional odds ratios are illustrated next for Example 3.1.

3.2.2.1 Example 3.1 (Continued)

We have seen in Sect. 2.1.6 that for the marginal 2×2 table the sample odds ratio was $\hat{\theta} = 2.149$. The sample odds ratio for each gender, i.e., for each partial table, along with the associated asymptotic confidence intervals (CI), can be computed in `R` by applying the `odds.ratio()` function (discussed in Sect. 2.1.6) on the 2×2 partial tables `depsmok3[„1]` and `depsmok3[„2]`, respectively. This way we get $\hat{\theta}_{(1)} = 1.876$ and $\hat{\theta}_{(2)} = 2.70$, while the associated 95% CI are $(0.93, 3.79)$ and $(1.85, 3.94)$ for males and females, respectively. These sample conditional odds ratios indicate that there is a smoking-depression association but it differentiates between males and females, being stronger for women. This issue will be discussed further in Sect. 3.3.3.

3.2.3 Odds Ratios for Tables of Higher Dimension

Conditional and marginal odds ratios can be defined for any two-way conditional and marginal probabilities table of a multi-way $I_1 \times I_2 \times \ldots \times I_q$ table with $I_\ell \geq 2$, $\ell = 1, \ldots, q$. In this case, the conditional and marginal odds ratios are defined as odds ratios for two-way tables of size $I \times J$, greater than 2×2. Thus, as defined for general two-way tables in Sect. 2.2.5, they will be a minimal set of odds ratios of nominal, local, cumulative, or global type.

For an $I \times J \times K$ table, based on (2.45), the XY conditional local odds ratios are defined as

$$\theta^{XY}_{ij(k)} = \frac{\pi_{ijk}\pi_{i+1,j+1,k}}{\pi_{i+1,j,k}\pi_{i,j+1,k}}\,, \quad i = 1, \ldots, I-1, \quad j = 1, \ldots, J-1, \quad k = 1, \ldots, K\,, \tag{3.4}$$

and the XY marginal local odds ratios as

$$\theta^{XY}_{ij} = \frac{\pi_{ij+}\pi_{i+1,j+1,+}}{\pi_{i+1,j,+}\pi_{i,j+1,+}}\,, \quad i = 1, \ldots, I-1, \quad j = 1, \ldots, J-1\,. \tag{3.5}$$

The conditional and marginal odds ratios of other types, like nominal, cumulative, or global, are defined analogously.

Conditional and marginal odds ratios play a crucial role in understanding the nature of the underlying association structure in multi-way tables and interpreting fitted models, as odds ratios do for two-way tables.

3.2.4 Example 3.2

The $5 \times 7 \times 2$ data table, given in Table 3.2, is from the General Social Survey basis for year 2008 (GSS 2008), cross-classifying responders by their educational level (D: highest degree obtained), political party affiliation (P), and gender (G).

For the DP partial tables (within gender) and the DP marginal table, the conditional and marginal sample local odds ratios are computed by (3.4) and (3.5), respectively, and provided in Table 3.3. In `R`, they can be derived applying the `local.odds.DM()` function (web appendix, see Sect. A.3.2), as shown below. The data are entered in a vector form by columns and transformed to a three-way array `party.tab` as follows:

```
> freq <- c(32,67,12,23,16,20,85,14,21,9,18,63,6,29,12,29,68,9,
            20,13,11,48,13,19,7,12,65,17,32,14,9,44,6,20,13,
            31,118,20,33,38,25,98,16,23,20,16,69,13,28,8,58,88,
            13,11,13,8,30,7,16,3,8,82,16,44,13,16,54,7,23,9)
names <- list(D=c("LT HSc","HSc","JunCol","Bachelor","Graduate"),
          P=c("1","2","3","4","5","6","7"), G=c("male","female"))
party.tab <- array (freq, c(5,7,2), dimnames=names)
```

Table 3.2 Respondents' cross-classification (*GSS* 2008) by degree (1: LT high school, 2: high school, 3: junior college, 4: bachelor, 5: graduate), political party affiliation (1: strong Democrat, 2: not strong Democrat, 3: independent (nearly Democrat), 4: independent, 5: independent (nearly Republican), 6: not strong Republican, 7: strong Republican), and gender (1: male, 2: female)

(G): males	Political party affiliation (P)						
Degree (D)	1	2	3	4	5	6	7
1: LT high school	32	20	18	29	11	12	9
2: High school	67	85	63	68	48	65	44
3: Junior college	12	14	6	9	13	17	6
4: Bachelor	23	21	29	20	19	32	20
5: Graduate	16	9	12	13	7	14	13
(G): Females	Political party affiliation (P)						
Degree (D)	1	2	3	4	5	6	7
1: LT high school	31	25	16	58	8	8	16
2: High school	118	98	69	88	30	82	54
3: Junior college	20	16	13	13	7	16	7
4: Bachelor	33	23	28	11	16	44	23
5: Graduate	38	20	8	13	3	13	9

The *DP* partial tables for male and female are respectively

```
> DP1 <- party.tab[„1];  DP2 <- party.tab[„2]
```

and the *DP* marginal (over gender) table is

```
> DPm <- margin.table(party.tab, c(1,2))
```

In this setup, the design matrix `C`, applied for the construction of the local odds ratios table in (2.54), requires the 5×7 frequency table in a vector form, expanded by rows. For this, the 4×6 table $\left(\hat{\theta}^{DP}_{ij(1)}\right)$ for males is derived by

```
> C <- local.odds.DM(5, 7)
> LOR1 <- as.vector(C%*%log(as.vector(t(DP1))))
> OR1 <- exp(t(matrix(LOR1, NJ-1)))
```

Changing table `DP1` by `DP2` and `DPm`, the $\left(\hat{\theta}^{DP}_{ij(2)}\right)$ and $\left(\hat{\theta}^{DP}_{ij}\right)$ tables are produced, respectively.

We observe that the conditional sample local odds ratios in Table 3.3 are diverse within each gender and in some cases far apart from 1. On the other hand, for given i, j, in most cases, the $\hat{\theta}^{DP}_{ij(1)}$ and $\hat{\theta}^{DP}_{ij(2)}$ are quite close to each other and close to the corresponding $\hat{\theta}^{DP}_{ij}$ marginal sample local odds ratio. This is an indication that there exists an association between education and party affiliation, but it seems not to differentiate between males and females. This indication will be verified by fitting and interpreting the appropriate model on this data set in Sect. 4.6.2.

Since D and P are both ordinal, we could analogously compute the partial or marginal global odds ratios.

Table 3.3 *DP* conditional and marginal sample local odds ratios for the data in Table 3.2. The partial $\hat{\theta}^{DP}_{ij(1)}$, $\hat{\theta}^{DP}_{ij(2)}$ and the marginal $\hat{\theta}^{DP}_{ij}$ are in Tables (a), (b), and (c), respectively

(a) $\hat{\theta}^{DP}_{ij(1)}$ (G: Males)	Political party affiliation (P)						
Degree (D)	1	2	3	4	5	6	7
1: LT high school	2.030	0.824	0.670	1.861	1.241	0.903	
2: High school	0.920	0.578	1.390	2.046	0.966	0.521	
3: Junior college	0.783	3.222	0.460	0.658	1.288	1.771	
4: Bachelor	0.616	0.966	1.571	0.567	1.187	1.486	
5: Graduate							
(b) $\hat{\theta}^{DP}_{ij(2)}$ (G: Females)	Political party affiliation (P)						
Degree (D)	1	2	3	4	5	6	7
1: LT high school	1.030	1.100	0.352	2.472	2.733	0.329	
2: High school	0.963	1.154	0.784	1.579	0.836	0.664	
3: Junior college	0.871	1.498	0.393	2.701	1.203	1.195	
4: Bachelor	0.755	0.329	4.136	0.159	1.576	1.324	
5: Graduate							
(c) $\hat{\theta}^{DP}_{ij}$	Political party affiliation (P)						
Degree (D)	1	2	3	4	5	6	7
1: LT high school	1.385	0.955	0.462	2.289	1.790	0.533	
2: High school	0.948	0.878	0.980	1.818	0.876	0.591	
3: Junior college	0.838	2.045	0.470	1.242	1.316	1.436	
4: Bachelor	0.684	0.532	2.390	0.341	1.243	1.441	
5: Graduate							

3.3 Analysis of K 2×2 Tables

Let X and Y be binary variables that are cross-classified across the K strata of a variable Z, forming thus 2×2 partial tables $\left(n_{ij(k)}\right)$, $k = 1, \ldots, K$. If X and Y are independent in each partial table, i.e., *given* the level k of Z, then X, Y are *conditionally independent, given Z*. In this case, it holds

$$\theta^{XY}_{(1)} = \theta^{XY}_{(2)} = \ldots = \theta^{XY}_{(K)} = 1 \; . \tag{3.6}$$

With respect to the marginal odds ratio θ^{XY}, condition (3.6) does not generally imply $\theta^{XY} = 1$, which corresponds to *marginal independence* of X and Y.

To visualize this, consider the following toy example. In the framework of a social survey, carried out at four different cities, responders are classified according to their opinion on an issue and gender. Is opinion independent of gender?

Opinion (Y)			
City (Z)	Gender (X)	Agree	Disagree
1	Male	8	4
	Female	12	6
2	Male	10	4
	Female	5	2
Total	Male	29	44
	Female	24	70

Opinion (Y)			
City (Z)	Gender (X)	Agree	Disagree
3	Male	2	30
	Female	4	60
4	Male	9	6
	Female	3	2

It is easy to verify that although the sample estimates of all the conditional odds ratios are all equal to 1

$$\hat{\theta}^{XY}_{(1)} = \frac{(8)(6)}{4)(12)} = 1,\ \hat{\theta}^{XY}_{(2)} = \frac{(10)(2)}{(4)(5)} = 1,\ \hat{\theta}^{XY}_{(3)} = \frac{(2)(60)}{(30)(4)} = 1,\ \hat{\theta}^{XY}_{(4)} = \frac{(9)(2)}{(6)(3)} = 1\ ,$$

indicating that the opinion is independent of the gender, the estimated marginal (over cities) odds ratio signals that the odds of agreeing for males is almost twice as high as that for females

$$\hat{\theta}^{XY} = \frac{(29)(70)}{(44)(24)} = 1.92\ .$$

This is a case under which the information obtained for the association between two classification variables *ignoring* a third grouping variable (here the city), i.e., the information obtained from the collapsed over the grouping variable table, contradicts information in the stratified tables. More generally, information in the stratified tables can even be of the opposite direction from that of the collapsed table (*Simpson's paradox*; see also Sects. 3.2.2 and 4.8).

For stratified 2×2 tables of the form discussed here, X and Y may not be conditional independent given Z, but the underlying XY association may be homogeneous across the levels of the conditioning variable. Then it holds

$$H_0:\ \theta^{XY}_{(1)} = \theta^{XY}_{(2)} = \ldots = \theta^{XY}_{(K)} = \theta\ . \tag{3.7}$$

The conditional independence of X and Y, (3.6), is a special case for $\theta = 1$.

An estimate for the common θ, proposed by Mantel and Haenszel (1959), is

$$\hat{\theta}_{MH} = \frac{\sum_k \left(\frac{n_{11k}n_{22k}}{n_{++k}}\right)}{\sum_k \left(\frac{n_{12k}n_{21k}}{n_{++k}}\right)}, \tag{3.8}$$

which gives more weight to layers of larger sample size. This becomes obvious by expressing $\hat{\theta}_{MH}$ in terms of the conditional sampling proportions $p_{ij|k} = n_{ijk}/n_{++k}$.

In case K is large and the data are sparse, then the estimate (3.8) of Mantel and Haenszel is preferred over the MLE of θ (see Agresti 2013, p. 229).

3.3.1 The Mantel–Haenszel Test

A popular test for testing the null hypothesis (3.6), i.e., conditional independence of two binary classification variables X and Y, over the strata defined by a third variable Z, is the test of Mantel and Haenszel (1959). It considers K stratified 2×2 tables and conditions on the row and column marginals of each of the K partial tables. Thus, for every partial table, n_{11k} follows the hypergeometric distribution (like Fisher's exact test) with mean and variance,

$$\propto_{11k} = \frac{n_{1+k} n_{+1k}}{n_{++k}} \quad \text{and} \quad \sigma^2_{11k} = \frac{n_{1+k} n_{2+k} n_{+1k} n_{+2k}}{n^2_{++k}(n_{++k} - 1)} ,$$

respectively. Then $\sum_k n_{11k}$ has mean $\sum_k \propto_{11k}$ and variance $\sum_k \sigma^2_{11k}$, since the partial tables are independent to each other, and the Mantel–Haenszel test statistic is defined as

$$T_{MH} = \frac{\left[\sum_k (n_{11k} - \propto_{11k})\right]^2}{\sum_k \sigma^2_{11k}} . \tag{3.9}$$

T_{MH} is asymptotically $\mathcal{X}^2_1$ distributed under the hypothesis (3.6). If $T_{MH(\text{obs})}$ is the observed value of the test statistic for a particular case, then p-value=$\mathsf{P}(\mathcal{X}^2_1 > T_{MH(\text{obs})})$. Mantel and Haenszel proposed it with a *continuity correction*. This way the test statistic approximates better an exact conditional test but is more conservative.

When the XY association is similar across the partial tables, then the test is more powerful and is similar to the test of conditional independence, given that the XY association is homogeneous across the strata, in the log-linear models framework (see Sect. 4.6.1). It loses in power when the underlying associations vary across strata, especially when they are of different direction, since the differences $n_{11k} - \propto_{11k}$ will then cancel out in the sum of the statistic (3.9).

Cochran (1954) proposed the test statistic (3.9) as well but with different σ^2_{11k} values. This difference in n_{11k}'s variance occurred because Cochran considered a different sampling scheme than that of Mantel and Haenszel. In particular, he assumed the two rows of each partial table to be independent binomials, deriving thus

$$\hat{\sigma}^2_{11k} = \frac{n_{1+k} n_{2+k} n_{+1k} n_{+2k}}{n^3_{++k}} .$$

However, the difference is of no practical importance.

3.3.2 Homogeneous Association Tests

Breslow and Day (1980) proposed a large-sample test for testing the H_0 of homogeneity of odds ratios (3.7). It is based on the conditional distribution under H_0 of n_{11k}, given $\mathbf{n}_c = (n_{1+k},\ n_{+1k},\ n_{++k})$, which is *noncentral hypergeometric* for $\theta \neq 1$, and the fact that n_{11k} from different strata are independent. Let $\propto_{11k}(\theta) = \mathrm{E}(n_{11k}|\mathbf{n}_c, \theta)$ and $\sigma^2_{11k}(\theta) = \mathrm{Var}(n_{11k}|\mathbf{n}_c, \theta)$ be the conditional mean and variance of n_{11k}, respectively. Then $\propto_{11k}(\theta)$ is estimated by the acceptable solution of the quadratic equation

$$\frac{\hat{\propto}_{11k}(n_{2+k} - n_{+1k} + \hat{\propto}_{11k})}{(n_{+1k} - \hat{\propto}_{11k})(n_{1+k} - \hat{\propto}_{11k})} = \theta\ , \tag{3.10}$$

where θ is the common odds ratio under (3.7). Furthermore,

$$\hat{\sigma}^2_{11k}(\theta) = \left(\frac{1}{\hat{\propto}_{11k}} + \frac{1}{(n_{2+k} - n_{+1k} + \hat{\propto}_{11k})} + \frac{1}{(n_{+1k} - \hat{\propto}_{11k})} + \frac{1}{(n_{1+k} - \hat{\propto}_{11k})} \right)^{-1}. \tag{3.11}$$

Based on the independence of the n_{11k}'s for $k = 1, \ldots, K$, they proposed the following test statistic

$$BD = \sum_k \frac{\left[n_{11k} - \hat{\propto}_{11k}(\hat{\theta}_{MH})\right]^2}{\hat{\sigma}^2_{11k}(\hat{\theta}_{MH})}, \tag{3.12}$$

where $\hat{\propto}_{11k}(\hat{\theta}_{MH})$ and $\hat{\sigma}^2_{11k}(\hat{\theta}_{MH})$ are evaluated by (3.10) and (3.11), respectively, substituting θ in (3.10) by the Mantel–Haenszel estimator $\hat{\theta}_{MH}$, given in (3.8). These $\hat{\propto}_{ijk}$'s lead to estimated partial tables that have the same marginals as the observed data and odds ratio equal to $\hat{\theta}_{MH}$. The hypothesis of homogeneous association (3.7) is rejected for high values of BD, based on its asymptotic $\mathscr{X}^2_{K-1}$ distribution under (3.7).

Tarone (1985) adjusted the test of Breslow and Day for the inefficiency of Mantel–Haenszel's $\hat{\theta}_{MH}$ as follows:

$$BDT = \sum_k \frac{\left[n_{11k} - \hat{\propto}_{11k}(\hat{\theta}_{MH})\right]^2}{\hat{\sigma}^2_{11k}(\hat{\theta}_{MH})} - \frac{\left[n_{11+} - \hat{\propto}_{11+}(\hat{\theta}_{MH})\right]^2}{\hat{\sigma}^2_{11+}(\hat{\theta}_{MH})}. \tag{3.13}$$

This is known as the Breslow–Day–Tarone test, is also asymptotically $\mathscr{X}^2_{K-1}$ distributed under (3.7), and is the most frequently used one.

Another test for heterogeneity in stratified 2×2 tables is the test proposed by Woolf (1955). He proposed one of the first estimators for the common θ as the weighted mean of the sample log odds ratios

$$\hat{\theta}_k = \frac{n_{11k} n_{22k}}{n_{12k} n_{21k}}, k = 1, \ldots, K ,$$

with weights the inverse of their variances. In particular,

$$\hat{\theta}_W = \exp\left(\frac{\sum_{k=1}^{K} \left(w_k \log \hat{\theta}_k \right)}{\sum_{k=1}^{K} w_k} \right) , \tag{3.14}$$

where

$$w_k = \left(\frac{1}{n_{11k}} + \frac{1}{n_{12k}} + \frac{1}{n_{21k}} + \frac{1}{n_{22k}} \right)^{-1} , k = 1, \ldots, K .$$

Based on this estimator he proposed the Woolf statistic for testing (3.7),

$$W = \sum_k w_k \left(\log \hat{\theta}_k - \log \hat{\theta}_W \right)^2 , \tag{3.15}$$

which is under the null hypothesis (3.7) asymptotically $\mathcal{X}^2_{K-1}$ distributed.

3.3.3 *Example 3.1 (Continued)*

Recall that for the example in Sect. 3.1.1, the sample conditional odds ratios between smoking and depression, conditioning on gender, were computed in Sect. 3.2.2.1. They were both far from 1, indicating that the underlying association is significant. Indeed, testing hypothesis (3.6) by the Mantel–Haenszel test, the test statistic (3.9) equals $T_{MH} = 30.62$ and the H_0 that the homogeneous conditional odds ratios equal 1 is rejected (p-value=3.141e-08).

The Mantel–Haenszel test can be performed in R by `mantelhaen.test()`, which tests (3.6), assuming that (3.7) holds. The corresponding output for the data in Table 3.1 is given below.

```
> mantelhaen.test(depsmok3, correct = FALSE)

        Mantel-Haenszel chi-squared test without continuity correction
data: depsmok3
Mantel-Haenszel X-squared=30.6184, df=1, p-value=3.141e-08
alternative hypothesis: true common odds ratio not equal to 1
95 percent confidence interval:
1.779109 3.463946
sample estimates:
common odds ratio
2.482486
```

Table 3.4 The Breslow–Day test (with and without Tarone's adjustment) and the Woolf test for Example 3.1

```
> BDT(depsmok3)
$MH.theta
common odds ratio
2.482486
Breslow-Day X2 statistic
$X2
[1] 0.8083629
$df
[1] 1
$p.value
[1] 0.3686047

Breslow-Day-Tarone X2 statistic
$T.X2
[1] 0.805296
$df
[1] 1
$p.value2
[1] 0.3695146
```

```
> woolf(depsmok3)
Woolf test statistic
$X2
[1] 0.8040594
$d.f
[1] 1
$p.value
[1] 0.3698824
$estim.theta
[1] 2.490770
```

The common for males and females odds ratio is estimated to be 2.48 with a 95% confidence interval (1.78, 3.46), indicating that the odds of major depression is almost 2.5 times higher for smokers than for nonsmokers. But are we justified to assume that they share a common odds ratio? The `BDT()` and the `woolf()` functions, provided in the web appendix (see Sect. A.3.3), perform the Breslow–Day test (with and without Tarone's adjustment) and the Woolf test, respectively, for testing (3.7). We verify that the underlying association is homogeneous for males and females. The related output for this data set is provided in Table 3.4.

We shall revisit this example to model the structure of the underlying association based on log-linear models (Sect. 5.4.2).

Table 3.5 Cross-classification of patients according to treatment and the presence of a prognostic effect in six clinics

Clinic: A	Prognostic factor		Clinic: B	Prognostic factor		Clinic: C	Prognostic factor	
Treatment	Yes	No	Treatment	Yes	No	Treatment	Yes	No
Success	79	5	Success	89	4	Success	141	6
Failure	68	17	Failure	221	46	Failure	77	18
Clinic: D	Prognostic factor		Clinic: E	Prognostic factor		Clinic: F	Prognostic factor	
Treatment	Yes	No	Treatment	Yes	No	Treatment	Yes	No
Success	45	29	Success	81	3	Success	168	13
Failure	26	21	Failure	112	11	Failure	51	12

3.3.4 Example 3.3

Consider the following hypothetical data set, provided in Table 3.5. Patients from six different clinics are cross-classified according to treatment's outcome and the presence or not of a prognostic factor. Interest lies in testing the strength of influence of the prognostic factor on the treatment's outcome.

The homogeneity of the association between the prognostic factor and the treatment's outcome across the clinics will be tested first, applying function `BDT(dat)`, where `dat` is the data table, entered in `R` as

```
> dat <- array(c(79,68,5,17,89,221,4,46,141,77,6,18,45,26,29,21,81,
                 112,3,11,168,51,13,12), c(2,2,6))
> dimnames(dat) <- list(Treatment=c("Success","Failure"),
                        Prognostic_Factor=c("Yes","No"),
                        Clinic=c("A","B","C","D","E","F"))
```

Verify that the hypothesis of homogeneous association across the clinics cannot be rejected ($BDT = 7.91$, $df = 5$, p-value=0.161). On the basis of homogeneous underlying association and applying `mantelhaen.test(dat)`, the common odds ratio is estimated to be $\hat{\theta}_{MH} = 2.96$, which is high significantly different than 1 ($MH = 32.703$, $df = 1$, p-value=1.074e-08). Hence, the conditional independence of the prognostic factor and the treatment's outcome given the clinic is rejected (based on the Mantel–Haenszel test) and the odds of success is estimated to be about 3 times higher for patients with the prognostic factor than patients without, homogeneous across the clinics.

The corresponding test in the framework of log-linear models is discussed in Sect. 4.6.1.1 while the ML estimate of the common odds ratio is computed in the framework of generalized log-linear models for odds ratios in Sect. 5.6.2.

3.4 Types of Independence for Three-way Tables

Let (n_{ijk}) be an $I \times J \times K$ contingency table of observed frequencies with row, column, and layer classification variables X, Y, and Z, respectively. The variables X, Y, and Z will be *independent* if and only if

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}, \quad i = 1,\ldots,I,\ j = 1,\ldots,J,\ k = 1,\ldots,K, \tag{3.16}$$

where π_{ijk} are the cell probabilities (3.1), while the marginal probabilities π_{i++}, π_{+j+}, and π_{++k} are also defined in Sect. 3.2.1.

If Y is *jointly* independent from X and Z (without these two being necessarily independent), then

$$\pi_{ijk} = \pi_{+j+}\pi_{i+k}, \quad i = 1,\ldots,I,\ j = 1,\ldots,J,\ k = 1,\ldots,K. \tag{3.17}$$

There are two more possible hypotheses of this type, since (3.17) can be expressed in a symmetric way for X or Z being jointly independent from the remaining two variables.

Finally, we shall see how the concepts of *conditional* and *marginal* independence of two variables X and Y over a third one Z, discussed in Sect. 3.3 in the framework of stratified 2×2 tables, extend to general $I \times J \times K$ tables.

Under multinomial sampling scheme, the joint probabilities of the three-way table cells $\pi_{ijk} = \mathsf{P}(X = i,\ Y = j,\ Z = k)$ can be expressed in terms of conditional probabilities as

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

which under conditional independence of X and Y given Z equals

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

Incorporating the standard notation we use for marginal probabilities (see also, Sect. 3.2.1), we conclude that variables X and Y are independent *conditionally* on Z, if and only if

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}, \quad i = 1,\ldots,I,\ j = 1,\ldots,J,\ k = 1,\ldots,K\ . \tag{3.18}$$

The analysis above assumed that Y is a response variable. The conditioning approach with X as response variable would also lead to (3.18), which is symmetric in terms of X and Y. The hypotheses of conditional independence of X and Z given Y, and of Y and Z, given X, are formed analogously to (3.18).

Under conditional independence of X and Y, given Z, the conditional XZ local odds ratios are

$$\theta^{XZ}_{i(j)k} = \frac{\pi_{ijk}\pi_{i+1,j,k+1}}{\pi_{i+1,j,k}\pi_{i,j,k+1}}, \quad i = 1,\ldots,I-1,\ k = 1,\ldots,K-1\ ,$$

for any $j = 1,\ldots,J$, and the marginal XZ local odds ratios

$$\theta^{XZ}_{ik} = \frac{\pi_{i+k}\ \pi_{i+1,+,k+1}}{\pi_{i+1,+,k}\ \pi_{i,+,k+1}}, \quad i = 1,\ldots,I-1,\ k = 1,\ldots,K-1\ .$$

Furthermore, by (3.18)

$$\theta^{XZ}_{i(j)k} = \frac{\frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}} \cdot \frac{\pi_{i+1,+,k+1}\pi_{+,j,k+1}}{\pi_{+,+,k+1}}}{\frac{\pi_{i+1,+,k}\pi_{+jk}}{\pi_{++k}} \cdot \frac{\pi_{i,+,k+1}\pi_{+,j,k+1}}{\pi_{+,+,k+1}}} = \frac{\pi_{i+k}\ \pi_{i+1,+,k+1}}{\pi_{i+1,+,k}\ \pi_{i,+,k+1}}\ ,$$

i.e.,

$$\theta^{XZ}_{i(j)k} = \theta^{XZ}_{ik}, \quad i = 1,\ldots,I-1,\ k = 1,\ldots,K-1,\ j = 1,\ldots,J\ . \tag{3.19}$$

Analogously, it can be proved that under (3.18) also

$$\theta^{YZ}_{(i)jk} = \theta^{YZ}_{jk}, \quad i = 1,\ldots,I,\ k = 1,\ldots,K-1,\ j = 1,\ldots,J-1\ . \tag{3.20}$$

Thus, when X and Y are conditionally independent given Z, then the XZ (as well as the YZ) marginal and conditional associations coincide. However, this is not the case for the XY marginal and conditional association.

For a three-way table, the *marginal* independence of X and Y is defined as

$$\pi_{ij+} = \pi_{i++} \cdot \pi_{+j+}\ \forall i,\ j\ . \tag{3.21}$$

Under conditional independence of X and Y, given Z, summing (3.18) over k, it holds

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

which does not simplify to the condition of marginal independence (3.21), confirming thus that *conditional independence does not imply marginal independence*. This fact is actually known from the properties of multivariate distributions.

Marginal independence (3.21) can be tested by the test of independence presented in Sect. 2.2.2 applied on the corresponding two-way marginal table. Hypotheses (3.16) through (3.18) could be tested analogously. These tests will not be presented here, since they can equivalently be treated in the context of *log-linear models* in Chap. 4 by fitting the associated model.

3.5 Graphs for Multi-way Contingency Tables

The graphs presented for two-way tables in Sect. 2.4 serve also for picturing association structures in multi-way tables. Marginal and conditional associations can be visualized, clarifying thus issues regarding their relationship.

All graphical displays for multi-way contingency tables share a common basis. They represent a multi-way table in the two-dimensional space, using areas to represent frequencies. As for two-way tables, they can represent the sample structure (applied on the observed frequency table), the expected structure under an assumed model (applied on the ML estimates), or the deviation from a model (applied on residuals). At this stage, we will illustrate sample structures by the graphs while model-based graphs will be presented in Chap. 4.

3.5.1 Fourfold Plots for $2 \times 2 \times K$ Tables

The fourfold plot, for visualizing odds ratios in 2×2 tables (Sect. 2.4.2), can be constructed also for stratified 2×2 tables. In the framework of the `R` package `vcd`, for example, the fourfold plots of Example 3.3 can be constructed by

```
> fourfoldplot(dat, color=c("#CCCCCC", "#999999"), mfcol=c(2,3))
```

This leads to the plots in Fig. 3.1.

Observing these fourfold plots, we conclude for a positive association between prognostic factor and treatment's response for all clinics, expecting the strongest to be in Clinic 3 and the weakest in Clinic 4. Furthermore, for Clinics 4 and 5, the sample estimates support the null hypothesis of no association, since the 95 % confidence rings for adjacent quartiles do overlap. These rings are calculated for the odds ratios in each stratum and are not adjusted for multiple testing. This plot assessment is confirmed by the individual asymptotic tests for the conditional odds ratios $\theta_{(k)}$, $k = 1, \ldots, 6$. The corresponding sample conditional odds ratios are $\hat{\theta}_{(1)} = 3.95$, $\hat{\theta}_{(2)} = 4.63$, $\hat{\theta}_{(3)} = 5.49$, $\hat{\theta}_{(4)} = 1.25$, $\hat{\theta}_{(5)} = 2.65$, and $\hat{\theta}_{(6)} = 3.04$, with $\max(\hat{\theta}_{(k)}) = \hat{\theta}_{(3)}$ and $\min(\hat{\theta}_{(k)}) = \hat{\theta}_{(4)}$. These sample conditional odds ratios along with their asymptotic tests of significance and confidence intervals can be calculated in `R` applying function `odds.ratio()` on each partial table. For example, the sample estimate and the associated asymptotic inferential results for $\theta_{(3)}$ are delivered by

```
> odds.ratio(dat[,,3])
```

Analogously, the fourfold plots for data in Table 3.1 (Example 3.1) can be derived and compared to the plot of Fig. 2.4 that corresponds to the marginal table (over gender) of this data set.

Furthermore for stratified $I \times J$ tables, conditional or marginal generalized odds ratios can be visualized by fourfold plots, applying the procedure described in Sect. 2.4.2 for $I \times J$ tables on each partial table or on the marginal table. For example, the conditional local odds ratios of Table 3.3 can be visualized by function `ffold.local()` of the web appendix (see Sect. A.3.2).

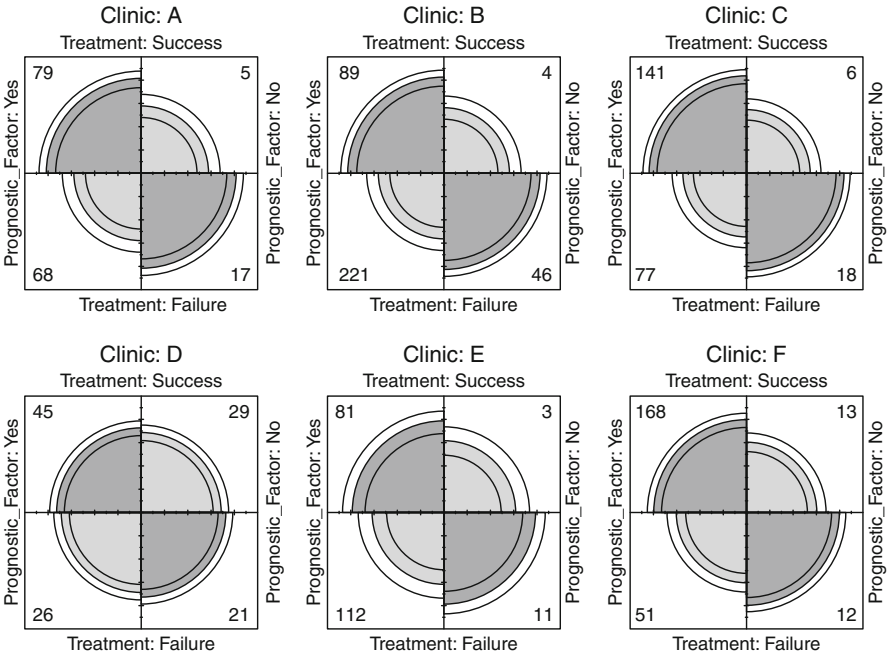


Fig. 3.1 Fourfold plots for the conditional odds ratios of Example 3.3 (Table 3.5)

3.5.2 Sieve Diagrams for Multi-way Tables

The sieve diagram for a multi-way table is an easy and direct adjustment of the two-way sieve diagram. The starting point is the sieve diagram for a two-way marginal table of the initial multi-way table, corresponding to variables, say X_1 and X_2. Then, a variable X_3 is added on the graph by subdividing the rectangles corresponding to the marginal horizontally to show how the counts of this cell are further classified by X_3. In the same way, these derived rectangulars are further split vertically by the categories of the next variable X_4 and the procedure continues till all variables are represented on the graph.

For Example 3.2, the data (Table 3.2) are already in R matrix `party.tab` (see Sect. 3.2.4). Before constructing the sieve diagram, we add labels to the variables and their categories, to make the display easier to be read:

```
> dimnames(party.tab) <- list(D=c("LT H.Sc", "H.Sc", "J.College",
+         "Bachelor", "Graduate"), P=c("1", "2", "3", "4", "5",
+         "6", "7"), G=c("M", "F"))
```

The sieve diagram is then derived in R package `cvd` by

```
> sieve(party.tab)
```

and presented in Fig. 3.2 (left). Note that the education level of the responders is given in the rows, subclassified by gender, and their party affiliation in columns.

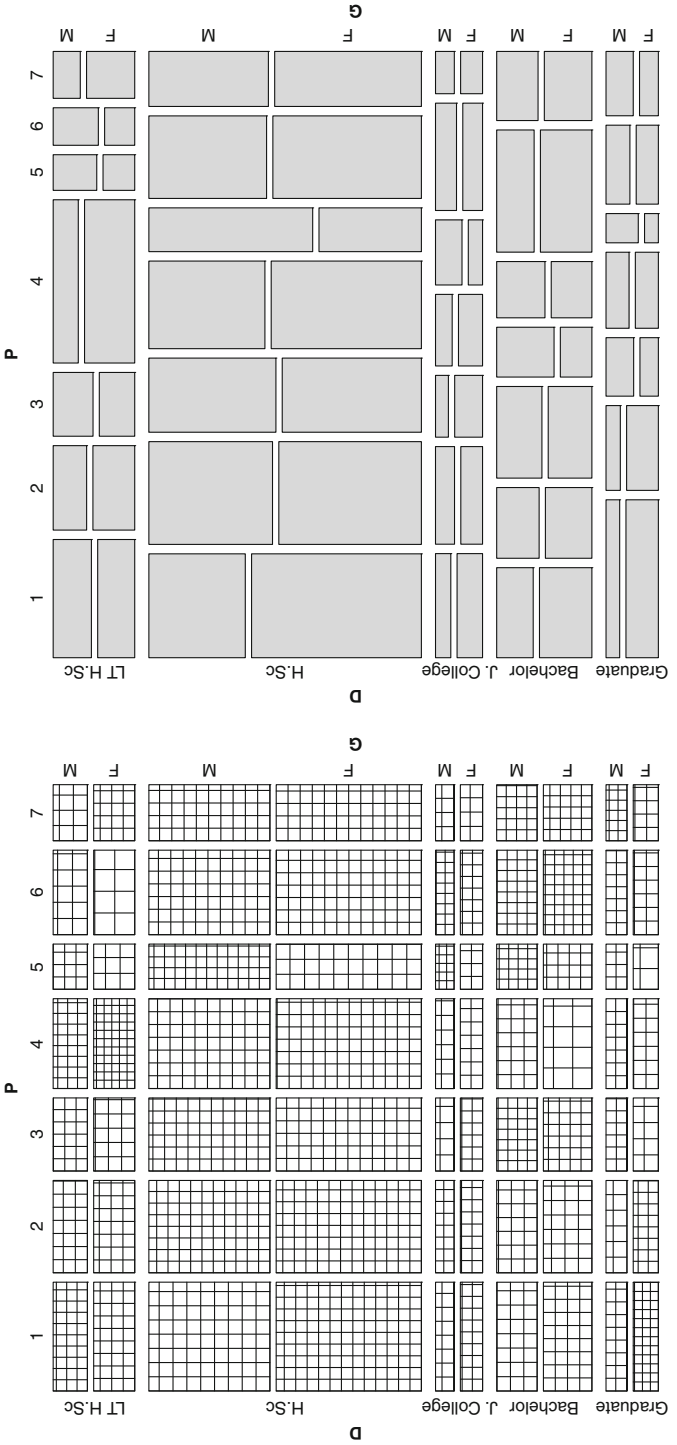


Fig. 3.2 Sieve diagram (*left*) and mosaic plot (*right*) of the observed frequencies for Example 3.2 (Table 3.2)

Each squared area corresponds to a cell of the table while the number of square in each area equals the corresponding observed cell frequency. Thus, cells or areas of the table of high or low frequencies are easily recognized.

3.5.3 *Mosaic Plots for Multi-way Tables*

The mosaic plots are extended to multi-way tables, the same way as the sieve diagrams. The mosaic plot for Example 3.2 is given in Fig. 3.2 (right) and is derived in `cvd` package of `R` by the function

```
> mosaic(party.tab)
```

Since the educational level variable is subclassified by gender, the rectangular areas corresponding to males and females for a specific educational level show the proportion of males and females in this educational level.

We shall reconsider mosaic plots in Chap. 4, when fitting models on the data. The most interesting mosaic plots for multi-way tables arise when visualizing the residuals corresponding to a particular model by colors or shadings. The location of the deviations of the observed frequencies from the expected under the assumed model may clarify the source of bad fit and guide to a model that represents better the data.

3.6 Overview and Further Reading

3.6.1 *Stratified* 2×2 *Contingency Tables*

Different Mantel–Haenszel type estimators for the common odds ratio of $2 \times 2 \times K$ contingency tables have been compared by Hauck (1984). Confidence intervals for the common odds ratio of stratified 2×2 tables, also based on the mid-p-value, are presented and compared in Mehta and Walsh (1992).

Conditional independence and homogeneous association in $2 \times 2 \times K$ tables have been considered so far in a non-model-based fashion. These issues will be reconsidered in terms of log-linear and logit models in Sects. 4.6.1.1 and 8.2, respectively, as well as via modeling odds ratios by the generalized log-linear model in Sect. 5.6.2, which will also provide the MLE of the expected common odds ratio value under the homogeneous association assumption.

An alternative approach for testing that K stratified 2×2 tables share a common odds ratio uses a random effects model to describe heterogeneity of the odds ratios. Consider the null hypothesis H_0 of homogeneous odds ratios (3.7). Then, assuming that the log odds ratios for the stratified tables $\log \theta_k$, $k = 1, \ldots, K$, are independent realizations of a random variable with mean $\log \theta$ and variance σ_θ^2, H_0 is equivalent to testing that the variance σ_θ^2 is zero. The likelihood function of the mixture

model for the parameters $\log\theta$ and σ_θ^2 has been approximated by Liu and Pierce (1993) by expanding the integrand in a Taylor series about its maximizing value using the Laplace's method. A similar approximation has been provided by Cox (1983), who expanded the integrand about the true mean of the random effect $\log\theta$. Based on Cox's approximation, Liang and Self (1985) proposed a score statistic for testing the hypothesis of equality of the odds ratios. The associated standardized test is asymptotically normal distributed and is of good performance in sparse table situations, as shown by simulation studies on its size (Jones et al. 1989). However, a skewness problem was detected in cases with K relative small and large θ values. Davison (1992) considered a simpler statistic, by eliminating a specific term of the score statistic, which has a skew distribution, approximated by a gamma distribution. The adequacy of the corrected-for-skewness score method is verified by numerical results of Gart and Nam (1988), who derived, for moderate sample sizes, confidence limits for the common (under the null hypothesis) odds ratio based on the Cornish–Fisher corrected score statistic of the unconditional likelihood.

3.6.2 *Generalized Mantel–Haenszel Test for $I \times J \times K$ Contingency Tables*

The T_{MH} statistic has been generalized to $I \times J \times K$ tables by Mantel (1963), Birch (1965), and Mantel and Byar (1978). In this case the null hypothesis of unit odds ratios across the stratified tables is extended to the null hypothesis of independence for the K $I \times J$ partial tables, conditional on the row and column marginals. The generalization of the T_{MH} statistic is

$$T_{GMH} = (\mathbf{n} - \mathbf{m})^T \mathbf{V}^{-1} (\mathbf{n} - \mathbf{m}) \,,$$

where $\mathbf{n} = \sum_{k=1}^{K} \mathbf{n}_k$, $\mathbf{m} = \sum_{k=1}^{K} \mathbf{m}_k$, and $\mathbf{V} = \sum_{k=1}^{K} \mathbf{V}_k$, with $\mathbf{n}_k$ the $(I-1)(J-1) \times 1$ vector of nonredundant cell frequencies of the kth partial table, i.e., $\mathbf{n}_k = (n_{11k}, \ldots, n_{1,J-1,k}, n_{21k}, \ldots, n_{I-1,J-1,k})^T$, $\mathbf{m}_k = E(\mathbf{n}_k) = (\frac{n_{1+k}n_{+1k}}{n_{++k}}, \ldots, \frac{n_{I-1,+,k}n_{+,J-1,k}}{n_{++k}})^T$ under the H_0 of conditional independence and $\mathbf{V}_k$ the $(I-1)(J-1) \times (I-1)(J-1)$ covariance matrix of $\mathbf{n}$ under H_0 with elements

$$\mathsf{Cov}(n_{i_1 j_1 k}, n_{i_2 j_2 k}) = \frac{n_{i_1+k}(\delta_{i_1 i_2} n_{++k} - n_{i_2+k}) n_{+j_1 k}(\delta_{j_1 j_2} n_{++k} - n_{+j_2 k})}{n_{++k}^2 (n_{++k} - 1)} \,,$$

where $\delta_{\rho\xi} = 1$ for $\rho = \xi$ and 0 otherwise. Under H_0, T_{GMH} is asymptotically $\mathscr{X}^2_{(I-1)(J-1)}$ distributed.

The T_{GMH} statistic assumes that the classification variables X and Y are both nominal. If they are ordinal, then appropriate is also the test statistic

$$M_K^2 = \frac{\left(\sum_k \left[\sum_{i,j} x_i y_j n_{ijk} - \mathrm{E}\left(\sum_{i,j} x_i y_j n_{ijk}\right)\right]\right)^2}{\sum_k \mathsf{Var}\left(\sum_{i,j} x_i y_j n_{ijk}\right)},$$

where x_i, $i = 1,\ldots,I$, and y_j, $j = 1,\ldots,J$, are the row (X) and column (Y) scores, respectively (Mantel 1963). The asymptotic distribution for M_k^2 under H_0 is $\mathcal{X}_1^2$. For $K = 1$, the statistic M_K^2 is the linear trend statistic (2.57) of the two-way contingency tables. Landis et al. (1978) presented a generalized test statistic having T_{GMH} and M_K^2 as special cases. This statistic applies also to stratified tables with one classification variable nominal and the other ordinal (see Agresti 2013, p. 317–319, 328). Related is also the work by Goodman (1969), who estimated the degree of partial association in K stratified $I \times J$ tables and proceeded to ML estimation, conditional (on the row and column marginals) and unconditional.

3.6.3 Visualization of Categorical Data

The fourfold plots for a 2×2 contingency table were first introduced by Fienberg (1975) and further developed (also for stratified 2×2 tables) by Friendly (1994, 1995). The sieve diagrams were proposed by Riedwyl and Schüpbach (1983, 1994). Mosaic plots for contingency tables in their current form were introduced by Hartigan and Kleiner (1981, 1984) and expanded by Friendly (1994, 1999) while mosaic type displays had already been used by the early 1800s. For a historical review and bibliography on mosaic plots, we refer to Friendly (2002).

Originally, the sieve diagrams and the mosaic plots had been considered for illustrating the fit of independence or visualizing departures from it. Their concept however applies to other hypotheses (or models). Thus they can be used to reveal the departure from the structure dictated by the assumed model, which can be in the GLM family (see Chap. 5) or not, since the underlying concept holds for a general model. In `R`, plots for GLM or also generalized nonlinear models (GNM) can be obtained in package `vcdExtra`, an extension of `vcd` for models fitted by `glm()` and `gnm()`. The package `gnm()` will be illustrated in Chap. 6, where we shall consider some special models, nonlinear in their parameters. A tutorial on `vcd` and `vcdExtra` is provided by Friendly (2013) while Meyer et al. (2006) exhibit the visualization of multi-way tables in `vcd`. Package `vcdExtra` offers also the feature of three-dimensional plots.

Graphical displays for categorical data (also for multivariate) are discussed in detail by Friendly (2000). Furthermore, interesting related contributions are to be found in the volume *Visualization of Categorical Data*, edited by Blasius and Greenacre (Academic Press, 1998).

Chapter 4
Log-Linear Models

Abstract The classical log-linear models are introduced for two-way and multi-way contingency tables. Estimation theory, goodness-of-fit testing, and model selection procedures are discussed. Characteristic examples are worked out in R and interpreted. Log-linear models for three-dimensional tables are illustrated through mosaic plots. Graphical models are shortly discussed. Finally the collapsibility in multi-way tables, in connection to Simpson's paradox, is addressed.

Keywords Hierarchical log-linear models • Model fit and selection • Dissimilarity index • Graphical models • Simpson's paradox

4.1 Log-Linear Models for Two-way Tables

4.1.1 Model of Independence

Independence (2.34) between the classification variables X and Y can equivalently be expressed in terms of the expected under independence cell frequencies m_{ij} in a *log-linear model* form as

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y \,, \qquad i = 1,\ldots,I,\ j = 1,\ldots,J \,, \tag{4.1}$$

where λ corresponds to the overall mean while λ_i^X, λ_j^Y are the ith row and jth column main (or marginal) effects, respectively.

Model (4.1) could equivalently be expressed in terms of the expected under the assumed model probabilities π_{ij}. The usual choice is in terms of m_{ij}, because expected cell frequencies are common for the different sampling schemes while the underlying probability structure changes (see Sect. 2.2.1). For this, all log-linear models considered in the sequel will be expressed in terms of expected cell frequencies.

M. Kateri, *Contingency Table Analysis: Methods and Implementation Using R*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4811-4_4,

Interpretation is carried out in terms of the odds. For given column category j, under model (4.1), the odds of being in row i_1 instead of row i_2 ($i_1 \neq i_2$), $i_1, i_2 = 1, \ldots, I$, is

$$\frac{m_{i_1 j}}{m_{i_2 j}} = \frac{\exp(\lambda + \lambda_{i_1}^X + \lambda_j^Y)}{\exp(\lambda + \lambda_{i_2}^X + \lambda_j^Y)} = \exp(\lambda_{i_1}^X - \lambda_{i_2}^X) \,, \quad j = 1, \ldots, J \,, \tag{4.2}$$

independent of j. Similarly, for columns j_1 and j_2 ($j_1 \neq j_2$, $j_1, j_2 = 1, \ldots, J$),

$$\frac{m_{i j_1}}{m_{i j_2}} = \exp(\lambda_{j_1}^Y - \lambda_{j_2}^Y) \,, \quad i = 1, \ldots, I \,, \tag{4.3}$$

i.e., the odds of being in column j_1 instead of j_2 is determined only by the distance of the corresponding column main effect values and is independent of i. By (4.3), the conditional j_1 and j_2 column probabilities (within row i)

$$\frac{\mathsf{P}(Y = j_1 | X = i)}{\mathsf{P}(Y = j_2 | X = i)} = \exp(\lambda_{j_1}^Y - \lambda_{j_2}^Y) \,, \quad i = 1, \ldots, I \,,$$

relate the same for all rows and this is true for any pair of columns j_1 and j_2. Thus, the conditional column distribution is the same for all rows, as should be for independent X and Y.

Using (4.3), the expected under independence local odds ratios are

$$\theta_{ij}^L = \frac{m_{ij}/m_{i.j+1}}{m_{i+1.j}/m_{i+1.j+1}} = \frac{e^{\lambda_j^Y - \lambda_{j+1}^Y}}{e^{\lambda_j^Y - \lambda_{j+1}^Y}} = 1 \,, \quad i = 1, \ldots, I-1, \; j = 1, \ldots, J-1,$$

i.e., all equal to 1, as expected by (2.52).

The parameters in model (4.1) are $1 + I + J$ while we know that under independence the parameters are $(I-1) + (J-1)$. Hence, parameters in (4.1) are not uniquely determined unless constraints are imposed on the main effects. The traditionally used identifiability constraints are the sum to zero constraints:

$$\sum_{i=1}^{I} \lambda_i^X = \sum_{j=1}^{J} \lambda_j^Y = 0 \,. \tag{4.4}$$

Due to computational convenience, software applications replace (4.4) by the constraints that set a category effect to zero, usually the last ($\lambda_I^X = \lambda_J^Y = 0$) or the first ($\lambda_1^X = \lambda_1^Y = 0$).

The different set of constraints are equivalent and they affect only the reference point for physical interpretation. Thus, λ_i^X compares the ith row category to the overall mean or to the first category, depending on whether model (4.1) is fitted under (4.4) or under $\lambda_1^X = 0$. The differences $\lambda_{i_1}^X - \lambda_{i_2}^X$ and $\lambda_{j_1}^Y - \lambda_{j_2}^Y$ are constraints invariant; thus, comparisons between categories are not affected by the identifiability constraints used.

Model (4.1) will be illustrated in Sect. 4.2.1, after we discuss technical matters on parameter estimation and model fit checking.

4.1.2 The Saturated Model

In case the classification variables X and Y are *not* independent, their interaction is significant and the corresponding *XY-interaction term* has to be added in the log-linear model expression, leading to the *saturated* model

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \ , \ \ i = 1, \ldots, I, \ j = 1, \ldots, J \ . \tag{4.5}$$

Identifiability constraints are also required for model (4.5). Under the sum to zero identifiability constraints, additional to (4.4) the following constraints hold for the interaction parameters:

$$\sum_{i=1}^{I} \lambda_{ij}^{XY} = \sum_{j=1}^{J} \lambda_{ij}^{XY} = 0 \ . \tag{4.6}$$

Analogous to model (4.1), the (4.4) and (4.6) constraints can be equivalently replaced by constraints equating the last (or first) row and column parameters to zero. For the interaction parameters this would be

$$\lambda_{Ij}^{XY} = \lambda_{iJ}^{XY} = 0, \ \ i = 1, \ldots I - 1, \ \ j = 1, \ldots J - 1$$

(or $\lambda_{1j}^{XY} = \lambda_{i1}^{XY} = 0$, for $i = 2, \ldots I$, $j = 2, \ldots J$).

The saturated model (4.5), under (4.4) and (4.6), has as many parameters as the number of cells, i.e., IJ. Thus, it does not impose any structure on the underlying association. It just reparametrizes the table's cells in an interpretational meaningful way. The local odds ratios are directly derived from the interaction parameters, since

$$\log \theta_{ij}^{L} = \lambda_{ij}^{XY} + \lambda_{i+1,j+1}^{XY} - \lambda_{i+1,j}^{XY} - \lambda_{i,j+1}^{XY} \ , \tag{4.7}$$
$$i = 1, \ldots, I-1, \ \ j = 1, \ldots, J-1 \ .$$

For a simple 2×2 table and for the first category set to zero constraints ($\lambda_{11}^{XY} = \lambda_{12}^{XY} = \lambda_{21}^{XY} = 0$), it holds

$$\log \theta = \lambda_{22}^{XY} \ .$$

Evidently, the λ^{XY} term indeed expresses the association between X and Y. Furthermore, model (4.1) is derived by (4.7), setting

$$\lambda_{ij}^{XY} = 0 \ , \ \ i = 1, \ldots, I, \ j = 1, \ldots, J \ , \tag{4.8}$$

i.e., by eliminating the association between X and Y. This means that (4.1) is *nested* in (4.7). We shall refer in detail to nested models in the context of log-linear models for multi-way tables in Sect. 4.4.

An example of the saturated model's implementation in practice is provided in Sect. 4.2.2.

Overall, log-linear models describe the way the involved categorical variables and their association (if significant) influence the count at each of the IJ cells of the cross-classification of these variables. They are the discrete analogue of analysis of variance, where for each cell of the cross-classification, there is modeled the mean of a continuous variable, instead of a count. The analogy to classical analysis of variance is obvious once the log-linear model's parameters, subject to the sum to zero constraints (4.4) and (4.6), are identified in terms of expected cell frequencies:

$$\lambda = \frac{1}{IJ}\sum_{i,j}\log m_{ij} \tag{4.9}$$

$$\lambda_i^X = \frac{1}{J}\sum_{j}\log m_{ij} - \lambda\ , \quad i = 1,\ldots,I\ , \tag{4.10}$$

$$\lambda_j^Y = \frac{1}{I}\sum_{i}\log m_{ij} - \lambda\ , \quad j = 1,\ldots,J\ , \tag{4.11}$$

$$\lambda_{ij}^{XY} = \log m_{ij} - \lambda - \lambda_i^X - \lambda_j^Y\ , \quad i = 1,\ldots,I,\ \ j = 1,\ldots,J\ . \tag{4.12}$$

4.2 On Inference and Fit of Log-Linear Models

We have seen in Sect. 2.2.1 that the three common sampling schemes for contingency tables are inferential equivalent. For this, the ML estimates of the expected cell frequencies m_{ij} under a log-linear model can be equivalently derived under any of these sampling assumption. For simplicity reasons, the Poisson log-likelihood function is usually considered. Assuming thus an independent Poisson distribution for each cell, $N_{ij} \sim \mathscr{P}(m_{ij})$, and upon observing a sample table $(n_{ij})_{I\times J}$, the Poisson log-likelihood *kernel* ℓ (ignoring the constants) is

$$\ell = \sum_{i,j}\left(n_{ij}\log m_{ij} - e^{\log m_{ij}}\right)\ . \tag{4.13}$$

Under a particular log-linear model assumption, substituting $\log m_{ij}$ in (4.13) by the model's formula, ℓ will be a function of the log-linear models parameters. Maximizing (4.13) with respect to these parameters, the sets of corresponding

likelihood equations are derived. Their solution is the set of ML estimates of the parameters and consequently the ML estimates $\hat{m}_{ij}$ of the expected under this model cell frequencies.

Thus, for the independence model, substituting in (4.13) the $\log m_{ij}$ by (4.1) and maximizing with respect to λ_i^X and λ_j^Y, the sets of *likelihood equations* are derived, respectively, as follows:

$$\hat{m}_{i+} = n_{i+}\ ,\ \ i = 1,\ldots,I, \qquad \text{and} \qquad \hat{m}_{+j} = n_{+j}\ ,\ \ j = 1,\ldots,J. \tag{4.14}$$

Their solution is the ML estimates of the expected cell frequencies $\hat{m}_{ij}$, provided in (2.35). The ML estimates of the λ parameters in (4.1), under the sum to zero constraints (4.4), are

$$\hat{\lambda} = \frac{1}{I}\sum_s \log n_{s+} + \frac{1}{J}\sum_s \log n_{+s} - \log n \tag{4.15}$$

$$\hat{\lambda}_i^X = \log n_{i+} - \frac{1}{I}\sum_s \log n_{s+}\ ,\quad i = 1,\ldots,I\ , \tag{4.16}$$

$$\hat{\lambda}_j^Y = \log n_{+j} - \frac{1}{J}\sum_s \log n_{+s}\ ,\quad j = 1,\ldots,J\ , \tag{4.17}$$

and are obtained by (4.9)–(4.11), substituting the m_{ij}'s by the corresponding $\hat{m}_{ij}$'s.

The goodness of fit of a log-linear model is assessed asymptotically by the classical X^2 and G^2 test statistics, which are under the assumed model asymptotically $\mathscr{X}^2$ distributed with degrees of freedom (df) equal to the dimension of the sample space reduced by the number of the parameters estimated under the model. Note that the dimension of the sample space of a contingency table depends on the underlying sampling scheme. Thus, for an $I \times J$ table, for example, it is $IJ - 1$ if the table is derived by a multinomial distribution (total n is fixed), while it is IJ when independent Poisson distributions are considered for each cell (n is random). For this, the λ of a log-linear model is a parameter only under Poisson sampling (counting for n). Consequently, the df of the model are the same under both sampling schemes and the sampling schemes, given n are inferentially equivalent.

For the independence model (4.1), the X^2 and G^2 tests are (2.36) and (2.37), respectively, with $\hat{m}_{ij}$ given by (2.35) or by (4.1), with the parameters being substituted by their ML estimates (4.15)–(4.17). The saturated model (4.5) fits the data perfectly ($X^2 = G^2 = 0, df = 0$).

The classical goodness-of-fit tests X^2 and G^2 are sensitive in sample size n, as already mentioned in Sect. 2.2.2. It is evident that for large n, they tend to reject even "good" models. For this, in the framework of log-linear models and in cases of large sample size n, a *dissimilarity index* is used that assesses the *practical* significance of the assumed model's lack of fit. This index $\hat{\Delta}$ is common in social sciences applications where also cross-tabulations of large sample sizes occur and is defined as

$$\hat{\Delta} = \frac{1}{2n} \sum_{i=1}^{I} \sum_{j=1}^{J} |n_{ij} - \hat{m}_{ij}| = \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |p_{ij} - \hat{\pi}_{ij}| \tag{4.18}$$

The dissimilarity index $\hat{\Delta}$ ranges in the interval $[0, 1]$ and expresses the percentage of observations that have to be moved to different cells in order to achieve a perfect fit. Thus, small values of $\hat{\Delta}$ are indicative of good fit with $\hat{\Delta} < 0.02$ or < 0.03 being the limit for a satisfying representation of the data by the assumed model. The sample index $\hat{\Delta}$ estimates the corresponding population index

$$\Delta = \frac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |\pi_{ij} - \pi_{ij}^{*}| \,,$$

which measures the dissimilarity between the population probability distribution $\pi = (\pi_{ij})$ and the probability distribution under the assumed model $\pi^{*} = (\pi_{ij}^{*})$. The approximate variance of the statistic $\hat{\Delta}$ and the associated confidence interval has been given by Kuha and Firth (2011). They also provide an updated review of literature on $\hat{\Delta}$, which has a long history.

In practice, log-linear models for two-way (and multi-way) contingency tables are fitted very easily in any software. In `R`, there are several options for getting log-linear models analysis. They can be fitted by `loglin` (of `stats`) or `loglm` (of the `MASS` package). Log-linear models will be fitted for Examples 2.4 and 2.3 by `loglm` in Sects. 4.2.1 and 4.2.2, respectively. However, the predominant approach is to analyze log-linear models in the *generalized linear model* (GLM) framework. Thus, Example 2.4 will be revisited in Sect. 5.4.1, after discussing the GLM and its connection to log-linear models.

4.2.1 Example 2.4 (Continued)

The log-linear model of independence (4.1) will be fitted on Table 2.3 in `R`, by the `loglm` function of the package `MASS`. The parameter estimates derived by `loglm` are under the sum to zero constraints. The data can be either in matrix form or in a data frame.

The data of Table 2.3 are to be found in matrix `natfare`, constructed in Sect. 2.4.1.

After loading the `MASS` package, model (4.1) is then fitted by

```
> I.fit <- loglm( ~ WELFARE + DEGREE, data=natfare)
```

The model formula of the fitted model and the corresponding G^2 and X^2 goodness-of-fit tests is the standard output, obtained by

```
> I.fit
```

```
Call:
loglm(formula = ~ WELFARE + DEGREE, data = natfare)

Statistics::
                      X^2  df   P(> X^2)
Likelihood Ratio  10.36287   8  0.2404748
Pearson           10.52048   8  0.2303766
```

The goodness-of-fit tests above suggest not to reject the independence model. Thus we conclude that the respondents' belief about national funds for welfare does not depend significantly on their educational level. Recall that independence was visualized in the conditional barplot in Fig. 2.3, where the conditional distributions of educational levels within each category of opinion about welfare spending were similar.

Naturally, we derived the same Pearson's X^2 as in Sect. 2.2.6 by the classical `chisq()`. However, in the log-linear models framework, a more detailed interpretation can be extracted by the parameter estimates $\hat{\lambda}_i^X$ and $\hat{\lambda}_j^Y$ in means of (4.2) and (4.3), respectively. All items saved in object `I.fit` can be viewed by `names(I.fit)` and we verify that the parameters' ML estimates, satisfying the sum to zero constraints (4.4), are saved in `I.fit` under `$param`. They can be printed by

```
> I.fit$param
```

```
'(Intercept)'
[1] 3.923732

$WELFARE
too little   about right     too much
-0.2254279     0.1041910    0.1212369

$DEGREE
     LT HS            HS        JColg           BA         Grad
-0.1517607     1.1139057   -0.5802153    0.1528232   -0.5347529
```

Alternatively, they can be saved in new vectors, convenient for further use, like

```
> L <- I.fit$param[1]     # λ̂
> L.X <- I.fit$param[2:4] # (λ̂_1^X, λ̂_2^X, λ̂_3^X)
> L.Y <- I.fit$param[5:9] # (λ̂_1^y, ..., λ̂_5^y)
```

Thus, it is estimated that in year 2008, it was 1.4 times more probable a responder to believe that the national welfare spending was too much than that it was too little, independent of his educational level, since

$$\frac{\hat{m}_{3j}}{\hat{m}_{1j}} = \exp(\hat{\lambda}_3^X - \hat{\lambda}_1^X) = e^{0.1212-(-0.2254)} = e^{0.347} = 1.41\ , \quad j = 1,\ldots,5\ ,$$

which is computed by

```
> exp(L.X [3]-L.X[1])
```

The ML estimates of the expected under independence cell frequencies are derived by

```
> fitted(I.fit)
```

	DEGREE				
WELFARE	LT HS	HS	JColg	BA	Grad
too little	34.69319	123.0031	22.60314	47.04607	23.65445
about right	48.23874	171.0283	31.42827	65.41466	32.89005
too much	49.06806	173.9686	31.96859	66.53927	33.45550

The dissimilarity index $\hat{\Delta}$ can now be easily calculated as

```
> D <- sum(abs(natfare-fitted(I.fit)))/(2*sum(natfare))
```

and we find that $\hat{\Delta} = 0.038$, stating that 3.8 % of the observations have to be moved to achieve a perfect fit.

The Pearsonian residuals are given by

```
> residuals(I.fit)
```

	DEGREE				
WELFARE	LT HS	HS	JColg	BA	Grad
too little	1.6724517	-0.6375826	-0.7794780	0.1386103	-0.1351891
about right	-1.2226377	-0.3092454	0.2780712	0.3175824	1.3612691
too much	-0.2973437	0.8277514	0.3555753	-0.4378190	-1.3419302

but there is no option in the `loglm` framework for getting the standardized residuals. The log-linear models can also be fitted in the GLM framework by `glm`, where the derived output is more informative (for example, the standard errors and significance of the parameters' ML estimates are also provided) and more options are available (standardized residuals calculation is one of them). This example is treated by `glm` in Sect. 5.4.1.

Function `loglm` applies also on a data frame. To construct the data frame for this example, the row and column factors, `WELFARE` and `DEGREE`, respectively, are defined and tied to the vector of observed frequencies `freq` in a data frame, named `nf.frame`, as shown below. The factors are defined for a frequency vector of length $IJ = 15$ that expands the cells of the table by rows.

```
> NI <- 3
> NJ <- 5
> row.lb <- c("too little","about right","too much")
> col.lb <- c("LT HS","HS", "JColg","BA", "Grad")
> WELFARE <- gl(NI,NJ,length=NI*NJ, labels=row.lb)
> DEGREE <- gl(NJ,1,length=NI*NJ, labels=col.lb)
> nt.frame <- data.frame(freq,WELFARE,DEGREE)
```

Then, the model is fitted as

```
> I.fit <- loglm( freq ~ WELFARE + DEGREE, data=nt.frame)
```

leading to the same output and options as described above.

4.2.2 Example 2.3 (Continued)

We have already seen in Sect. 2.2.3 that the independence hypothesis is rejected for the cross-classification in Table 2.2 of responders (in GSS2008) subject to their gender and confidence in banks and financial institutions. In the log-linear models framework, model (4.1) is fitted by

```
> I.fit <- loglm( ~ Gender + Conf, data=confinan)
```

giving the fit statistics that we already know from Sect. 2.2.3

```
> I.fit
```

```
Call:
loglm(formula = ~ Gender + Conf, data = confinan)

Statistics::
                        X^2  df      P(> X^2)
Likelihood Ratio   16.39847   2   .0002748643
Pearson            16.34136   2   .0002828258
```

Hence, the interaction between gender and confidence in banks is significant. The interaction between two variables X and Y is denoted in R by `X:Y`. Entering the term `Gender:Conf` in the model above, the saturated model is achieved

```
> sat.fit <- loglm( ~ Gender + Conf + Gender:Conf, data=confinan)
```

with $G^2 = X^2 = 0$ and $df = 0$ (perfect fit). Though no structure is imposed on the underlying probability table gaining in parsimony, the parameters are still informative for interpretational purposes. We get

```
> sat.fit$param
```

```
'(Intercept)'
[1] 5.260226

$Gender
      males     females
-0.09028955  0.09028955

$Conf
   great deal     only some     hardly any
 -0.4147691     0.7337607     -0.3189915

$Gender.Conf
                $Conf
Gender          great deal     only some     hardly any
  males        -0.1701995   -0.009293922     0.1794934
  females       0.1701995    0.009293922    -0.1794934
```

In log-linear models, only the highest factor interaction parameters are interpreted. Thus, in presence of λ^{XY}, the main effects are not interpreted. Odds ratios can be calculated by (4.7) and corresponding conclusions can be expressed. Thus, based on the λ^{XY} values of the output above, the odds of having hardly any instead of great confidence to banks is 2.01 times higher for men than for women, computed by

```
> L.XY <- sat.fit$param$Gender.Confinan
> 1/exp(L.XY[1,1]+L.XY[2,3]-L.XY[1,3]-L.XY[2,1])
[1] 2.012516
```

4.3 Log-Linear Models for Three-way Contingency Tables

Consider a three-way contingency table, cross-classifying the variables X, Y, and Z. In Sect. 3.2 we discussed on conditional and marginal distributions of such tables and their relations, preparing the field to introduce the various notions of independence in Sect. 3.4.

The hypothesis of *complete independence* of X, Y, and Z (or mutual independence), defined by (3.16), is equivalently expressed in log-scale as

$$\log \pi_{ijk} = \log \pi_{i++} + \log \pi_{+j+} + \log \pi_{++k}, \quad i = 1,\ldots,I,\ j = 1,\ldots,J,\ k = 1,\ldots,K,$$

which indicates that the logarithmic model of complete independence is

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z, \quad i = 1,\ldots,I,\ j = 1,\ldots,J,\ k = 1,\ldots,K, \tag{4.19}$$

with the main effect parameters λ_i^X, λ_j^Y, and λ_k^Z satisfying identifiability constrains as the main effects of the log-linear models for two-way tables, i.e.,

$$\sum_{i=1}^{I} \lambda_i^X = \sum_{j=1}^{J} \lambda_j^Y = \sum_{k=1}^{K} \lambda_k^Z = 0 \quad \text{or} \quad \lambda_1^X = \lambda_1^Y = \lambda_1^Z = 0 \tag{4.20}$$

Analogously, hypothesis (3.17) of *joint independence* of Y from X and Z is in log-scale equivalent to model

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}, \quad \forall\, i,j,k. \tag{4.21}$$

Additionally to constraints (4.20), the parameters of model (4.21) satisfy the identifiability constraints

$$\sum_{i=1}^{I} \lambda_{ik}^{XZ} = \sum_{k=1}^{K} \lambda_{ik}^{XZ} = 0 \quad \text{or} \quad \lambda_{1k}^{XZ} = \lambda_{i1}^{XZ} = 0, \tag{4.22}$$

for all possible values of the non-summing subscript (k or i).

The model of joint independence (4.21) involves only one two-factor interaction term, the λ^{XZ}, since Y is joint independent from X and Z, but X and Z can be dependent to each other. Obviously, on a three-way table two more models of joint independence can be defined, those having as single two-factor interaction the λ^{XY} or the λ^{YZ} term.

If X and Y are independent *conditionally* on Z, then the underlying probabilities structure is captured in (3.18) as

$$\pi_{ijk} = \pi_{ij|k}\pi_{++k} = \pi_{i+|k}\pi_{+j|k}\pi_{++k} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}},$$

which is equivalent to the log-linear model:

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}, \quad \forall\, i,j,k. \tag{4.23}$$

The identifiability constraints of this model are (4.20), (4.22), and

$$\sum_{j=1}^{J} \lambda_{jk}^{YZ} = \sum_{k=1}^{K} \lambda_{jk}^{YZ} = 0 \quad \text{or} \quad \lambda_{1k}^{YZ} = \lambda_{j1}^{YZ} = 0, \tag{4.24}$$

for all possible values of the non-summing subscript (k or j).

In model (4.23) are present two two-factor interaction terms (from the three possible for a three-way table). The missing interaction term, the λ^{XY}, is the one responsible for the physical interpretation of the model, signaling missing interaction, in the presence of the other variable. Thus, X and Y are conditionally independent, given Z. The model of conditional independence of X and Z, given Y (or of Y and Z, given X) is defined analogously.

Naturally, the next model to be considered is the one having all three possible two-factor interactions. Thus, consider the model

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ}, \quad \forall\, i,j,k. \tag{4.25}$$

Additional to (4.20), (4.22), and (4.24), constraints

$$\sum_{i=1}^{I} \lambda_{ij}^{XY} = \sum_{j=1}^{J} \lambda_{ij}^{XY} = 0 \quad \text{or} \quad \lambda_{1j}^{XY} = \lambda_{i1}^{XY} = 0, \tag{4.26}$$

for all possible values of the non-summing subscript (j or i), are imposed on the parameters of this model.

It can be easily verified that under model (4.25), all *conditional odds ratios* of the kth XY partial table for all pairs (i,i'), (j,j') with $i < i'$ and $j < j'$

$$\frac{\pi_{ij|k}\pi_{i'j'|k}}{\pi_{i'j|k}\pi_{ij'|k}},\ i = 1,\ldots,I-1,\ i' = 2,\ldots,I,\ j = 1,\ldots,J-1,\ j' = 2,\ldots,J\ ,$$

are independent of k, $k = 1,\ldots,K$. Indeed, we have

$$\log\left(\frac{\pi_{ij|k}\pi_{i'j'|k}}{\pi_{i'j|k}\pi_{ij'|k}}\right) = \log\left(\frac{m_{ijk}m_{i'j'k}}{m_{i'jk}m_{ij'k}}\right) = \lambda_{ij}^{XY} + \lambda_{i'j'}^{XY} - \lambda_{i'j}^{XY} - \lambda_{ij'}^{XY}\ . \tag{4.27}$$

Hence, the XY conditional association does not depend on k, i.e., is homogeneous across the levels of Z. Analogously it can be proved that also the YZ and XZ conditional associations are homogeneous across the levels of X and Y, respectively. For this, model (4.25) is called the models of *homogeneous association*.

If we set $i' = i+1$ and $j' = j+1$ (without loss of generality), the conditional odds ratios above become the $\theta^{XY}_{ij(k)}$ local conditional odds ratios, defined in (3.4), and (4.27) leads to

$$\log \theta^{XY}_{ij(k)} = \lambda^{XY}_{ij} + \lambda^{XY}_{i+1.j+1} - \lambda^{XY}_{i+1.j} - \lambda^{XY}_{i.j+1}\ ,\ i = 1,\ldots,I-1,\ j = 1,\ldots,J-1, \tag{4.28}$$

independent of k. For the conditional odds ratios $\theta^{XZ}_{i(j)k}$ and $\theta^{YZ}_{(i)jk}$ hold analogous results.

Finally, the *saturated* model has an additional term, the three-factor interaction term λ^{XYZ} that accounts for the more complex connection of all three variables:

$$\log m_{ijk} = \lambda + \lambda^X_i + \lambda^Y_j + \lambda^Z_k + \lambda^{XY}_{ij} + \lambda^{XZ}_{ik} + \lambda^{YZ}_{jk} + \lambda^{XYZ}_{ijk}, \quad \forall\, i,j,k. \tag{4.29}$$

All terms of saturated model satisfy identifiability constraints, of the type given above. Thus also for the three-factor interaction term it holds

$$\sum_{i=1}^{I} \lambda^{XYZ}_{ijk} = \sum_{j=1}^{J} \lambda^{XYZ}_{ijk} = \sum_{k=1}^{K} \lambda^{XYZ}_{ijk} = 0 \quad \text{or} \quad \lambda^{XYZ}_{1jk} = \lambda^{XYZ}_{i1k} = \lambda^{XYZ}_{ij1} = 0. \tag{4.30}$$

The parameters of the saturated model are in 1-1 correspondence with the m_{ijk}. Taking into consideration the appropriate constraints and solving simple equations we can express all λ parameters as functions of the m_{ijk}'s, analogously to (4.9)–(4.12) for two-way tables.

All possible main effect and interaction terms that can appear in a three-way log-linear model are listed in Table 4.1, along with their number of them being "free," after the identifiability constraints consideration. All these "free" parameters sum to $IJK-1$, which is the dimension of the parameter space when the contingency table $\left(m_{ijk}\right)_{I\times J\times K}$ is multinomial distributed. The fixed term λ is considered as a parameter only under the *Poisson* sampling scheme; in which case the number of possible "free" parameters is IJK (in analogy to two-way contingency tables).

All the log-linear models considered so far are of a special type. In all of them, whenever a higher-order effect is in the model, then all possible lower-order effects involving the variables of this higher-order effect term are also in the model. Such models are called *hierarchical log-linear models* and are parsimoniously symbolized by the set of the highest-order terms (with respect to all variables) that define them uniquely. For instance, model $\log m_{ij} = \lambda + \lambda^X_i + \lambda^{XY}_{ij}$ for two-way tables is nonhierarchical, since it includes the term λ^{XY}_{ij}, without having the term λ^Y_j. Analogously, the absence of the term λ^Z_k makes $\log m_{ijk} = \lambda + \lambda^X_i + \lambda^Y_j + \lambda^{XY}_{ij} + \lambda^{YZ}_{jk}$ nonhierarchical. The hierarchical log-linear models for three-way tables are given in Table 4.2, along with their notation.

Table 4.1 Number of "free" parameters for each log-linear model term (main effect or interaction) applied on an $I \times J \times K$ contingency table, due to the identifiability constraints

Term	Number of parameters	Number of "free" parameters	Identifiability constraints
Main effects			
λ_i^X	I	$(I-1)$	(4.20) for λ_i^X
λ_j^Y	J	$(J-1)$	(4.20) for λ_i^Y
λ_k^Z	K	$(K-1)$	(4.20) for λ_i^Z
Two-factor interactions			
λ_{ik}^{XZ}	IK	$(I-1)(K-1)$	(4.22)
λ_{jk}^{YZ}	JK	$(J-1)(K-1)$	(4.24)
λ_{ij}^{XY}	IJ	$(I-1)(J-1)$	(4.26)
Three-factor interaction			
λ_{ijk}^{XYZ}	IJK	$(I-1)(J-1)(K-1)$	(4.30)

Table 4.2 Hierarchical three-way log-linear models

Model	Description	$\log m_{ijk} =$
(X,Y,Z)	Independence of X, Y, Z	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$
	Jointly independence of	
(Y,XZ)	Y from X and Z	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$
(X,YZ)	X from Y and Z	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$
(Z,XY)	Z from X and Y	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$
	Conditional independence of	
(XZ,YZ)	X and Y, given Z	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$
(XY,XZ)	Y and Z, given X	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$
(XY,YZ)	X and Z, given Y	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$
(XY,XZ,YZ)	Homogeneous association	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$
(XYZ)	Saturated	$\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$

4.4 Hierarchical Log-Linear Models for Multi-way Tables

Log-linear models can be defined for contingency tables of dimension higher than three, in a similar manner as for three-way tables. Log-linear models for multi-way tables include higher-order interactions, up to interactions of order equal to the dimension of the table. The number of possible models increases with the dimension of the table, involving the procedure of deciding for the one appropriate to describe the underlying structure of association. In order to impose a structure on model building, especially helpful in model selection, log-linear modeling is usually restricted to the family of *hierarchical log-linear models*.

Furthermore, the presence of nonhierarchical interaction terms in a model causes interpretational inconveniences. For example, in a 4-way table, cross-classifying variables X, Y, Z, and W, how can we understand and explain that variable X does not interact with Y (absence of the λ_{ij}^{XY} term from the model) but it interacts simultaneously with Y, Z, and W (model includes the $\lambda_{ijk\ell}^{XYZW}$ term)? Even among the

hierarchical log-linear models, the physical interpretation of the models becomes more involved as the dimension of the table increases. It is easier to understand and interpret a high-dimensional model by focusing on its missing terms. Missing interaction terms refer to variables that are conditional independent and conditional independence statements are easier to understand and express.

To clarify this, consider the hierarchical log-linear model (XYZ, YW) applied on the 4-way table described above. The formula of this model would be

$$\log m_{ijk\ell} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_\ell^W + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{j\ell}^{YW} + \lambda_{ijk}^{XYZ} .$$

Note that the missing two-factor interaction terms are XW and ZW, while W is associated to Y and X, Z are associated to each other and both to Y (also in a three-factor interaction). This signals that X and W are conditionally independent, given Y and Z. Indeed, the conditional XW log local odds ratios

$$\log \theta_{i(jk)\ell}^{XW} = \log m_{i(jk)\ell} + \log m_{i+1(jk)\ell+1} - \log m_{i+1(jk)\ell} - \log m_{i(jk)\ell+1}$$

under the above model turn out to be

$$\log \theta_{i(jk)\ell}^{XW} = 0, \quad \forall i = 1, \ldots, I-1, \; \ell = 1, \ldots, L-1,$$

for all j ($j = 1, \ldots, J$) and k ($k = 1, \ldots, K$), fact that verifies the conditionally independence of X and W, given Y, Z. In a symmetric manner, also Z and W are conditionally independent, given the other two.

For a higher-order example, let the variables $X_1, \ldots, X_7$ be cross-classified to form a $I_1 \times I_2 \times \ldots \times I_7$ contingency table. Then, model $(X_1X_2,\ X_1X_5,\ X_3X_4X_5,\ X_5X_6X_7)$ equates $\log m_{i_1 i_2 \ldots i_7}$ to the sum of the fixed term, λ, plus the sum of the seven main effects $\lambda_{i_k}^{X_k}$, $k = 1, \ldots, 7$, plus the sum of the eight two-factor interactions $\lambda_{i_k i_\ell}^{X_k X_\ell}$ from the 21 possible (the terms corresponding to the pairs $(k, \ell) =$ (1,2), (1,5), (3,4), (3,5), (4,5), (5,6), (5,7), (6,7) are in the model), plus the three-factor interactions terms $\lambda_{i_3 i_4 i_5}^{X_3 X_4 X_5}$ and $\lambda_{i_5 i_6 i_7}^{X_5 X_6 X_7}$. Observing the terms not included in the model, we can see that variables $X_1,\ X_2$ are jointly independent from $X_3,\ X_4$, conditional on $X_5,\ X_6,\ X_7$.

4.5 Maximum Likelihood Estimation for Log-Linear Models

For multi-way tables, the ML estimation procedure for a log-linear model $\mathscr{M}$ is analogous to the procedure followed in Sect. 4.2 for the two-way independence model (4.1). The log-likelihood function is of the (4.13) form, with the subscripts and the indices in the sum appropriately adjusted. Thus, for an $I_1 \times I_2 \times \ldots \times I_s$ table, cross-classifying variables $X_1, X_2 \ldots, X_s$, the kernel of the log-likelihood is

$$\ell(\boldsymbol{\lambda}) = \sum_{i_1, \ldots, i_s} \left(n_{i_1, \ldots, i_s} \log(m_{i_1, \ldots, i_s}) - e^{\log(m_{i_1, \ldots, i_s})} \right) , \tag{4.31}$$

where $m_{i_1,\ldots,i_s}$ are the expected frequencies under the assumed model $\mathscr{M}$ and $\boldsymbol{\lambda}$ the vector of all its parameters. It is then maximized with respect to every parameter in $\boldsymbol{\lambda}$ and the set of the associate likelihood equations is derived.

For the three-way hierarchical log-linear model (XZ, YZ), for example, the parameter vector $\boldsymbol{\lambda}$ (4.31) by (4.23) becomes

$$\ell(\boldsymbol{\lambda}) = \sum_{i_1,\ldots,i_s} \left(n_{i_1,\ldots,i_s} (\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}) - e^{\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}} \right) .$$

Then, solving $\frac{\partial \ell(\boldsymbol{\lambda})}{\partial \lambda_i^X} = 0$ leads to

$$\hat{m}_{i++} = n_{i++}, \quad i = 1, \ldots, I ,$$

which are the likelihood equations corresponding to the X main effect parameters. Analogously, with respect to the XZ interaction parameters, $\frac{\partial \ell(\boldsymbol{\lambda})}{\partial \lambda_{ik}^{XZ}} = 0$ leads to

$$\hat{m}_{i+k} = n_{i+k}, \quad i = 1, \ldots, I, \quad k = 1, \ldots, K .$$

The remaining sets of likelihood equations are $\hat{m}_{+j+} = n_{+j+}$ $(j = 1, \ldots, J)$ and $\hat{m}_{++k} = n_{++k}$ $(k = 1, \ldots, K)$, for the Y and Z main effects, respectively, and $\hat{m}_{+jk} = n_{+jk}$ (for all j, k), corresponding to the YZ interaction.

In general, log-linear models oppose some nice properties regarding their likelihood-based inference. It has been proved that the minimal sufficient statistics of a model $\mathscr{M}$ is the set of sample marginals, corresponding to the highest-order terms in the model, with respect to each variable. Thus, for (XZ, YZ), the sufficient statistics are (n_{i+k}, n_{+jk}), for all i, j, k, while for (X, YZ), they would be (n_{i++}, n_{+jk}), for all i, j, k. The likelihood equations of the model are then equating the sufficient statistics to their corresponding expecting values under $\mathscr{M}$ (Birch 1963).

The ML estimates under the independence model (4.1) are derived in closed-form expression but this is not the case in general. For most log-linear models for higher-dimensional tables, the likelihood equations do not lead to closed-form expressions for the ML estimates and have to be solved iteratively. The first algorithm applied for this was the *iterative proportional fitting* (IPF) algorithm. Predominant is now the *Newton–Raphson* (NR) algorithm, which will be presented in the context of the GLMs (Sect. 5.3.1).

Log-linear models for which closed-form MLEs exist are the *decomposable* models. The joint probability of a decomposable model can be factorized in a closed form in terms of marginal probabilities. This factorization is due to Goodman (1970, 1971c) while the term decomposable was introduced by Andersen (1974). Decomposable log-linear models received special attention in the 1970s and are treated in detail in Bishop et al. (1975, Sect. 3.4). They exhibit nice properties, connected also to *graphical log-linear models* (see Sect. 4.7.2).

4.6 Model Fit and Selection

The classical goodness-of-fit statistics to evaluate the fit of a multi-way log-linear model $\mathscr{M}$ are Pearson's X^2 and the LR statistic G^2, defined for an $I_1 \times I_2 \times \ldots \times I_s$ table as

$$X^2 = \sum_{i_1,\ldots,i_s} \frac{(n_{i_1,\ldots,i_s} - \hat{m}_{i_1,\ldots,i_s})^2}{\hat{m}_{i_1,\ldots,i_s}}, \tag{4.32}$$

$$G^2 = 2 \sum_{i_1,\ldots,i_s} n_{i_1,\ldots,i_s} \log(\frac{n_{i_1,\ldots,i_s}}{\hat{m}_{i_1,\ldots,i_s}}). \tag{4.33}$$

The asymptotic distribution for X^2 and G^2 under model $\mathscr{M}$ is $\chi^2_{d-d_0}$, where $d = \prod_{k=1}^{s} I_k - 1$ is the total number of "free" cells of the table under consideration under the multinomial sampling scheme, d_0 the number of "free" parameters of the assumed model $\mathscr{M}$ (overall λ is not considered as a parameter), and $\hat{m}_{i_1,\ldots,i_s}$ the ML estimate of the expected under $\mathscr{M}$ frequency for cell $(i_1,\ldots,i_s)$.

The residual degrees of freedom $df = d - d_0$ of the hierarchical log-linear models for three-way tables are given in Table 4.3. In this case $d = IJK - 1$ and d_0 is calculated by adding the number of "free" parameters for the terms in model from Table 4.1.

Evaluation of the model fit to the data includes also inspection of the residuals. The types of residuals discussed in Sect. 2.2.4 for two-way tables apply also to tables of higher dimension. The dissimilarity index $\hat{\Delta}$ in (4.18) is also defined for multi-way tables. It does not share the nice properties of G^2 but its relative reduction between models $\mathscr{M}_1$ and $\mathscr{M}_2$ can be used to compare practically the models, even if they are not nested.

The number of possible log-linear models increases with the dimension of the table, corresponding to different types of dependencies among the classification

Table 4.3 Hierarchical three-way log-linear models and their residual df

Model	Formula	df
(X,Y,Z)	(4.19)	$IJK - I - J - K + 2$
(Y,XZ)	(4.21)	$(J-1)(IK-1)$
(X,YZ)		$(I-1)(JK-1)$
(Z,XY)		$(K-1)(IJ-1)$
(XZ,YZ)	(4.23)	$K(I-1)(J-1)$
(XY,XZ)		$I(J-1)(K-1)$
(XY,YZ)		$J(I-1)(K-1)$
(XY,XZ,YZ)	(4.25)	$(I-1)(J-1)(K-1)$
(XYZ)	(4.29)	0

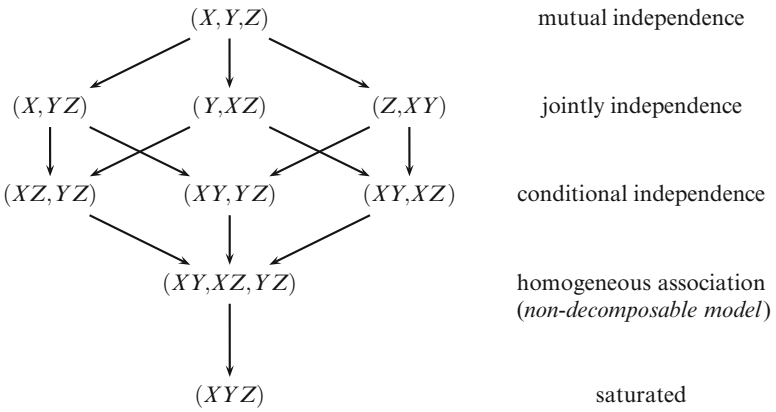


Fig. 4.1 Sequences of nested models for three-way tables, from the saturated (XYZ) to the model of mutual independence (X, Y, Z)

variables. Thus, model selection becomes a basic issue as the dimension of the table rises. The model selection procedure is based on the concept of "nested" models. In general, a model $\mathscr{M}_1$ is *nested* in model $\mathscr{M}_2$, denoted as $\mathscr{M}_1 \subset \mathscr{M}_2$, if $\mathscr{M}_1$ is derived from $\mathscr{M}_2$ by eliminating some of $\mathscr{M}_2$'s parameters. Thus $\mathscr{M}_2$ contains all the terms of $\mathscr{M}_1$ plus at least one more not present in $\mathscr{M}_1$.

Nested models are compared by conditional testing. Model $\mathscr{M}_1$ is more parsimonious than $\mathscr{M}_2$, but for this $G^2(\mathscr{M}_1) \geq G^2(\mathscr{M}_2)$. Given that model $\mathscr{M}_2$ holds, the adequacy of $\mathscr{M}_1$ is tested by

$$G^2(\mathscr{M}_1|\mathscr{M}_2) = G^2(\mathscr{M}_1) - G^2(\mathscr{M}_2)\,, \tag{4.34}$$

which under $\mathscr{M}_1$ is asymptotically $\mathscr{X}^2_{df(\mathscr{M}_1)-df(\mathscr{M}_2)}$ distributed.

The possible sequences of nested models for three-way tables are illustrated in Fig. 4.1. Conditional tests of the type (4.34) can be performed between models connected with arrows, not necessarily directly (see also Tutz 2012, Sect. 12.4).

The log-linear model selection procedure consists of a sequential search between hierarchical nested model, starting from the saturated model and removing terms (one at a time) by conditional testing the significance of the term removed. The process stops and decides for the model for which the next term to be removed leads to a significant increase of the test statistic (4.34). For each level of interaction, say k-factor interactions, the order the interaction terms are removed from the model is the order of their significance, less significant being removed first. The procedure described is a *step algorithm* of *backward* elimination. Alternatively, *forward* elimination algorithms start from the model of complete independence and continue to add terms, as long as they improve significantly the fit, according to the conditional test (4.34).

However, we should not let an algorithm decide blindly for the model. Sometimes, the nature of the problem or experimental conditions dictate the presence of nonsignificant terms in the model. For example, suppose in a survey responders are cross-classified according to their educational level (X_1), marital status (X_2), gender (X_3), and age in categories (X_4). From the experimental design it is controlled over gender and age, in the sense that the number of males and females participating in the survey is prespecified for each of the K age categories. This means that the table marginals $n_{++i_3i_4}$ for $i_3 = 1,2$ and $i_4 = 1,\ldots,K$ are set fixed by the design. If the X_3X_4 interaction term is not found to be significant by the selection algorithm and is thus not included in the model, then the corresponding likelihood equations, $m_{++i_3i_4} = n_{++i_3i_4}$ ($i_3 = 1,2$, $i_4 = 1,\ldots,K$), are missing. Consequently, the number of males and females assigned by the adopted model to each age group will not agree with the known prespecified numbers. Thus, the X_3X_4 interaction term should be included in the model, even if it is nonsignificant. In this case, the $\lambda_{i_3i_4}^{X_3X_4}$ terms signal the underlying product multinomial sampling design and not the physical significance of this interaction.

We have already mentioned the crucial role the concept of conditional independence plays in understanding and recording structures of associations in multi-way contingency tables. An important application of the above described model selection procedure is for testing for conditional independence structures, which is exposed next for three-way tables.

4.6.1 Conditional Test of Conditional Independence

In the context of a $I \times J \times K$ contingency table with classification variables X, Y, and Z, if the model of homogeneous association (XY,XZ,YZ) fits the data well, we can test for conditional independence between any two of them, given the third. This test will be *conditional* on homogeneous association. For example, the test of

$$H_0 : X, Y \text{ are independent, conditional on } Z \quad \text{vs.} \quad H_1 : \text{not } H_0$$

can be expressed as

$$H_0 : \text{model } (XZ,YZ) \quad \text{vs.} \quad H_1 : \text{model } (XY,XZ,YZ) \ ,$$

since we already know that the underlying association is homogeneous. The H_0 and H_1 models are nested; thus, the associated test can be based on the difference

$$G^2(XZ,YZ) - G^2(XY,XZ,YZ) \tag{4.35}$$

which, under H_0 and given that model (XY,XZ,YZ) holds, is asymptotically distributed as $\chi^2_{(I-1)(J-1)}$, since $df_{(XZ,YZ)} - df_{(XY,XZ,YZ)} = (I-1)(J-1)$.

For stratified 2×2 contingency tables, the conditional test (4.35) applied on the $2 \times 2 \times K$ table has $df = 1$ and is analogue to the Mantel–Haenszel test (3.9). This provides an intuitive justification to the fact that the Mantel–Haenszel test works best when the partial associations across the stratification levels are similar (remarked in Sect. 3.3.1).

4.6.1.1 Log-Linear Model Selection for Example 3.3

Reconsidering Example 3.3 of Sect. 3.3.4, if T, F, and C stand for the treatment outcome, the prognostic factor, and the clinic variables, then the homogeneous association log-linear model (TF, TC, FC)) and the model of treatment-factor conditional independence within clinic (TC, FC)) are of 5 and 6 df, respectively, and fitted in `MASS` as follows

```
: > hom.assoc<-loglm(~Treatment*Prognostic_Factor+
+           Prognostic_Factor*Clinic+Treatment*Clinic, data=dat)
> cond.ind.TF<-loglm(~Prognostic_Factor*Clinic+Treatment*Clinic)
```

The homogeneous association model is adequate, since $G^2(TF, TC, FC) = 7.950$ (p-value= 0.159) and $X^2(TF, TC, FC) = 7.894$ (p-value= 0.162), very close to the Breslow–Day–Tarone test statistic (3.15), which is equal to $BDT = 7.91$ ($df = 5$, p-value=0.161).

The conditional test (4.35) in this case is $G^2(TC, FC) - G^2(TF, TC, FC) = 34.845$, with associated p-value=3.570184e-09 ($df = 1$), computed as

```
> DG2 <- cond.ind.TF$deviance - hom.assoc$deviance
> p.value <- 1 - pchisq(DG2, 1)
```

while the corresponding difference in the X^2 statistics

```
> DX2 <- cond.ind.TF$pearson - hom.assoc$pearson
```

is $X^2(TC, FC) - X^2(TF, TC, FC) = 33.177$, also indicative of the inappropriateness of the conditional independence model considered (though not asymptotically $\mathscr{X}^2_1$ distributed). Thus the "treatment–prognostic factor" association is homogeneous across the clinics but conditional independence is rejected, based on the above G^2 conditional test. Recall that the Mantel–Haenszel test gave for this example $MH = 32.703$ ($df = 1$, p-value=1.074e-08), very close to the difference in X^2 statistics value above.

4.6.2 Log-Linear Model for Example 3.2

Reconsider the $5 \times 7 \times 2$ contingency table of the example introduced in Sect. 3.2, which is already given in the `R` array `party.tab`. The three-way log-linear model that describes this data table best will be achieved by the backward stepwise algorithm. The stepwise model selection algorithms, forward or backward, presented in Sect. 4.6, are implemented in `R` by the `step` function. In `step` the contribution

Table 4.4 Backward stepwise procedure of log-linear model selection for Example 3.2

Start: AIC=140
~D*P*G

	Df	AIC	LRT	Pr(Chi)
- D:P:G	24	120.82	28.818	0.2271
<none>		140.00		

Step: AIC=120.82
~ D + P + G + D:P + D:G + P:G

	Df	AIC	LRT	Pr(Chi)
- D:G	4	113.32	0.505	0.9730008
<none>		120.82		
- P:G	6	132.77	23.951	0.0005333 ***
- D:P	24	174.34	101.523	1.650e-11 ***

—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Step: AIC=113.32
~ D + P + G + D:P + P:G

	Df	AIC	LRT	Pr(Chi)
<none>		113.32		
- P:G	6	125.84	24.519	0.000419 ***
- D:P	24	167.41	102.091	1.319e-11 ***

—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Call:
loglm(formula = ~ D + P + G + D:P + P:G,
data = party.tab, evaluate = FALSE)

Statistics:

	X^2	df	P(> X^2)
Likelihood Ratio	29.32320	28	0.3962751
Pearson	29.17456	28	0.4037142

of a term is evaluated by the change its removal causes on Akaike's information criterion (AIC) value of the model. The saturated model is applied on `party.tab` and saved under `sat`. Then, with model `sat` as starting point, nonsignificant terms of this model are eliminated by the procedure `step`, as shown below. Recall that we work in library `MASS`.

```
> sat <- loglm(~ D*P*G, data=party.tab)
step(sat, direction="backward", test="Chisq")
```

The derived output is provided in Table 4.4.

Thus, according to the backward stepwise algorithm and based on conditional testings (4.34) between nested hierarchical log-linear models, the three-factor interaction is nonsignificant (p-value$= 0.227$). Further on, the two-factor interaction

Table 4.5 Conditional testing between nested hierarchical log-linear models for Example 3.2

LR tests for hierarchical log-linear models

Model 1:
~ D + P + G
Model 2:
~ D + P + G + D:P
Model 3:
~ D + P + G + D:P + P:G
Model 4:
~ D + P + G + D:P + P:G + D:G
Model 5:
~ D * P * G

	Deviance	df	Delta(Dev)	Delta(df)	P(> Delta(Dev)
Model 1	155.93392	58			
Model 2	53.84259	34	102.0913317	24	0.00000
Model 3	29.32320	28	24.5193932	6	0.00042
Model 4	28.81808	24	0.5051182	4	0.97300
Model 5	0.00000	0	28.8180773	24	0.22705
Saturated	0.00000	0	0.0000000	0	1.00000

DG is also nonsignificant with $G^2_{(PG,DP)} - G^2_{(DG,PG,DP)} = 0.505$ and associated (p-value$= 0.973$), based on the $\mathcal{X}^2_4$ approximation for the test statistic. The interaction terms DP and PG are both highly significant with $G^2_{(D,PG)} - G^2_{(PG,DP)} = 102.091$ and $G^2_{(G,DP)} - G^2_{(DG,PG,DP)} = 24.519$, respectively. Thus, the backward elimination procedure concludes to the (DP, PG) model, i.e., the responder's educational level (D) is conditional independent from his/her gender (G), given his/her party affiliation (P).

The procedure above provides at each stage the value of the AIC for the corresponding model as well. This criterion will be discussed in Sect. 5.3.2.

The successive conditional testings between nested hierarchical log-linear models, from the model of complete independence up to the saturated, adding terms according to their significance order, is summarized in the corresponding analysis of variance table, which is possible in `R` by function `anova`.

```
> I <- loglm(~D+P+G);  as_1 <- loglm(~D+P+G+D:P)
> as_2 <- loglm(~D+P+G+D:P+P:G); as_3 <- loglm(~D+P+G+D:P+P:G+D:G)
> anova(I, as_1, as_2, as_3, sat)
```

In the derived output (in Table 4.5), *deviance* coincides for log-linear models with the G^2 test statistic for the corresponding model (see Sect. 5.3.2). The conditional G^2 values between successive nested models are in column `Delta(Dev)`, followed in next columns by the difference between their df and the asymptotic p-value of the associated conditional test.

The mosaic plot of the observed frequencies for this example is provided in Fig. 3.2 (right). This mosaic plot can be enriched by displaying on it the residuals of each cell as well. Thus, the mosaic plot derived by

```
> mosaic(party.tab, gp = shading_Friendly,
        labeling= labeling_residuals)
```

is to be seen in Fig. 4.2 (left). It differs from Fig. 3.2 (right) in that the tiles are colored according to the value of the corresponding residuals for the model of complete independence. Negative significant residuals are red shaded while the positive significant are blue shaded, with the depth of the color strengthening for larger (in absolute value) residuals. We asked additional to label the tiles with the significant residual value, so red-shaded tiles are those with the negative residual values and blue with the positive ones. Cells with nonsignificant residuals are non-shaded (white) with red (dashed) frame for negative residuals and blue (solid) frame for positive ones. Thus, we observe that the highest positive residual corresponds to females with educational level less than high school, who are more political "independent" ("4") than expected under independence. The highest negative residual is for females with a bachelor degree, who are less political "independent" than expected under independence.

The residuals illustrated in the mosaic plot above were for the independence model (default). To refer to residuals of a different model, the output object of the assumed model has to take the position of the data matrix as input in `mosaic()`. Thus, the mosaic plot in Fig. 4.2 (left) can equivalently be obtained as

```
> mosaic(I, gp=shading_Friendly, labeling=labeling_residuals)
```

The residuals of the (PG, DP) model are incorporated in the mosaic plot by

```
> mosaic(as_2, gp=shading_Friendly, labeling=labeling_residuals)
```

The derived plot is provided in Fig. 4.2 (right) and we can easily verify that the Pearsonian residuals for (PG, DP) vary between -1.58 and 1.81, without anyone being significant.

The residuals pictured so far are the Pearsonian residuals (default in `mosaic()`). Alternative option is the deviance residuals, controlled by the option `residuals=`. Thus, the deviance residuals for model (D, P, G) are considered in the mosaic plot as

```
> mosaic(party.tab, gp = shading_Friendly, residuals="deviance",
        labeling= labeling_residuals)
```

For other type of residuals, they have to be calculated ahead and be read in `mosaic()`. This option will be illustrated in the context of GLMs for Example 2.4 in Sect. 5.4.1.

The ML estimates of the expected under the adopted model (PG, DP) cell frequencies are saved in array `MLE` by

```
> MLE <- fitted(as_2)
```

In order to visualize the structure of association dictated by each model, the mosaic plots based on the ML estimates of the expected cell frequencies under characteristic models are provided in Fig. 4.3. In particular, the mosaic plot of the ML estimates under the complete independence model (P, D, G) is in (a), while under (DP, G) and (DP, PG) in (b) and (c), respectively. For comparison reasons, in (d) is located the mosaic plot of the sample values, also given in Fig. 3.2 (right). Observe in (a) that under the complete independence all rectangular tiles

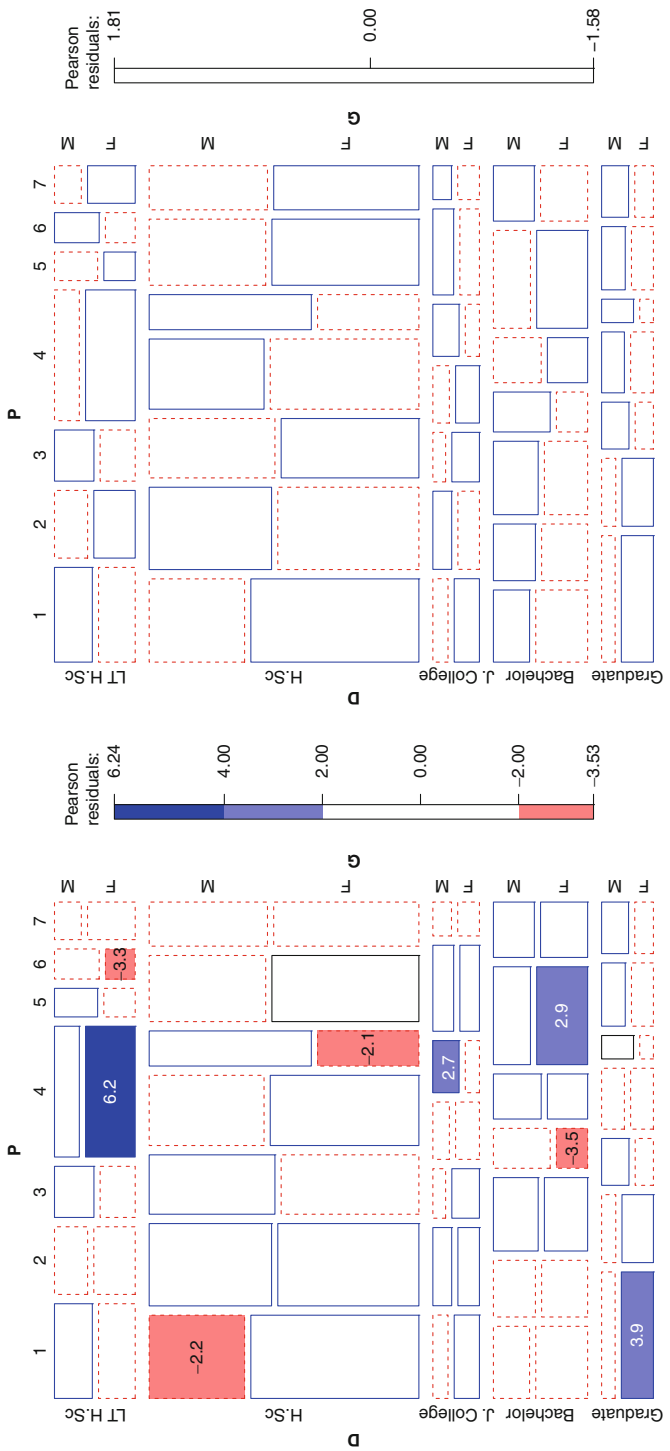


Fig. 4.2 Mosaic plots for Example 3.2 (Table 3.2), with tiles shaded by the Pearsonian residuals for models (P, D, G) (*left*) and (DP, PG) (*right*)

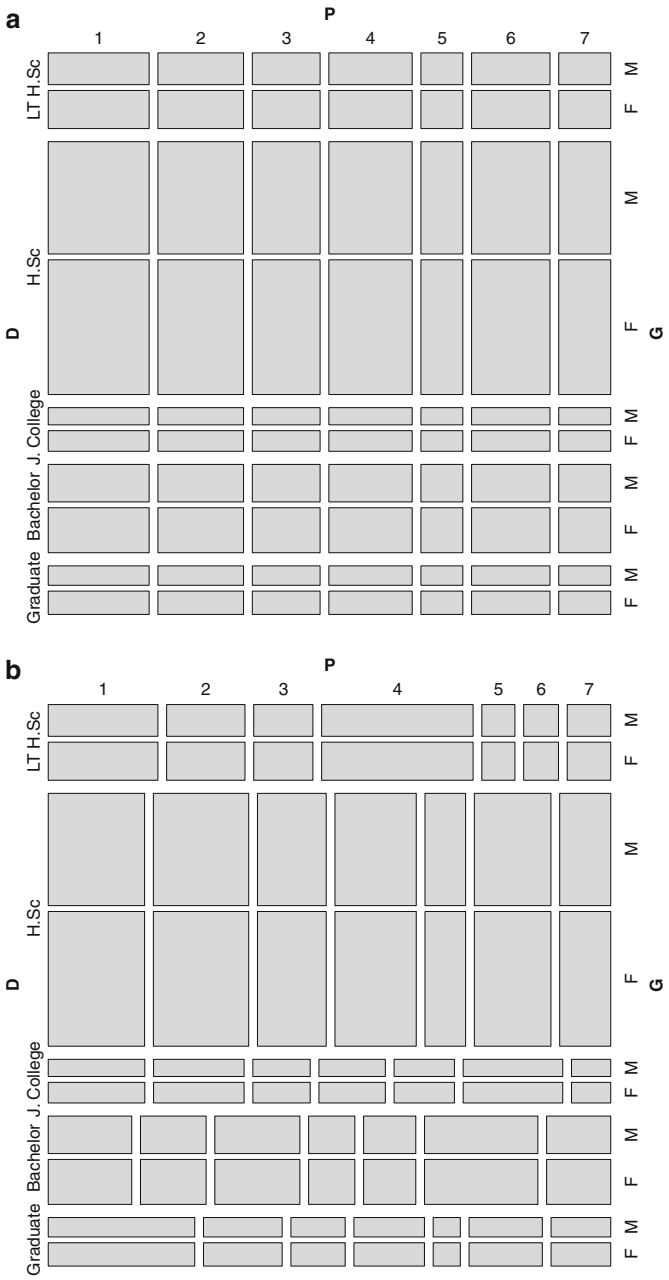


Fig. 4.3 Mosaic plots of the ML estimates of the expected cell frequencies for Example 3.2 (Table 3.2) under models (**a**) (P,D,G), (**b**) (DP,G), (**c**) (DP,PG) and of the observed cell frequencies in (**d**)

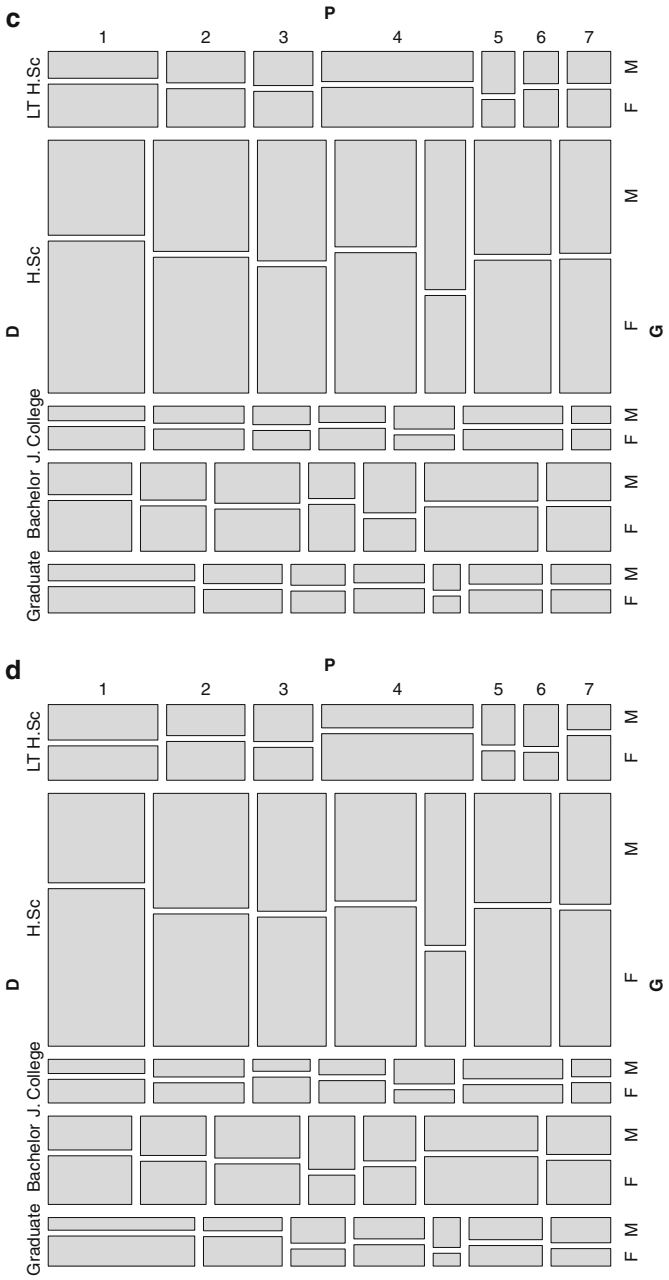


Fig. 4.3 (continued)

are perfectly aligned. Adding the *DP* interaction in (b), the alignment of the *DP* rectangles is disturbed while the within *P* by *G* division of the rectangles remains aligned, since the *PG* term is missing. This alignment is also lost in (c), which resembles closely to the mosaic plot of the observed frequencies in (d).

Mosaic plot in Fig. 4.3c is obtained by

```
> mosaic(MLE)
```

while replacing `MLE` with the array of estimates under (P,D,G) or (DP,G), plots (a) or (b) are derived, respectively.

4.7 Graphical Models

Log-linear models can also be defined as graphical models. Not all log-linear models are graphical, as we shall see next. Graphical models are useful whenever the detection of conditional independencies among the involved variables is of interest. They are a wide class of models whose conditional independence structure can be deduced by a graph. In the context of multi-way contingency tables, such graphs for log-linear models were introduced by Darroch et al. (1980), who called them *first order interaction graphs*. They are undirected graphs and in the context of graphical models they are known as *independence graphs* or *conditional independence graphs*. For reasons explained below, we shall use the term *conditional independence graphs*.

In case of high-dimensional contingency tables, graphical log-linear models provide guidance for possible collapsing over one or more classification variables without losing the relevant information. This dimension reduction problem is faced through the factorization criterion and model's decomposability.

To describe graphical models, one needs the basic notion of graph theory, the link between graph theory and probability models, and the group of models which are graphical, that is, whose conditionally independencies can be depicted by a graph.

Graphical models are not in the scope of this book but we shall introduce briefly the basic terminology on undirected graphs (Sect. 4.7.1) and the class of graphical log-linear models in order to connect them with classical log-linear models (Sect. 4.7.2) and use them in the discussion on dimension reduction of multi-way contingency tables by collapsing over one or more of the classification variables (Sect. 4.8).

4.7.1 Undirected Graphs

An undirected graph consists of a finite set of nodes (or vertices) V and a set of edges E, connecting some (or all) of the nodes in pairs. Consider, for example, a set of five notes $V = \{v_1, v_2, v_3, v_4, v_5\}$. Then, an undirected graph $\mathcal{G} = (V, E)$ consists

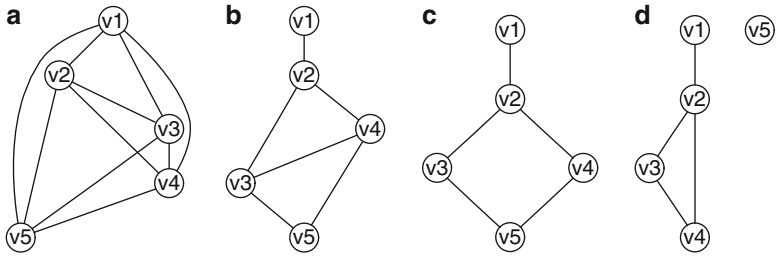


Fig. 4.4 Undirected graphs $\mathcal{G} = (V,E)$ for $V = \{v_1, v_2, v_3, v_4, v_5\}$ and
(a) $E = \{\{v_1,v_2\},\{v_1,v_3\},\{v_1,v_4\},\{v_1,v_5\},\{v_2,v_3\},\{v_2,v_4\},\{v_2,v_5\},\{v_3,v_4\},\{v_3,v_5\},\{v_4,v_5\}\}$,
(b) $E = \{\{v_1,v_2\},\{v_2,v_3\},\{v_2,v_4\},\{v_3,v_4\},\{v_3,v_5\},\{v_4,v_5\}\}$,
(c) $E = \{\{v_1,v_2\},\{v_2,v_3\},\{v_2,v_4\},\{v_3,v_5\},\{v_4,v_5\}\}$,
(d) $E = \{\{v_1,v_2\},\{v_2,v_3\},\{v_2,v_4\},\{v_3,v_4\}\}$

of the five nodes in V and up to ten edges (the elements of E) connecting pairwise the nodes. Possible graphs for this setup are provided in Fig. 4.4.

Next, we shall briefly refer to some terminology for undirected graphs. Two nodes $v_1,\ v_2 \in V$ are said to be *adjacent* in $\mathcal{G}$ if the edge $\{v_1,v_2\}$ belongs to E, so that they are connected by a line in the corresponding graph. A subset $A \subset V$ is *complete* if all pairs of nodes in A are adjacent. A graph $\mathcal{G} = (V,\ E)$ is *complete* if its set of nodes V is complete. A complete subset of nodes C induces a complete subgraph of the graph of $\mathcal{G}$. If this subgraph becomes incomplete by the addition of a further node, then C is maximally complete and is said to be a *clique*. A *path* is a sequence of distinct nodes $v_1, v_2, \ldots, v_k$ for which each successive pair of nodes are adjacent. Two nodes v_i and v_j are *connected* if there exists a path joining them. A *cycle* is a path with $v_1 = v_k$ and is said to be *chordless* if only the successive pairs of nodes in the cycle are adjacent. An edge of a cycle that connects to non-successive nodes in the cycle is characterized as a *chord*. A graph is *chordal* (or triangulated) if each of its cycles of four or more nodes has a chord. A subset of nodes B *separates* two nodes v_i and v_j if every path joining them contains at least one node from B. A subset B separates two subsets of N, A, and C, if it separates every pair of nodes $v_i \in A$ and $v_j \in C$.

4.7.2 Graphical Log-Linear Models

Graphical models are a family of probability models, simplified through *conditional independencies* represented in graphs. Focusing on contingency tables, the family of graphical log-linear models is a subset of the hierarchical log-linear models that utilizes undirected graphs to represent conditional independencies.

The connection between graphical log-linear models for a K-way contingency table (with cross-classifying variables $X_1, \ldots, X_K$) and undirected graphs is achieved by assuming that (i) the set V of a graph consists of K nodes $(v_1, \ldots, v_K)$, one for

each classification variable of the table, and (ii) the set of edges E connects some (or all) of the nodes in pairs, indicating a lack of independence between the variables. There is a one-to-one correspondence between models and graphs. In particular, given an undirected graph, the corresponding graphical log-linear model is defined as the hierarchical log-linear model with generators the *cliques* of the graph. For this reason, graphical log-linear models are not always parsimonious models.

Thus, for a five-way table ($K = 5$), the graph provided in Fig. 4.4a is a *complete graph* and corresponds to the saturated model $(X_1X_2X_3X_4X_5)$, while the graphs of Fig. 4.4b–d correspond to the graphical log-linear models $(X_1X_2, X_2X_3X_4, X_3X_4X_5)$, $(X_1X_2, X_2X_3, X_2X_4, X_3X_5, X_4X_5)$, and $(X_2X_3X_4, X_1X_2, X_5)$, respectively. For instance, verify for model $(X_1X_2, X_2X_3X_4, X_3X_4X_5)$ that its three maximal interaction terms correspond to the cliques of the graph in Fig. 4.4b. Only log-linear models with this correspondence are graphical. Thus, the hierarchical log-linear model $(X_1X_2, X_2X_3, X_2X_4, X_3X_4, X_3X_5, X_4X_5)$ is not graphical. Such exclusions from the class of graphical log-linear models ensure the one-to-one correspondence between models and graphs mentioned above.

Conditional independence is the key concept for defining graphical log-linear models. Thus, a representative example of a non-graphical log-linear model is the model of homogeneous association for three-way tables, since it has no conditional independence interpretation. The set of conditional independencies involved in a graphical log-linear model are ruled by three Markov properties, whose description is out of the scope of this section. See Lauritzen (1996) or Højsgaard et al. (2012) for details.

Graphs of graphical log-linear models are interpreted in terms of their missing edges, which are indicative of the underlying conditional independence structure, justifying thus that they are referred to as *conditional independence graphs* (see also Agresti 2013). In particular, the variables corresponding to two nonadjacent nodes in a graph are *conditionally independent*, given the nodes (variables) in the paths connecting them.

Conditional independence is connected to separation of nodes' subsets. If subsets of nodes A and C are separated by subset B in the graph, then, under the corresponding model, variables in A are conditionally independent to variables in B, given C.

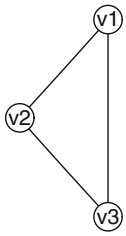
We have seen that an important subset of the hierarchical log-linear models are the *decomposable* models, which lead to MLEs of closed-form expression (see Sect. 4.5). Graphical decomposable log-linear models are graphical models with chordal graphs. The graphical models pictured in Fig. 4.4 are all decomposable except case (c).

Inference for graphical models is beyond the scope of this book. We shall only illustrate them briefly in the following section's examples, mostly to highlight their role in collapsing over one or more classification variables of a high-dimensional contingency table.

For constructing conditional independence graphs and fitting graphical models in `R` one can consult Højsgaard et al. (2012, Chaps. 1 and 2). For example, graphs (a) and (d) of Fig. 4.4 are derived in `gRbase` as shown below.

```
> library(gRbase)
> ag.a <- ug(~v1:v2:v3:v4:v5);  plot(ag.a)
> ag.d <- ug(~v1:v2+v2:v3:v4+v5;  plot(ag.d)
```

Often the association structure of a high-dimensional hierarchical log-linear model (not necessarily graphical) is visualized in terms of a graph, known as *association graphs*. Note however that there is not a one-to-one correspondence between log-linear models and association graphs. More than one log-linear models may have the same graph. For example, considering the graphs in Fig.4.4 as association graphs of hierarchical log-linear models, (b) is also the graph for the model $(X_1X_2, X_2X_3, X_2X_4, X_3X_4, X_3X_5, X_4X_5)$, while (a) is also (among others) the conditional independence graph of the hierarchical log-linear model including all possible two-factor interactions and none of higher order. In general, a triangle subgraph



of a conditional independence graph expresses the association structure between X_1, X_2 and X_3 of a hierarchical log-linear model containing the corresponding three-factor interaction as well as of a model without this three-factor but all associated pairwise interactions.

The graphs presented so far are *undirected* graphs and are applicable when the classification variables are treated in a symmetric manner in terms of the underlying associations. In the case of one or more response variables, the association structures are visualized through the *directed acyclic graphs* and the *chain graphs*.

4.8 Collapsibility in Multi-way Tables

An intuitive way to treat multi-way tables is to reduce their dimension by collapsing over classification variables that are not of direct interest. This way, the collapsing variables are ignored, though they may correspond to covariates that influence the relationship among the variables of interest. Such variables are characterized as *confounding* variables and should be controlled (through conditioning on their levels) instead of ignored. Thus, the association structure among the variables of interest studied on the marginal table produced by collapsing over a confounding variable does not necessarily express their interrelationships but reflects possibly a confounded effect (that of the collapsing variable on the variables of interest).

Furthermore, collapsing over a confounding variable can falsify the structure of the underlying associations, since partial associations can differ substantially (even in direction) from the corresponding marginal ones, as already stated in the context

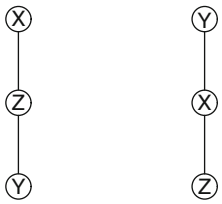


Fig. 4.5 Conditional independence graphs for models (XZ, YZ) (*left*) and (XY, XZ) (*right*)

of $2 \times 2 \times K$ tables in Sect. 3.2.2. This phenomenon is known as *Simpson's paradox* (Yule 1912; Simpson 1951), which states that the association in a marginal table can be of different direction than conditional association at each corresponding partial table.

Hence, the dimension of a table should not be reduced without ensuring that confounding does not occur. Conditions under which collapsing is possible in three-way tables are exposed next, while a discussion on conditions for multi-way tables follows.

Consider a $I \times J \times K$ table, cross-classifying the variables X, Y, and Z and suppose that we are interested in the XZ association. The XZ marginal and the XZ conditional (given Y) local odds ratios coincide (and thus we could collapse over Y without affecting the XZ association), if either X, Y are conditional independent, given Z, or Y, Z are conditional independent, given X, i.e., if the underlying model is the (XZ, YZ) or (XY, XZ), respectively. These patterns of conditional independencies can easily be visualized in the conditional independence graphs of these models in terms of separated variables (see Fig. 4.5). Thus, under both models we can collapse over Y, since it is separated from X (Z) by Z (X) for model (XZ, YZ) (XY, XZ). With similar arguments we can verify in Fig. 4.5 (left) that for (XZ, YZ) we could also collapse over X but not over Z.

In general for multi-way contingency tables, conditions under which they can be collapsed are provided by Bishop et al. (1975, Chap. 2), who defined the classical parametric collapsibility. It is based on the condition that if a model for a multi-way tables partitions the classification variables into three mutually exclusive subsets A, B, and C, such that B separates A and C, then parameters relating variables in A and variables in A to variables in B remain unchanged when collapsing over the variables in set C. This means that the association structure of a contingency table is not affected by collapsing over a variable (or a set of variables), only if it is conditionally independent to another variable (or set of variables) of the contingency table, conditioning on the rest of the variables. Since the concept of conditional independence is the fundamental kernel of graphical log-linear models (Sect. 4.7.2) and due to the "separation–conditional independence" connection, graphical models and the associated graphs are extremely useful in detecting patterns of conditional independencies and take decisions for collapsing, especially in high-dimensional contingency tables. For a discussion on the alternative approaches on collapsibility, see Sect. 4.9.4.

Table 4.6 *DP* ML estimates of the expected under (PG, DP) conditional and marginal local odds ratios for the data in Table 3.2

(G): males	Political party affiliation (P)						
Degree (D)	1	2	3	4	5	6	7
1: LT high school	1.385	0.955	0.462	2.289	1.790	0.533	
2: High school	0.948	0.878	0.980	1.818	0.876	0.591	
3: Junior college	0.838	2.045	0.470	1.242	1.316	1.436	
4: Bachelor	0.684	0.532	2.390	0.341	1.243	1.440	
5: Graduate							

4.8.1 Collapsing for Example 3.2

Recall that for Example 3.2, model (PG, DP) was selected. As expected due to (3.19) and the discussion above, since under (PG, DP) variables D and G are conditionally independent given P, it holds

$$\hat{\theta}^{DP}_{ij(1)} = \hat{\theta}^{DP}_{ij(2)} = \hat{\theta}^{DP}_{ij}\ ,\ \ i = 1, \ldots, 4,\ j = 1, \ldots, 6,$$

and their common estimated expected values are provided in Table 4.6.

The estimates of the expected under (PG, DP) conditional *DP* local odds ratios can be calculated in `R` following the procedure described in Sect. 3.2 for the corresponding observed local odds ratios just by replacing the `party.tab` by the `MLE` array. The *DP* partial fitted tables for male and female are respectively

```
> eDP1 <- MLE[„1];  eDP2 <- MLE[„2]
```

and the *DP* fitted marginal (over gender) table is

```
> eDPm <- margin.table(MLE, c(1,2))
```

The 4×6 table of fitted under (PG, DP) conditional (for males) local odds ratios $\left(\hat{\theta}^{DP}_{ij(1)}\right)$ is then derived by

```
> eOR<-exp(t(matrix(as.vector(C%*%log(as.vector(t(eDP1)))),NJ-1)))
```

Replacing table `eDP1` by `eDP2` and `eDPm` in the command above, the conditional $\left(\hat{\theta}^{DP}_{ij(2)}\right)$ and the marginal $\left(\hat{\theta}^{DP}_{ij}\right)$ fitted tables are produced, respectively, which under (PG, DP) coincide.

Alternatively, (PG, DP) can be fitted as a graphical model in `gRim`. In Sect. 3.2.4 the data were stored in the array `part.tab`. In `gRim`, if the data are in a contingency table format, they need to be defined as table. Thus, the graphical model is fitted as follows.

```
> library(gRim)
> party <- as.table(party.tab)
> graph.PG.DP <- dmod(~P*G+D*P, data=party)
```

The conditional independence graph of the model, given in Fig. 4.6 (right), is derived by

```
> plot(graph.PG.DP)
```

Based on the graph, we observe that we could collapse over gender (G).

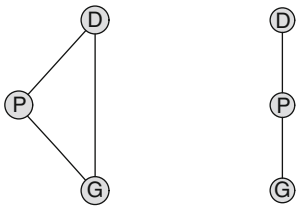


Fig. 4.6 Conditional independence graphs for Example 3.2 (Table 3.2) for the saturated model (DPG) (*left*) and for the graphical log-linear model (DP, PG) (*right*)

Table 4.7 Cross-classification of a sample of 2,228 responders according to their age, presence of depression, their gender (G), and whether they are living alone

Living alone (L): no

Gender (G) Males	Depression (D)	
Age (A)	No	Yes
≤ 45	283	16
> 45	270	13

Gender (G) Females	Depression (D)	
Age (A)	No	Yes
≤ 45	310	44
> 45	262	63

Living alone (L): yes

Gender (G) Males	Depression (D)	
Age (A)	No	Yes
≤ 45	212	34
> 45	113	63

Gender (G) Females	Depression (D)	
Age (A)	No	Yes
≤ 45	291	46
> 45	138	70

Analogously, collapsing over the educational level (D) is also possible but not over the party affiliation (P).

4.8.2 *Example 4.1*

Consider the $2 \times 2 \times 2 \times 2$ contingency table produced by cross-classifying a sample of 2,236 responders according to presence of depression (D), their gender (G), and whether they are living alone (L) and are aged above 45 (A), given in Table 4.7 (artificial data).

If we are interested in the association between depression and age, the marginal AD sample odds ratio is

$$\hat{\theta}^{AD} = \frac{1,096 \cdot 209}{140 \cdot 783} = 2.09 \,,$$

indicating that the odds of depression is 2.1 times higher for people over 45. But, looking at the conditional AD sample odds ratio, for all possible combinations of G and L, we get

$$\left(\hat{\theta}^{AD(LG)}\right) = \begin{pmatrix} 0.852 & 1.694 \\ 3.476 & 3.209 \end{pmatrix},$$

realizing that Simpson's paradox occurs. Indeed, we observe that for men the AD association changes direction with respect to the living conditions (first column). In particular, for men not living alone, the odds of depression is 1.2 $(= 1/0.852)$ times higher for people up to 45 than older while for men living alone it is 3.5 times higher for people older than 45.

Studying the underlying association structure, we proceed to log-linear model selection via the backward stepwise procedure, implemented in `R` as follows.

```
> freq<-c(283,270,16,13,310,262,44,63,212,113,34,63,291,138,46,70)
> names<-list(A=c("<45",">=45"), D=c("no","yes"), G=c("M","F"),
+             L=c("no","yes"))
> dat <- array(freq, c(2,2,2,2), dimnames=names)
> sat <- loglm(~A*D*G*L, data=dat)
> step(sat, direction="backward" , test="Chisq")
```

The proposed model is the (ADL, DGL) with $G^2 = 3.886$ and p-value=0.4216 (based on the $\mathscr{X}^2_4$ approximation), which is graphical.

In the graphical models framework, the saturated model is fitted in `GRim` as

```
> depression <- as.table(dat)
> graph.sat <- dmod(~A*D*G*L, data=depression)
```

while

```
> plot(graph.sat)
```

produces its conditional independence graph, pictured in Fig. 4.7 (left). Based on the saturated model, the backward model selection procedure

```
mod.sel <- backward(graph.sat)
```

```
. BACKWARD: type=decomposable search=all, criterion=aic(2.00), alpha=0.00
. Initial model:  is graphical=TRUE  is decomposable=TRUE
  change.AIC -4.1140  Edge deleted: G A
```

suggests to delete the GA edge (based on the AIC, see Sect. 5.3.2). This leads, as expected, to (ADL, DGL). As a graphical model, it is fitted by

```
> graph.model <- dmod(~A*D*L+D*G*L, data=depression)
```

The derived output

```
> graph.model
```

```
Model: A dModel with 4 variables

graphical   : TRUE          decomposable  : TRUE
-2logL      : 10940.92      mdim          : 11        aic : 10962.92
ideviance   : 171.80        idf           : 7         bic : 11025.72
deviance    : 3.89          df            : 4
```

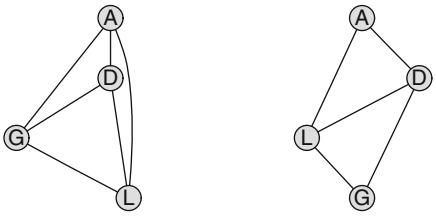


Fig. 4.7 Conditional independence graphs for Example 4.1 (Table 4.7) for the saturated model $(ADGL)$ (*left*) and for the graphical log-linear model (GLD, DLA) (*right*)

provides information on whether the fitted model is graphical and decomposable as well as on its goodness of fit. `-2logL` is minus twice the maximized log-likelihood and `mdim` the number of parameters in the model. `deviance` and `df` give the likelihood ratio statistic value and the associated degrees of freedom of the fitted model while `ideviance` is $G^2((A,D,G,L)|(ADL,DGL))$, that is, the likelihood ratio statistic for testing independence conditional on the fitted model. The degrees of freedom corresponding to this conditional test are `idf`. The term "deviance" and the other two criteria (AIC and BIC) will be introduced in Sect. 5.3.2.

The fact that (GLD, DLA) is a graphical decomposable log-linear model can easily be verified also by its association graph, derived by

```
> plot(graph.model)
```

and provided in Fig. 4.7 (right). By this graph, we verify that Simpson's paradox may occur when collapsing over L. On the other hand, when marginalizing over G, the AD and AL association structures are still well estimated, i.e., Simpson's paradox does not occur. Also collapsing over A is possible (the DG and GL association structures are not affected). Thus, the Simpson paradox we noticed for the AD marginal table is due to the marginalization over L.

4.9 Overview and Further Reading

4.9.1 On Log-Linear Models Analysis

The contribution of Birch (1963) in the analysis of multi-way tables was essential. He was the first who pointed out the equivalence of multinomial and Poisson log-linear models. Furthermore, his result that the ML estimates are the sole estimates that equate a log-linear model's sufficient statistics to their fitted values was a milestone for the log-linear models analysis. He generalized earlier work by Roy, Mitra, and Kastenbaum. It is fair to mention that the first who worked on the interaction structure for multi-way tables was Bartlett (1935), who considered the $2 \times 2 \times 2$ case. A review of these early results is provided by Goodman (1963b, 1964). Fundamental in the multi-way log-linear models establishment was the contribution of Cox, Darroch, Good, Goodman, Bishop, and Fienberg. Seminal to the theoretical development of the topic is the contribution of Haberman, who generalized Birch's

results and provided a formal investigation of MLEs for log-linear models and their properties (see Haberman 1974a). He also developed Newton–Raphson iterative algorithms for their fit (Haberman 1973a), which gave estimates of the standard error of the MLEs as well, and made thus their asymptotic significance testing feasible. The algorithm applied by then was the IPF algorithm of Deming and Stephan (1940), adjusted for log-linear models by Bishop (1969), Fienberg (1970a), and Darroch and Ratcliff (1972). Ku and Kullback (1974) approached log-linear models by indicating the analogies to linear models for continuous variables. Lang (1996b) provided a detailed discussion on the comparison of multinomial and Poisson log-linear models. For an extended historical review on the development of the inferential methods for log-linear models we refer to Fienberg and Rinaldo (2007).

Zero frequencies of a contingency table need special consideration and are distinguished between two types, the *sampling* and the *structural zeros*. Sampling zeros refer to cells of low but positive probability that may lead to zero observed frequencies in a certain realization. They are thus *random zeros*, and the corresponding cells are included in the analysis leading to nonzero expected frequencies estimates. Sampling zeros need no special consideration, in general. Traditionally, it has been suggested either to add a small positive constant ε only to the zero cells (see Grizzle et al. 1969) or to add it always (see Cox 1970b; Goodman 1970). Bishop (1969), Fienberg (1970b), and Goodman (1971b) dealt further with the problem of log-linear models' ML estimation in the presence of sampling zeros. Glonek et al. (1988) proved that for hierarchical log-linear models for multi-way contingency tables, the positivity of the sufficient statistics (i.e., corresponding marginal totals of the table) ensure the existence of the MLEs if and only if the model is decomposable. For non-decomposable models, they discuss the additional conditions required.

Tables with many sampling zeros (*sparse tables*) require special consideration, since the standard asymptotic theory does not apply and technical problems may arise in the estimation procedure. A contingency table with large number of cells and relative small total sample size will contain many zero cells and is called *sparse*. The basic asymptotic theory for testing nonparametric null hypotheses for multinomial data under sparseness assumption has been developed by Holst (1972) and Morris (1975). In case of sparse two-way tables, Mehta and Patel (1983) show that Fisher's exact test and Pearson's X^2 can lead to contradictory conclusions. Zelterman (1987) indicated that X^2 can show significant bias in testing independence for sparse tables and proposed a new statistic, D^2, which is also supported by Haberman (1988) in the context of null hypotheses defining unequal cell probabilities. Goodness-of-fit tests for sparse multinomials are reviewed and compared in Kim et al. (2009).

A class of test statistics for sparse tables with ordered categories are proposed by Burman (2004), which under certain conditions are asymptotically more powerful tests than Pearson's chi-square. Classes of goodness-of-fit tests under sparseness for multidimensional multinomial contingency tables are considered by Maydeu-Olivares and Joe (2005, 2006), based on low-order marginal proportions. Koehler (1986) and Dale (1986) studied the fit of log-linear models on sparse tables. Fienberg and Rinaldo (2012) studied ML estimation in log-linear models, conditions

of their existence, and the role of the sampling zeros. An alternative approach to treat sparse tables is the Bayesian (Sect. 10.5). Sparseness is also met in high-dimensional data (see Sect. 10.6). On the other hand, structural zeros are cells of zero probability that must be excluded from the analysis and thus not estimated. Structural zeros will be faced in Sect. 5.5.

Statistical inference for categorical data is mostly asymptotic, based on large sample approximations. For log-linear models, Haberman (1977) provided conditions for the asymptotic normality of linear functions of the MLEs and for the asymptotic chi-squared distribution of Pearson's X^2 and the G^2 goodness-of-fit statistics. He further pointed out that they remain applicable even if individual cell frequencies are small, provided the sample size and the number of cells of the table are large. The analysis of small sample contingency tables is briefly reviewed in Sect. 10.4.

Friendly (1994) connected mosaic plots to log-linear models, visualizing on mosaic displays beyond the observed cell frequencies (by the area of the cell rectangular) also the residuals (through shadings of the cell areas). For more on visualizing log-linear models via mosaic plots, we refer to Theus and Lauer (1999). Zeileis et al. (2007) visualized on mosaic plots departures of independence in two-way tables and models of conditional independence for three-way tables through residual shadings that code also significance of associations.

Beyond MLEs, the broad class of best asymptotic normal (BAN) estimators has been developed for the multinomial distribution by Neyman (1949), which share optimal large sample properties. In this class belong the *weighted least squares* (WLS) estimators, which are simpler to compute than the MLEs. The basic reference on WLS estimation for categorical data models is Grizzle et al. (1969).

Early contributions on treating misclassification of categorical data are by Bross (1954), facing the problem in 2×2 tables, and by Mote and Anderson (1965), considering its effect on X^2 tests. Espeland and Odoroff (1985) proposed a log-linear model for misclassified categorical data, fitted by the EM algorithm, generalizing earlier results by Chen (1979). A review on methods of categorical data analysis subject to misclassification is provided by van den Hout and van der Heijden (2002) while Buonaccorsi (2010, Chap. 2) treats two-way tables under misclassification extensively.

4.9.2 Residual Analysis: Outlier Detection

Residuals for two-way tables were introduced by Anscombe and Tukey (1963), who proposed graphical and analytical procedures to analyze the residuals. Later on, Cox and Snell (1968) defined residuals in a more general setup and studied their asymptotic properties. They did not deal with contingency tables but discussed problems concerning Poisson and binomial distributed samples. Haberman (1973b) developed methods of residual analysis for log-linear models in two-way tables, complete and incomplete. In particular, he considered the models of independence

and quasi-independence, supporting the use of the standardized residuals over the Pearsonian. Pearsonian and standardized residuals were compared in terms of the type I error rates of post hoc cellwise tests for two-way tables under independence and homogeneity models by MacDonald and Gardner (2000) and García-Pérez and Núñez-Antón (2003). The conclusions of MacDonald and Gardner (2000) were in favor of the standardized residuals. García-Pérez and Núñez-Antón (2003) considered the moment-corrected Pearsonian residuals and concluded that they behave the same as the standardized when the marginal distributions of the table are uniform while standardized residuals behave slightly better for peaked marginal distributions. The residuals presented in Sect. 2.2.4 are the most known and widely used. However, a variety of alternative residuals have been suggested in the literature. For example, Brown (1974) and Simonoff (1988) introduced the deleted residuals, for which each expected cell frequency is estimated by the model of quasi-independence, fitted on the data table with this particular cell replaced by a structural zero.

Residuals are a crucial tool for detecting outliers in a contingency table (Simonoff 1988). On outlier detection for two-way tables see, among others, Fuchs and Kennet (1980), Kotze and Hawkins (1984), and Lee and Yick (1999). For outlier detection and measures of influence, see Hastie and Pregibon (1992) and Lee and Fung (1997). For outlier identification in multi-way contingency tables, see Kuhnt (2004) and references cited therein. Alternatively, outliers are treated via algebraic statistics (see Sect. 10.4) by Rapallo (2012).

4.9.3 On Graphical Log-Linear Models

The connection of log-linear to graphical models is due to Darroch et al. (1980), while important early contributions are by Edwards and Kreiner (1983) and Wermuth and Lauritzen (1983). Classical reference sources on graphical models are Whittaker (1990) and Lauritzen (1996). Graphical models with missing data are dealt in Lauritzen (1995). Conditional independence graphs for multi-way log-linear models along with more complex multigraphs and the construction of fundamental conditional independencies for non-decomposable log-linear models are discussed in Khamis (2011). For graphical models with causal motivation, distinguishing between explanatory and response variables, see in Sect. 8.4.2.

4.9.4 On Collapsibility

Collapsibility, discussed in Sect. 4.8, is an important concept associated with the dimension reduction of multi-way contingency tables without affecting the underlying association structure information. Issues of collapsibility are tied related to Simpson's paradox. For more on Simpson's paradox we refer to Simpson (1951), Blyth (1972), and Samuels (1993).

There exist various notions of collapsibility, starting with the classical *parametric collapsibility* (Bishop et al. 1975, Chap. 2). Further necessary and sufficient conditions of parametric collapsibility, less restrictive than those by Bishop et al. (1975), are provided by Whittemore (1978), who introduced also the term of *strict collapsibility*. Additional to strict collapsibility, Ducharme and Lepage (1986) considered the *pseudo collapsibility* and tested the various types of collapsibility based on the table's nominal odds ratios. A geometric approach for exploring collapsibility is provided by Shapiro (1982). Vellaisamy and Vijay (2007) stated the results of Whittemore in an alternative form using the technique of Möbious inversion and further established new results on collapsibility and strict collapsibility.

Asmussen and Edwards (1983) approached collapsibility via graphical models and defined the *model-collapsibility*. They linked collapsibility to model's decomposability and to the idea of invariance of models when some variables are unobserved. They also showed that model-collapsibility is often equivalent to estimate-collapsibility. The different types of collapsibility conditions are reviewed in Whittaker (1990, Sect. 12.5). Model-collapsibility is also considered in Khamis (2011). Vellaisamy and Vijay (2010) obtained necessary and sufficient conditions for the strict collapsibility based on the interaction parameters of the conditional log-linear model adopted for the layers of the conditioning variables. They considered also the model-collapsibility for hierarchical log-linear models under the conditioning framework and provided connections between the strict and the model-collapsibility. Model- and estimate-collapsibility and their equivalence for conditional graphical models for multi-way contingency tables are considered by Liu and Guo (2012).

4.9.5 *Information-Theoretic Approach in Contingency Table Analysis*

A pioneering approach in categorical data analysis is the *minimum discrimination information* (MDI) approach, based on information theory. It is based on the discrimination information function of Kullback (1959), which is defined on two probability distributions and is a measure of closeness between them.

Based on the principle of MDI, the MDI estimates are BAN estimates obtained by minimizing the discrimination information function between the observed frequencies and the expected under the assumed model or hypothesis. For the cell probabilities of two-way tables with fixed marginals, Ireland and Kullback (1968a) proposed the MDI estimators, illustrating also how their procedure is extended to multi-way tables. Further applications of the results derived in Ireland and Kullback (1968a) and connections to previously ad hoc considered estimators by Fisher (1934) for the 2×2 case are given in Ireland and Kullback (1968b).

The MDI approach offers a complete treatment for categorical data inference. The corresponding statistic is asymptotically X^2 distributed under the assumed

model and is used for testing model fit. Furthermore, the procedure can be applied for testing hypotheses about parameters of the model or linear combinations of them and provides indication of outlier cells and the analysis of information table, in analogy to the analysis of variance table. It is a platform of unified treatment for contingency tables of any order and dimension as well as for categorical data not in a contingency table form. For applications of this approach on contingency tables see Ku and Kullback (1974), references cited therein, and the book by Gokhale and Kullback (1978a). A clarifying short review is given by Gokhale and Kullback (1978b). The MDI approach is identical to the ML approach for *internal constrained problems* (ICP) while for *external constrained problems* (ECP) the two approaches are equivalent in probability under the null hypothesis or the assumed model. For more on ICP and ECP, we refer to Gokhale and Kullback (1978b) and Read and Cressie (1988, Sect. 3.5).

The MDI approach is itself a special case of the *minimum power divergence* approach. The *power divergence* family is introduced by Cressie and Read (1984) and unifies all major approaches considered for discrete multivariate data analysis. Its dynamism lies on the fact that the individual special cases are obtained through a single parameter λ. The power divergence goodness-of-fit statistic for comparing the frequency vector $\mathbf{Y} = (Y_1, \ldots Y_{n_y})'$ to the estimated of the expected under the assumed null hypothesis (or model) $\hat{\alpha} = (\hat{\alpha}_1, \ldots \hat{\alpha}_{n_y})'$ is defined as

$$2I^{\lambda}(\mathbf{Y} : \hat{\alpha}) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{n_y} Y_i \left[\left(\frac{Y_i}{\hat{\alpha}_i} \right)^{\lambda} - 1 \right], \; -\infty < \lambda < \infty, \; \lambda \neq -1, 0 . \quad (4.36)$$

The cases $\lambda = -1$ and $\lambda = 0$ are defined by the continuous limits of (4.36) for $\lambda \to -1$ and $\lambda \to 0$. It forms a parametric family of goodness-of-fit statistics, controlled by the parameter λ. Pearson's X^2 is (4.36) with $\lambda = 1$, while (4.36) converges to the LR statistic G^2 for $\lambda \to 0$. Further, for $\lambda \to -1$, it converges to the MDI statistic mentioned above. The Neyman-modified X^2 statistic (Neyman 1949) is obtained for $\lambda = -2$ and the Freeman–Tukey statistic (Freeman and Tukey 1950) for $\lambda = -1/2$. Under the null hypothesis tested and under certain regularity conditions, (4.36) is asymptotically X^2 distributed and all members of this goodness-of-fit statistics family are asymptotically equivalent. In terms of test power and of small sample approximation, Cressie and Read (1984) suggested the value $\lambda = 2/3$. Statistical inference for multivariate discrete data based on the power divergence is studied extensively in Read and Cressie (1988), also under sparseness assumptions.

Associated with statistic (4.36) is the *power divergence measure*, which measures the divergence of two probability distributions. If $\pi = (\pi_1, \ldots \pi_K)'$ and $\mathbf{q} = (q_1, \ldots q_K)'$ are two probability vectors, then the power divergence specifies their divergence as

$$2I^{\lambda}(\pi : \hat{\mathbf{q}}) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{K} \pi_i \left[\left(\frac{\pi_i}{q_i} \right)^{\lambda} - 1 \right], \; -\infty < \lambda < \infty, \; \lambda \neq -1, 0, \quad (4.37)$$

with the cases $\lambda = -1$ and $\lambda = 0$ being defined as above.

The power divergence belongs to the even broader family of ϕ-divergences. For π and $\mathbf{q}$ as above, the ϕ-divergence between π and $\mathbf{q}$ (or Csiszar's measure of information in $\mathbf{q}$ about π) is defined by

$$\mathrm{I}^{\mathrm{C}}(\pi, \mathbf{q}) = \sum_{i=1}^{K} q_i \phi(\pi_i / q_i), \tag{4.38}$$

where ϕ is a real-valued strictly convex function on $[0, \infty)$ with $\phi(1) = \phi'(1) = 0$, $0\phi(0/0) = 0$, $0\phi(y/0) = \lim_{x\to\infty} \phi(x)/x$ (see Pardo 2006). Setting $\phi(x) = x \log x$, (4.38) is reduced to the Kullback–Leibler divergence measure that corresponds to the LR statistic G^2. For $\phi(x) = (1-x)^2$, Pearson's divergence is derived, related to Pearson's X^2 statistic. If $\phi(x) = \frac{x^{\lambda+1}-x}{\lambda(\lambda+1)}$, (4.38) becomes the power divergence measure (4.37).

For ϕ-divergence-based inference and for special applications to log-linear models and categorical data analysis, we refer to Pardo (2006), references therein, and to Martìn and Pardo (2008). Minimum power divergence and minimum ϕ-divergence estimators generalize the MLEs, retaining their properties and meanwhile exhibiting robustness properties (see Basu et al. 1998 and Pardo 2006). In Sect. 7.4 we discuss generalized association models, connected to ϕ-divergence.

Chapter 5
Generalized Linear Models and Extensions

Abstract The generalized linear model (GLM) is reviewed and the log-linear models are integrated in this family. For GLMs, maximum likelihood estimation, model fit, and model selection are discussed. In the GLM framework the analysis of incomplete tables is more straightforward. The quasi-independence model is defined and illustrated in R. Furthermore, the family of generalized log-linear models (GLLMs) is briefly presented and a GLLM is illustrated with a representative example in R.

Keywords Generalized linear models • Exponential family • Maximum likelihood estimation • Model selection and fit • Log-linear models • Quasi independence • Multinomial Poisson homogeneous model

5.1 The Generalized Linear Model (GLM) in Keywords

Log-linear models for contingency tables are members of the family of *generalized linear models* (GLMs). The GLM is a broad class of statistical models, introduced by Nelder and Wedderburn (1972), that allows for unified consideration and treatment of many models of different types of response variables and error structures. Characteristic special cases of the GLM are the models of regression, logistic regression, Poisson regression, and the log-linear models. The GLM is an extension of the classical regression model that relates a *response variable* Y to a set of q *explanatory variables* X_j, $j = 1, \ldots, q$, by equating a function of the expected response $\mathrm{E}(Y)$ to a linear predictor based on $\mathbf{X} = (X_1, \ldots, X_q)$.

Under the classical linear regression model, if $\mathbf{y} = (y_1, \ldots, y_{n_y})'$ is a sample of size n_y of the response variable Y and $\mathbf{x} = (x_{ij})_{n_y \times q}$ is the $n_y \times q$ matrix with the corresponding sample values on the explanatory variables X_j, $j = 1, \ldots, q$, then in matrix notation we have

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \,,$$

M. Kateri, *Contingency Table Analysis: Methods and Implementation Using R*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4811-4_5,

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)'$ is the parameter vector and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_{n_y})'$, the vector of errors. The distributional assumptions are that (i) Y_i are independent normal distributed with $\mathrm{E}(Y_i) = \propto_i$ $(i = 1, \ldots, n_y)$ and common variance $\mathrm{Var}(Y_i) = \sigma^2$ and (ii) the errors are also independent normal distributed with zero mean and common variance σ_ε^2. In summary, the regression model has a *random component*, the response variable Y, and a *systematic component*, the linear combination of the explanatory variables $\mathbf{X}\boldsymbol{\beta}$, that links to the vector of the expected response values, i.e.,

$$\boldsymbol{\propto} = \mathrm{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \ , \tag{5.1}$$

with $\boldsymbol{\propto} = (\propto_1, \ldots, \propto_{n_y})'$ and $\mathbf{Y} = (Y_1, \ldots, Y_{n_y})'$.

The GLM extends the regression models by relaxing the assumption about normal distributed response variable Y and by linking the systematic component not directly to $\boldsymbol{\propto}$ but to a function of it $g(\boldsymbol{\propto})$. Thus, the systematic component of the GLM is

$$\boldsymbol{\eta} = g(\boldsymbol{\propto}) = \mathbf{X}\boldsymbol{\beta} \ , \tag{5.2}$$

with $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{n_y})'$. Function g is called the *link function*. The linear model (5.1) is a special case of (5.2) for the *identity link*, i.e. for $\boldsymbol{\eta} = g(\boldsymbol{\propto}) = \boldsymbol{\propto}$.

Under GLM, the distribution of the response Y may be any member of the *exponential family*. For univariate responses, as considered in this book, the corresponding density function is

$$f(y_i;\ \theta_i, \psi, \omega_i) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{\psi}\omega_i + c(y_i, \psi, \omega_i) \right\} \ , \tag{5.3}$$

where ω_i is a weight with

$$\omega_i = \begin{cases} 1, & \text{ungrouped data } (i = 1, \ldots, n_y) \\ n_i^c, & \text{grouped data} \quad (i = 1, \ldots, g) \end{cases} \ ,$$

and $c = 1$ or -1, according to whether as group response is considered the average or the sum of the individuals' responses in a group, respectively. Parameter θ is called *natural parameter*, because it determines the mean, since

$$\propto = \mathrm{E}(Y) = b'(\theta) \ . \tag{5.4}$$

Parameter ψ controls the variance

$$\sigma^2 = \mathrm{Var}(Y) = \frac{\psi}{\omega_i} b''(\theta) \tag{5.5}$$

and is therefore called the *dispersion parameter*. $b(\cdot)$ and $c(\cdot)$ are specific functions determined by the type of the exponential family.

Many commonly used distributions are members of the exponential family, like the normal, the gamma, the binomial, the multinomial, and the Poisson. For one-parameter families the dispersion parameter ψ is fixed. For example, the Poisson $\mathscr{P}(\theta)$ and the binomial $\mathscr{B}(n,\theta)$, for fixed n, have $\psi = 1$. These distributions are in the simpler *natural exponential family*. Furthermore, for the Poisson $\omega = 1$ while for the binomial $\omega = n$ or n^{-1}, according to whether as response y is considered the success proportion or the number of successes.

The link function $\eta_i = g(\propto_i)$ can theoretically be any monotonic and differentiable function. However, the link options are practically limited, since the link is chosen so that the inverse $\propto_i = g^{-1}(\eta_i)$ leads to admissible values for $\propto_i$ and simple functions of θ_i. Characteristic example is the case of a binomial response $\mathscr{B}(n,\pi_i)$. Then $\propto_i = \pi_i$ and it must be in $(0,1)$. The three links that are more often used for binomial data are the *logit*, the *probit*, the *complementary log–log* , and the *complementary log*. In Chap. 8, we will apply the logit link $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$ and refer briefly to the other options. The link function specifies the nature of the distribution considered for the error ε_i. A convenient link with nice properties is the *canonical link* that expresses $\propto_i$ in terms of the parameter θ_i, i.e., the canonical link is $g(\propto_i) = B^{-1}(\theta_i)$, where $B = b'$. Under the canonical link, $\mathbf{X}'\mathbf{Y}$ is a *sufficient* statistic for $\boldsymbol{\beta}$.

In summary, GLM is a framework that unifies a wide range of models, flexible through the choices for the distribution of its random component, for the link and eventually the error distribution. Beyond the powerful theoretical setup, it is practically attractive because it allows to draw inference for all possible GLM models by the same algorithm, simplifying thus their implementation in statistical software.

5.2 Log-Linear Model: Member of the GLM Family

Classical log-linear models, presented in Chap. 4, can be viewed in the framework of GLM for specific selection of the link function and the error distribution, as will be stated next. Doing so has specific advantages. Beyond convenience in model selection and inference by adopting the procedures developed for the GLM family, it allows for easy handling of the structural zeros in log-linear modeling (see Sect. 5.5) and it provides a platform for extending the log-linear model to model the marginals as well (see Sect. 5.6).

In order to adjust to GLM's notation, contingency tables are expanded to vectors. Thus, the $I \times J$ table $\mathbf{n} = (n_{ij})$ is expanded (by rows) to the $n_y \times 1$ vector $\mathbf{y}$ as

$$\mathbf{y} = (y_1, y_2, \ldots, y_{n_y})' = (n_{11}, n_{12}, \ldots, n_{1J}, n_{21}, \ldots, n_{IJ})' ,$$

with $n_y = IJ$. Additionally, this vector approach ensures unified treatment for tables of any dimension. Throughout this book whenever tables are expanded in vectors, expansion is considered by rows, followed by columns, layers, etc.

Under the GLM setup, the log-linear models for contingency tables are easier derived considering the Poisson distribution for the random component, i.e., $Y_i \sim \mathscr{P}(\theta_i)$ and for link the $g(\propto_i) = \log \propto_i$, $i = 1, \ldots, n_y$. The *log link* is the canonical link for the Poisson distribution. They are referred as *Poisson log-linear models*. Considering Poisson sampling is not restrictive due to the equivalence of the three possible sampling schemes (see Sect. 2.2.1). Recall that also in the classical log-linear framework, estimation was based on the Poisson likelihood (2.33).

Thus, the log-linear models for $I \times J$ tables discussed in this section can be expressed in matrix notation, as follows:

$$\log(\boldsymbol{\propto}) = \mathbf{X}\boldsymbol{\beta} \,, \tag{5.6}$$

where $\boldsymbol{\propto}$ is the $IJ \times 1$ vector of expected cell frequencies under the model, $\boldsymbol{\beta}$ is the $q \times 1$ vector of parameters, and $\mathbf{X}$ is the $IJ \times q$ associated design matrix. The table of expected cell frequencies $\mathbf{m}_{I\times J}$ is expanded the same way as the table of observed frequencies.

For example, the model of independence (4.1) subject to last category zero constraints is equivalently expressed by (5.6), where the $IJ \times 1$ vector of expected frequencies is $\propto = (m_{11}, m_{12}, \ldots, m_{1J}, m_{21}, \ldots, m_{IJ})'$, the $(I + J - 1) \times 1$ vector of parameters is $\beta = (\lambda, \lambda_1^X, \ldots, \lambda_{I-1}^X, \lambda_1^Y, \ldots, \lambda_{J-1}^Y)'$, and

$$\mathbf{X} = \begin{pmatrix} \mathbf{1} & \mathbf{1}^{(1)} & \mathbf{I}^* \\ \mathbf{1} & \mathbf{1}^{(2)} & \mathbf{I}^* \\ \vdots & \vdots & \vdots \\ \mathbf{1} & \mathbf{1}^{(I-1)} & \mathbf{I}^* \\ \mathbf{1} & \mathbf{0}_{J\times(I-1)} & \mathbf{I}^* \end{pmatrix}$$

is the $IJ \times (I + J - 1)$ design matrix, with $\mathbf{1}$ the $J \times 1$ matrix of 1's, $\mathbf{1}^{(i)}$ the $J \times (I - 1)$ matrix with 1's at the ith column and 0's in all other entries, $\mathbf{0}_{s\times t}$ the $s \times t$ matrix of 0's, and

$$\mathbf{I}^* = \begin{pmatrix} \mathbf{I}_{J-1} \\ \mathbf{0}_{1\times(J-1)} \end{pmatrix} \,,$$

where $\mathbf{I}_s$ is the $s \times s$ identity matrix.

The application of the independence model through local odds ratios (2.52), though simpler in expression, is more advanced and computationally involved, because it is not in the GLM family. It does not apply to the expected cell frequencies directly but to a function of them. For this, a generalization of the GLM is needed, briefly discussed in Sect. 5.6.

5.3 Inference for GLMs

5.3.1 ML Estimation for GLMs

For the maximum likelihood estimation of $\boldsymbol{\beta}$ for model (5.2), the log-likelihood of a given sample needs to be maximized with respect to $\boldsymbol{\beta}$. Thus, for a random sample $\mathbf{y}$ of size n_y, from a population distributed by (5.3), the log-likelihood is

$$\ell = \sum_{i=1}^{n_y} \log f(y_i;\ \theta_i, \psi, \omega_i) = \sum_{i=1}^{n_y} \frac{y_i\theta_i - b(\theta_i)}{\psi}\omega_i + \sum_{i=1}^{n_y} c(y_i, \psi, \omega_i) \tag{5.7}$$

and is a function of $\boldsymbol{\beta}$ due to (5.2) and (5.4).

The first derivative of the log-likelihood function is the *Fisher's score function*

$$s(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \left(\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_1}, \ldots \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_q} \right)' .$$

Equating the score function's components to zero, the corresponding likelihood equations are obtained

$$s(\beta_j) = \frac{\partial \ell}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} \left(\sum_{i=1}^{n_y} \log f(y_i;\ \theta_i, \psi, \omega_i) \right) = 0\ , \quad j = 1, \ldots, q,$$

and are finally equal to

$$\sum_{i=1}^{n_y} \left(\frac{y_i - \mathrm{E}(Y_i)}{\mathrm{Var}(Y_i)} \cdot \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \cdot x_{ij} \right) = 0\ , \quad j = 1, \ldots, q\ , \tag{5.8}$$

where $\eta_i = \sum_{j=1}^{q} \beta_j x_{ij}$. The likelihood equations (5.8) are derived applying the chain rule, since $\theta_i = (b')^{-1}(\propto_i)$, $\propto_i = g^{-1}(\eta_i)$, and using (5.4) and (5.5).

For certain distributional assumption for Y_i and particular link function g, the likelihood equations (5.8) take their explicit form and specify the MLE $\hat{\boldsymbol{\beta}}$. For the canonical link, $\eta_i = \theta_i$ and $g^{-1} = b'$, leading to $\frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} = b''(\theta_i)$. Thus, by (5.5), (5.8) are simplified to

$$\sum_{i=1}^{n_y} [y_i - \mathrm{E}(Y_i)]\, x_{ij} = 0\ , \quad j = 1, \ldots, q\ , \tag{5.9}$$

stating that the likelihood equations for the canonical link equate the β_j's sufficient statistic $\sum_{i=1}^{n_y} y_i x_{ij}$ to its expected value, for $j = 1, \ldots, q$.

The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is derived from the second derivative of the log-likelihood, since it is equal to

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \mathscr{I}_F^{-1}\ ,$$

where $\mathscr{I}_F = \mathrm{Cov}(s(\boldsymbol{\beta}))$ is the *expected Fisher information matrix*. In our case

$$\mathscr{I}_F = \mathrm{Cov}(s(\boldsymbol{\beta})) = \mathrm{E}\left(\frac{\partial \ell}{\partial \boldsymbol{\beta}} \frac{\partial \ell}{\partial \boldsymbol{\beta}'}\right) = \mathrm{E}\left(-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right) = \mathbf{X}'\mathbf{W}\mathbf{X} \,,$$

where $\mathbf{W}$ is a diagonal matrix with diagonal entries

$$w_i = (\partial \circ_i / \partial \eta_i)^2 [\mathrm{Var}(Y_i)]^{-1} \,. \tag{5.10}$$

For large n_y,

$$\hat{\boldsymbol{\beta}} \sim \mathscr{N}_q(\boldsymbol{\beta},\ \mathscr{I}_F^{-1}) \,.$$

The matrix of the negative second derivatives of the score function is the *observed information matrix*

$$\mathscr{I}_F^{obs} = -\mathbf{H} = -\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \,,$$

where the matrix of second derivatives $\mathbf{H}$ is usually referred as the *Hessian* matrix. It holds that

$$\mathscr{I}_F = \mathrm{E}\left(\mathscr{I}_F^{obs}\right) = \mathrm{E}\left(-\mathbf{H}\right) \,. \tag{5.11}$$

For GLMs with canonical link functions, $\eta_i = \theta_i$ implies $\frac{\partial \circ_i}{\partial \eta_i} = \frac{\partial \circ_i}{\partial \theta_i}$ and the Hessian matrix becomes

$$\mathbf{H} = -\mathbf{X}'\mathbf{W}\mathbf{X} \,, \tag{5.12}$$

with $\mathbf{W}$ a diagonal matrix with entries $w_i = \omega_i \left[g^{-1}(\theta_i)\right]' / \psi$, $i = 1, \ldots, n_y$, independent of $\mathbf{y}$. Hence

$$\mathscr{I}_F = \mathrm{E}\left(-\mathbf{H}\right) = -\mathbf{H} = \mathscr{I}_F^{obs} \,,$$

i.e., for canonical link functions, the expected and observed information matrices are identical.

The likelihood equations (5.8) or (5.9) do not usually lead to closed form expressions for the $\hat{\boldsymbol{\beta}}$ and have to be solved iteratively. The two algorithms usually applied for solving the likelihood equations are the *Newton–Raphson* and the *Fisher scoring*.

If $\boldsymbol{\beta}^{(t)}$ is the value assigned to $\hat{\boldsymbol{\beta}}$ at stage t of the iterative procedure ($t = 0, 1, 2, \ldots$), then the updating equations of the Newton–Raphson algorithm at stage $t+1$ are

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t - \left(\mathbf{H}^{(t)}\right)^{-1} s(\boldsymbol{\beta}^{(t)}) \,, \tag{5.13}$$

where $s(\boldsymbol{\beta}^{(t)})$ and $\mathbf{H}^{(t)}$ are the score function $s(\boldsymbol{\beta})$ and the Hessian matrix $\mathbf{H}$ evaluated at $\boldsymbol{\beta}^{(t)}$. For matrix inversion to be possible, $\mathbf{H}^{(t)}$ has to be non-singular.

The algorithm converges and stops when a termination criterion is met, say after t_c iterations, leading to $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(t)}$. A termination criterion checks whether $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\beta}^{(t+1)}$ are sufficiently close, for example, whether

$$|\ell(\boldsymbol{\beta}^{(t_c+1)}) - \ell(\boldsymbol{\beta}^{(t_c)})| \leq \varepsilon \quad \text{or} \quad \frac{\left\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\right\|}{\left\|\boldsymbol{\beta}^{(t)}\right\|} \leq \varepsilon \,,$$

for a pre-chosen small positive ε.

The Fisher's scoring algorithm is similar to the Newton–Raphson algorithm with the only difference being that it is based on the expected information matrix, instead of the observed information matrix. In particular, the updating equations for the Fisher scoring algorithm are

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^{t} + \left(\mathscr{I}_F^{(t)}\right)^{-1} s(\boldsymbol{\beta}^{(t)}) \,, \tag{5.14}$$

where $\mathscr{I}_F^{(t)}$ is $\mathscr{I}_F$ evaluated at $\boldsymbol{\beta}^{(t)}$.

The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is estimated for the Fisher's scoring algorithm by $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = \hat{\mathscr{I}}_F^{-1}$ and for the Newton–Raphson algorithm by $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = (-\hat{\mathbf{H}})^{-1}$, where $\hat{\mathscr{I}}_F$ and $\hat{\mathbf{H}}$ are $\mathscr{I}_F$ and $\mathbf{H}$, respectively, evaluated at $\hat{\boldsymbol{\beta}}$.

Due to (5.11), the Newton–Raphson and the Fisher scoring algorithm coincide for GLMs of canonical link function. For noncanonical link functions, the choice between the algorithms relates to issues of ease of application, algorithm's convergence, and efficiency of implementation. It is a choice between observed and expected information matrix. For a related discussion, we refer to the classical discussion paper by Efron and Hinkley (1978) and Palmgren (1981). Alternatively, other methods have been proposed like the *Quasi-Newton* (or *Newton's unidimensional*) method that is easier to apply since it does not require matrix inversion but does not provide estimate of the asymptotic covariance matrix. We will illustrate the Newton's unidimensional method for association models in Sect. 6.2.

The solutions of the likelihood equations correspond actually to local maxima and not to the global maximum of the log-likelihood function ℓ, as is expected for the MLE $\hat{\boldsymbol{\beta}}$. Whenever ℓ is concave, the local and global maxima are identical. For non-concave ℓ, the choice of the initial estimate $\boldsymbol{\beta}^{(0)}$ is important, to ensure that it is in the region of the global maxima.

5.3.2 *Evaluating Model Fit for GLMs*

Given a sample $\mathbf{y}$ of n_y observations, let $\hat{\boldsymbol{\alpha}}$ denote the corresponding ML estimate of $\boldsymbol{\alpha} = \text{E}(\mathbf{Y})$ under a model $\mathscr{M}$ of q parameters. The quality of the model fit is assessed by comparing the maximum log-likelihood for the model $\ell(\hat{\boldsymbol{\alpha}}; \mathbf{y})$ to the maximum log-likelihood for the model that describes the data perfectly, i.e., the

saturated model. A saturated model has as many parameters as the observations in the sample. We have seen so far saturated models in the context of log-linear models. For the saturated GLM, the number of parameters is n_y, $\hat{\boldsymbol{\alpha}} = \mathbf{y}$ and the corresponding log-likelihood is $\ell(\mathbf{y};\mathbf{y})$. It is obvious that always $\ell(\hat{\boldsymbol{\alpha}};\mathbf{y}) < \ell(\mathbf{y};\mathbf{y})$ with model $\mathscr{M}$ fitting as better as its log-likelihood approaches the saturated log-likelihood. Hence, the goodness of fit of a model is expressed in terms of their difference by the test statistic

$$-2\left[\ell(\hat{\boldsymbol{\alpha}};\mathbf{y}) - \ell(\mathbf{y};\mathbf{y})\right] ,$$

which for the exponential family (5.3) becomes

$$\frac{D(\mathbf{y};\hat{\boldsymbol{\alpha}})}{\psi} = \frac{2}{\psi}\sum_{i=1}^{n_y} \omega_i \left(y_i(\tilde{\theta}_i - \hat{\theta}_i) - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)]\right) , \tag{5.15}$$

where $\hat{\theta}_i$ is the ML estimate of parameter θ_i under the model $\mathscr{M}$ and $\tilde{\theta}_i$ is the estimate under the saturated model. The statistic $D(\mathbf{y};\hat{\boldsymbol{\alpha}})$ is known as *deviance*. Analogously, the Pearson's X^2 statistic can be used for testing the adequacy of model $\mathscr{M}$. In this context

$$X^2(\mathscr{M}) = \sum_{i=1}^{n_y} \frac{(y_i - \hat{\alpha}_i)^2}{\hat{\alpha}_i} . \tag{5.16}$$

For Poisson and binomial GLMs, the deviance (5.15) turns out to equal the LR statistic for testing the null hypothesis that model $\mathscr{M}$ holds against the saturated model

$$G^2(\mathscr{M}) = 2\sum_{i=1}^{n_y} y_i \log(\frac{y_i}{\hat{\alpha}_i}) . \tag{5.17}$$

The statistics above can be used for testing goodness of fit of $\mathscr{M}$, if their asymptotic distribution can be specified. For this to be possible, the data have to be grouped (each y_i occurs n_i times) with the number of observations in each group n_i being sufficiently large. In this case, the distribution for the statistics (5.15)–(5.17) is approximately $\mathscr{X}^2_{df}$, with $df = n_y - q$, the difference between the number of parameters for the saturated model (n_y) and the model under testing (q). For more on the test statistics refer to McCullagh and Nelder (1989).

These goodness-of-fit tests do not account for model complexity while they are increasing in sample size n_y, giving thus significant values even for good models if the sample size is large. Alternatively, the fit of a model $\mathscr{M}$ can be evaluated by *Akaike's information criterion* (Akaike 1974)

$$AIC = -2\ell(\hat{\boldsymbol{\alpha}};\mathbf{y}) + 2q . \tag{5.18}$$

It is based on the maximum likelihood under $\mathscr{M}$ but penalizes its value for model complexity. Furthermore, the *Bayesian information criterion* (Schwarz 1978)

$$BIC = -2\ell(\hat{\boldsymbol{\alpha}};\mathbf{y}) + (\log n)q \tag{5.19}$$

is another maximum likelihood-based measure, incorporating Bayesian thinking, that beyond complexity takes into account also the sample size n. The *AIC* and *BIC* are used for comparing models, with smaller values indicating better models. They can be used to compare also non-nested models. They will be illustrated in the log-linear model context in Sect. 5.4.1.

5.3.3 Residuals

Residuals are critical for diagnosing lack of model fit and identifying possible underlying patterns. The types of residuals used in GLM analysis are the same as those discussed in the context of independence testing for two-way tables (see Sect. 2.2.4). In the GLM setup, the raw residuals $e_i = y_i - \hat{\alpha}_i$ $(i = 1,\ldots,n_y)$ are transformed to the Pearsonian residuals

$$e_i^P = \frac{y_i - \hat{\alpha}_i}{\sqrt{\widehat{\mathrm{Var}}(y_i)}}\,, \quad i = 1,\ldots,n_y\,. \tag{5.20}$$

For the Poisson GLM, $\widehat{\mathrm{Var}}(y_i) = \hat{\alpha}_i$ in (5.20) above, while for testing independence in two-way tables, (5.20) is (2.40), expressed in vector form. Pearson's residuals are asymptotic normal distributed but not standard normal, as explained in Sect. 2.2.4. Thus, dividing the raw residuals by their asymptotic standard errors, the standardized residuals are derived

$$e_i^s = \frac{e_i^P}{\sqrt{1-\hat{h}_i}} = \frac{e_i}{\sqrt{\widehat{\mathrm{Var}}(y_i)(1-\hat{h}_i)}}\,, \quad i = 1,\ldots,I,\;\; j = 1,\ldots,J\,, \tag{5.21}$$

where $\hat{h}_i$ is the estimate of the diagonal element h_i, $i = 1,\ldots,n_y$ of the $n_y \times n_y$ matrix

$$\mathbf{Hat} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{W}^{1/2}\,,$$

known as *hat matrix*, with $\mathbf{W}$ the diagonal matrix with entries (5.10).

The *deviance residuals* decompose the deviance to the individual contributions of each observation i. Hence, for the exponential family (5.3), they are equal to

$$e_i^d = \mathrm{sign}(y_i - \hat{\alpha}_i)\cdot\left[2\omega_i\left(y_i(\tilde{\theta}_i - \hat{\theta}_i) - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)]\right)\right]^{1/2}\,, \quad i = 1,\ldots,n_y\,, \tag{5.22}$$

satisfying $D(\mathbf{y};\hat{\boldsymbol{\alpha}}) = \sum_i^{n_y}\left(e_i^d\right)^2$. For testing independence in two-way tables, (5.22) simplify to (2.43).

5.3.4 Model Selection in GLMs

Deviance plays a predominant role in comparing GLMs, via the likelihood ratio criterion, for responses y_i, $i = 1, \ldots, n_y$, in the exponential family with $\psi = 1$. In this case, by (5.15), the deviance of a model is equal to the corresponding LR statistic (4.33) for testing its fit.

Let $\mathscr{M}_1$ be a GLM of q_1 parameters. Let also $\mathscr{M}_0$ be a simpler GLM, produced from $\mathscr{M}_1$ by eliminating r of its q_1 parameters. Then, $\mathscr{M}_0$ is said to be *nested* in $\mathscr{M}_1$ and denoted by $\mathscr{M}_0 \subset \mathscr{M}_1$. Model $\mathscr{M}_0$ has $q_0 = q_1 - r$ parameters and is more parsimonious than $\mathscr{M}_1$.

If $\hat{\boldsymbol{\alpha}}_0$ and $\hat{\boldsymbol{\alpha}}_1$ are the ML estimates of $\boldsymbol{\alpha}$ under $\mathscr{M}_0$ and $\mathscr{M}_1$, respectively, then, for $\psi = 1$, the deviances of models $\mathscr{M}_0$ and $\mathscr{M}_1$ are

$$\begin{aligned} D(\mathbf{y}; \hat{\boldsymbol{\alpha}}_0) &= -2\left[\ell(\widehat{\boldsymbol{\alpha}_0}; \mathbf{y}) - \ell(\mathbf{y}; \mathbf{y})\right] \\ D(\mathbf{y}; \hat{\boldsymbol{\alpha}}_1) &= -2\left[\ell(\widehat{\boldsymbol{\alpha}_1}; \mathbf{y}) - \ell(\mathbf{y}; \mathbf{y})\right] . \end{aligned}$$

Since reducing the number of model's parameters implies increase of model's distance from the perfect fit of the saturated model, it will always be $D(\mathbf{y}; \hat{\boldsymbol{\alpha}}_0) > D(\mathbf{y}; \hat{\boldsymbol{\alpha}}_1)$.

Models $\mathscr{M}_0$ and $\mathscr{M}_1$ apply both on the same $\mathbf{y}$, thus their difference is

$$D(\mathbf{y}; \hat{\boldsymbol{\alpha}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\alpha}}_1) = -2\left[\ell(\widehat{\boldsymbol{\alpha}_0}; \mathbf{y}) - \ell(\widehat{\boldsymbol{\alpha}_1}; \mathbf{y})\right] = \mathrm{LRS}(\mathscr{M}_0, \mathscr{M}_1) ,$$

where $\mathrm{LRS}(\mathscr{M}_0, \mathscr{M}_1)$ is the LR statistic for testing the null hypothesis that $\mathscr{M}_0$ holds against the alternative that $\mathscr{M}_1$ holds. In particular, by (5.15), the difference in deviances equals

$$D(\hat{\boldsymbol{\alpha}}_0; \hat{\boldsymbol{\alpha}}_1) = D(\mathbf{y}; \hat{\boldsymbol{\alpha}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\alpha}}_1) = 2\sum_{i=1}^{n_y} \omega_i \left(y_i(\hat{\theta}_{i1} - \hat{\theta}_{i0}) - [b(\hat{\theta}_{i1}) - b(\hat{\theta}_{i0})]\right) . \tag{5.23}$$

Under $\mathscr{M}_0$, (5.23) is approximately $\mathscr{X}^2_r$ distributed, where $r = q_1 - q_0$ is the difference between the number of parameters of the two compared models. This asymptotic result is the key for models' comparison.

For Poisson log-linear models, (5.23) simplifies to (4.34), i.e.,

$$G^2(\mathscr{M}_0 | \mathscr{M}_1) = 2\sum_{i=1}^{n_y} \hat{\alpha}_{i1} \log\left(\frac{\hat{\alpha}_{i1}}{\hat{\alpha}_{i0}}\right) = G^2(\mathscr{M}_0) - G^2(\mathscr{M}_1) ,$$

where $G^2(\mathscr{M}_0)$ and $G^2(\mathscr{M}_1)$ are as in (5.17).

Upon considering a sequence of nested models from a very simple $\mathscr{M}_0$ up to the saturated $\mathscr{M}_{\mathrm{sat}}$,

$$\mathscr{M}_0 \subset \mathscr{M}_1 \subset \mathscr{M}_2 \subset \ldots \subset \mathscr{M}_{\mathrm{sat}} ,$$

the importance of the parameters added gradually can be evaluated by successive comparisons of neighbor models. Thus, the appropriate model can be built by selecting this model $\mathscr{M}_s$ for which $D(\hat{\boldsymbol{\alpha}}_s; \hat{\boldsymbol{\alpha}}_{s+1})$ is nonsignificant and $D(\hat{\boldsymbol{\alpha}}_{s-1}; \hat{\boldsymbol{\alpha}}_s)$ is significant. This means that adding more parameters would complicate the model without improving its fit significantly, while removing any parameters further would lead to a model of significantly poorer fit. Hence, comparisons of nested models serve for developing procedures of "best model" selection. Furthermore, once the "best model" is selected, model comparison can serve as a tool for evaluating the individual importance of each parameter or group of parameters. Model selection can also be based on *AIC* and *BIC*. For a comparative study of *AIC* and *BIC* and a corrected for finite samples version of *AIC* with emphasis on their role in model selection, we refer to Burnham and Anderson (2004). These criteria will be illustrated in the context of log-linear models for multi-way tables next (see Sect. 5.4.1).

5.4 Software for GLMs

All general-purpose statistical packages (like SAS, SPSS, Stata, and SYSTAT) have procedures for GLM analysis. For example, GLMs are fitted in SAS by the procedure GENMOD. The corresponding `R` function is `glm`, which is based on the S-function "glm" (Hastie and Pregibon 1992). The basic form for calling the `glm` function is

```
> Mfit <- glm(formula, family=..., data=...)
```

where `formula` defines the model to be fitted, `family` determines the error distribution and link function of the model, and `data` specifies the data frame on which the model will be applied. `Mfit` is the object where output of `glm` is saved. `formula` is provided in a form of the type `Y~X1+X2+X3+X1:X2`, where `Y` is the dependent variable, `X1, X2, X3` the independent, and `X1:X2` denotes the interaction between `X1` and `X2`. The expression above is equivalent to `Y~X3+X1*X2`, where `X1*X2` stands for the generating term of a hierarchical model, i.e., it is equivalent to `Y~X1+X2+X1:X2`. For log-linear models the choice for `family` is `family=poisson (link = "log")`. The specification of data frame is optional. If it is omitted, the variables are taken from the environment from which `glm` is called.

The minimum output is printed on screen by simply typing `Mfit` while more detailed output is provided by `summary(Mfit)`. The content of object `Mfit` can be viewed by `names(Mfit)`. An item, say `A`, of `Mfit` is located in `Mfit$A` and can be saved in a variable for further use (e.g., `V1 <- Mfit$A`). Due to the predominant role deviance plays in GLM's analysis, the residuals saved in `Mfit`, the output object of `glm`, are the deviance residuals. For results not provided in `Mfit`, a variety of special functions is available that apply on the `glm` output. Function `step()` for model selection between nested models and `anova()` for analysis of variance can be activated also in `glm` framework, as will be illustrated in the examples that follow.

Table 5.1 Summary output of the independence model applied on Table 2.3, fitted by `glm`

```
Call:
glm(formula = freq ~ WELFARE + DEGREE, family = poisson,data = nt.frame)

Deviance Residuals:
Min                  1Q  Median              3Q  Max
-1.3419         -0.5377  -0.1352          0.3366  1.6724

Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  3.54654    0.10253    34.590    < 2e-16 ***
WELFARE2     0.32962    0.08276     3.983    6.81e-05 ***
WELFARE3     0.34666    0.08247     4.204    2.63e-05 ***
DEGREE2      1.26567    0.09855    12.843    < 2e-16 ***
DEGREE3     -0.42845    0.13858    -3.092     0.00199 **
DEGREE4      0.30458    0.11473     2.655     0.00793 **
DEGREE5     -0.38299    0.13670    -2.802     0.00508 **
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

( Dispersion parameter for poisson family taken to be 1 )

   Null deviance: 478.046 on 14 degrees of freedom

Residual deviance: 10.363 on 8 degrees of freedom

AIC: 110.74

Number of Fisher Scoring iterations: 4
```

For historical reasons, let us note that GLIM (generalized linear interactive modeling) was the first package with the ability of fitting a variety of GLMs in a unified manner. It was developed by the GLIM working party of the Royal Statistical Society in 1974. GLIM4, the latest release (1993), had many links as standard options and was convenient for GLM fit and model selection. A rich macro library was available while users could write their own macros in GLIM language. The associated journal *GLIM Newsletter*, issued from 1979 to 1998, was publishing GLIM macros.

5.4.1 Example 2.4 by `glm`

The log-linear model of independence (4.1) will be fitted on Table 2.3, by `glm` of `R`. The variables are required in vector form; thus we apply `glm` on the data frame `nt.frame`, constructed in Sect.4.2.1. Model (4.1) is then fitted by

```
> I.glm <- glm(freq ~ WELFARE+DEGREE,family=poisson,data=nt.frame)
```

and the extended output (provided in Table 5.1) is obtained by

```
> summary(I.glm)
```

The value of the G^2 statistic is reported under "Residual Deviance" and is saved in `I.glm$deviance`, as can be verified by typing `names(I.glm)`. Its asymptotic p-value is not provided but can easily be calculated by

```
> p.value <- 1-pchisq(I.glm$deviance, I.glm$df.residual)
```

We find p-value = 0.240; thus the independence model describes adequately this data set. Furthermore the value of the AIC is given ($AIC = 110.74$) while the BIC, defined by (5.19), can be computed as

```
> n <- sum(ntfare$freq);  q <- I.glm$df.null-I.glm$df.residual
> BIC <- I.glm$aic-(2-log(n))*q
```

giving $BIC = 139.91$. The level of the AIC and BIC values can be judged in comparison to alternative models. In this case, for the saturated model

```
> sat <- glm(freq ~ WELFARE*DEGREE,family=poisson,data=nt.frame)
```

$AIC = 116.4$, while for the models of only one main effect

```
> welfr <- glm(freq ~ WELFARE,family=poisson,data=nt.frame)
```

and

```
> degr <- glm(freq ~ DEGREE,family=poisson,data=nt.frame)
```

we get $AIC = 548.1$ and $AIC = 129$, respectively. Hence, the choice of the independence model is justified.

Function `glm` produces parameter estimates subject to the first category zero constraints. Recall that only the effect differences between different categories are of interest and these remain invariant under different types of constraints. Observe that $\hat{\lambda}_3^X - \hat{\lambda}_1^X = 0.347 - 0$, equal to the corresponding value derived in Sect. 4.2.1 subject to the sum to zero constraints.

The residuals saved in object `I.glm` are the working residuals. The Pearsonian residuals are calculated by `residuals(I.glm, type = c("pearson"))` and the deviance by changing the type option to `"deviance"`. Standardized residuals are obtained by `rstandard(I.glm)`.

The items of the output object are all in vector form but can easily be transformed to the more friendly table form by `xtabs()`. For example, the ML estimates of the expected cell frequencies under independence and the standardized residuals are derived in table form by

```
> MLEs <- xtabs(I.glm$fitted.values ~ WELFARE+DEGREE,data=natfare)
> stdres <- xtabs(rstandard(I.glm) ~ WELFARE+DEGREE,data=natfare)
```

Thus, the standardized residuals are

```
> stdres
```

	DEGREE				
WELFARE	LT HS	HS	JColg	BA	Grad
too little	2.0983151	-1.039894	-0.9517240	0.1790943	-0.1654438
about right	-1.6533505	-0.543633	0.3659428	0.4422752	1.7955727
too much	-0.4040979	1.462390	0.4702723	-0.6127615	-1.7788921

The only standardized residual that exceeds in absolute value 1.96 corresponds to cell $(1,1)$. That is, responders with educational level lower than high school tend to believe that welfare spending is too little with higher probability than expected under the independence model.

The sequence of commands followed above is unified in function `fit.I()` of the web appendix (see Sect. A.3.4), which additionally provides the values for Pearson's X^2 along with its p-value, the dissimilarity index (4.18) and the *BIC*. The function requires the vector of frequencies (by rows) and the number of rows and columns of the table. For this example, it is called as `fit.I(freq,3,5)`.

The standardized residuals can be displayed on the mosaic plot as shown below. We apply

```
> mosaic(natfare, gp=shading_Friendly, residuals=stdres,
+   residuals_type="Std\nresiduals",labeling = labeling_residuals)
```

where `stdres` is the table of standardized residuals derived above. The mosaic plot derived is given in Fig. 5.1 (right). The figure on the left is the mosaic plot for standardized residuals for Example 2.2 and is derived analogously.

The residuals illustrated in the mosaic plots so far were all for the independence model (default). To refer to residuals of a different model, the output object of the assumed model has to take the position of the data matrix as input in `mosaic()`. Thus,

```
> mosaic(natfare, gp=shading_hcl, residuals_type="deviance")
```

is equivalent to

```
> mosaic(I.glm, gp=shading_hcl, residuals_type="deviance")
```

To incorporate the residuals of the model with only the row (opinion) main effect

```
> X.glm <- glm(freq ~ WELFARE+DEGREE,family=poisson,data=nf.frame)
```

the mosaic plot function should be

```
> mosaic(X.glm, gp=shading_hcl, residuals_type="deviance")
```

From the ML estimates it can be verified that the estimated under independence $\hat{\theta}_{ij}$ ($i = 1,2$, $j = 1,\ldots,4I$) are, as expected, all equal to 1. The same holds also for the global and cumulative odds ratios. The ML estimates of any set of generalized odds ratios expected under the assumed model can be calculated in `R`, using the corresponding functions of the web appendix (see Sect. A.3.2). The procedure is that described for the sample generalized odds ratios at the end of Sect. 2.2.5 and illustrated in the example of Sect. 2.2.6. Only the vector of observed frequencies has to be replaced with the vector of ML estimates of the expected cell frequencies under the assumed model. The equivalent independence model (2.52) in terms of the local odds ratios will be illustrated for this example in Sect. 5.6.

5.4.2 *Example 3.1 (Revisited)*

For the example of Table 3.1, we have seen in Sect. 3.3, applying the Breslow–Day test (or the Woolf test), that the association between smoking and depression is homogeneous for males and females. At this point, we shall select the appropriate log-linear model for describing the underlying association structure of Table 3.1. The data are available in `R` in matrix `depsmok3`. In order to fit the models in the GLM setup applying `glm`, the data have to be expanded from a matrix to a vector and the factors corresponding to the classification variables have to be defined. This is carried out easily as follows:

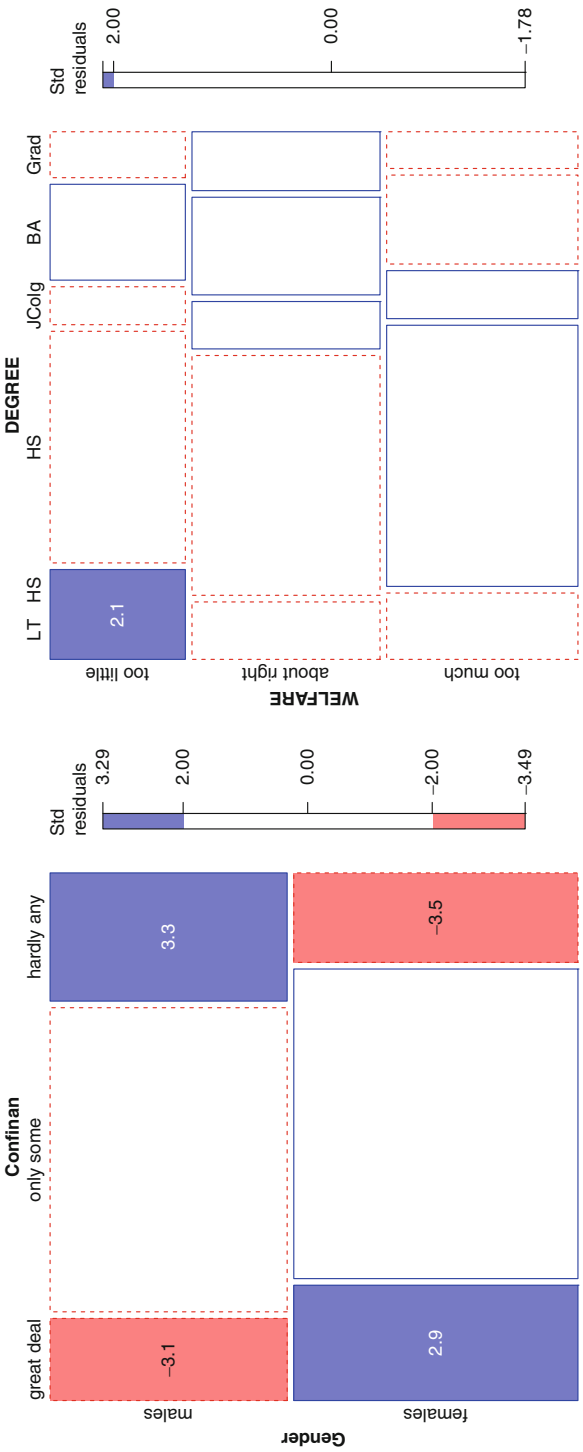


Fig. 5.1 Mosaic plots of standardized residuals for the independence model applied on Table 2.2 (*left*) and Table 2.3 (*right*)

```
> obs <- as.vector(depsmok3)
> row <- rep(1:2, 4);  col <- rep(1:2, each=2,2)
> lay <- rep(1:2, each=4);  row.lb <- c("yes","no")
> col.lb <- c("yes","no");  lay.lb <- c("male", "female")
> S <- factor(row,labels=row.lb);  D <- factor(col,labels=col.lb)
> G <- factor(lay, labels=lay.lb)
> depres.fr <- data.frame(obs,S,D,G)
```

The appropriate log-linear model is selected via the backward stepwise procedure based on *AIC*. Thus, we first save the saturated model under object `saturated` and then proceed with the backward model selection procedure as follows:

```
> saturated <- glm(freq S*D*G, poisson, data = depres.fr)
> step(saturated, direction="backward")
```

The stepwise procedure concludes to the model of no three-factor interaction (*SD*, *DG*, *SG*), giving the following output:

```
Start: AIC=71.38
freq  S * D * G:
                Df          Deviance   AIC
- S:D:G          1           0.77135   70.155
<none>                       0.00000   71.384
Step: AIC=70.16
freq ~ S + D + G + S:D + S:G + D:G
                Df          Deviance     AIC
<none>                         0.771     70.155
- S:D            1            33.024     100.408
- D:G            1            34.386     101.769
- S:G            1           112.298     179.682
Call: glm(formula = freq~S+D+G+S:D+S:G+D:G, family=poisson,
data=depres.fr)

Coefficients:

Intercept       Sno               Dno   Gfemale
3.7393        -1.6684          3.0485   0.8850
Sno:Dno     Sno:Gfemale   Dno:Gfemale
0.9187        0.7834         -0.9369

Degrees of Freedom: 7 Total (i.e. Null); 1 Residual
Null Deviance:  3315
Residual Deviance: 0.7713   AIC: 70.16
```

The (*SD*, *DG*, *SG*) is also the model of *homogeneous association* since under this model the association in all two-way partial tables is homogeneous across the levels of the remaining third classification variable, as explained in Sect. 4.3. This model is fitted in R, as shown below, giving the output provided in Table 5.2.

```
> hom.assoc <- glm(freq~S*D+S*G+D*G, poisson,data=depres.fr);
summary(hom.assoc)
```

The p-value of testing the model fit based on G^2 statistic is 0.380, which is close to the corresponding p-values of the Woolf's or the Breslow–Day test (Sect. 3.3.3).

Table 5.2 Output for model (*SD*, *DG*, *SG*), fitted on data in Table 3.1

```
Call:
glm(formula = freq~S*D+S*G+D*G, family=poisson, data=depres.fr)

Deviance Residuals:
      1       2       3        4       5       6        7
-0.32157 0.70555 0.06943 -0.10112 0.20418 -0.32157 -0.07131
      8
0.07006
Coefficients:
             Estimate   Std. Error   z value     Pr(>|z|)
(Intercept)   3.73930      0.14417    25.936     < 2e-16 ***
SNo          -1.66844      0.17668    -9.443     < 2e-16 ***
DNo           3.04847      0.14705    20.731     < 2e-16 ***
Gfemale       0.88501      0.16620     5.325     1.01e-07 ***
SNo:DNo       0.91871      0.17059     5.385     7.23e-08 ***
SNo:Gfemale   0.78344      0.07529    10.405     < 2e-16 ***
DNo:Gfemale  -0.93691      0.17055    -5.493     3.94e-08 ***
--
Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 3315.40325 on 7 degrees of freedom

Residual deviance: 0.77135 on 1 degrees of freedom

AIC: 70.155

Number of Fisher Scoring iterations: 4
```

Relation (4.27), adjusted in our setup, becomes

$$\log \theta_{(k)}^{SD} = \log\left(\frac{\pi_{11|k}\pi_{12|k}}{\pi_{21|k}\pi_{22|k}}\right) = \lambda_{22}^{SD} = \log \theta^{SD} \ , \ k = 1,2 \ ,$$

due to the identifiability constraints $\lambda_{11}^{SD} = \lambda_{12}^{SD} = \lambda_{21}^{SD} = 0$. Thus, the ML estimate of the common odds ratio θ^{SD} under the log-linear model of homogeneous association is

$$\hat{\theta}^{SD} = \exp\left(\hat{\lambda}_{22}^{SD}\right) = \exp(0.91871) = 2.506 \ ,$$

close in value to $\hat{\theta}_{MH}$ and $\hat{\theta}_W$, calculated in Sect. 3.3.3.

Furthermore, the asymptotic Wald $(1-\alpha)100\%$ CI for θ^{SD} is

$$\exp\left[\log \hat{\theta}^{SD} \pm z_{\alpha/2} s.e.\left(\log \hat{\theta}^{SD}\right)\right] \ ,$$

where $s.e.(\log \hat{\theta}^{SD})$ is the standard error of $\log \hat{\theta}^{SD}$ and is equal to $s.e.(\log \theta^{SD}) = s.e.(\lambda_{22}^{SD}) = 0.17059$.

This CI can easily be computed via the function

```
> CI <- function(t, SE, conf.level=0.95)
                  {exp(t+c(-1,1)*qnorm(0.5*(1+conf.level))*SE)}
```

with `t` and `SE` standing for $\log \hat{\theta}^{SD}$ and its standard error, respectively. Hence, the 95% CI for θ^{SD} in this case is computed as

```
> logSD <- 0.91871 ; SE.SD <- 0.17059
> CI(logSD, SE.SD)
[1] 1.793842  3.501041
```

The `xtabs()` function, used in the previous example (Sect. 5.4.1), is especially useful in multi-way tables, since it provides a straightforward way to extract marginal and partial tables of observed or expected cell frequencies. In this example for instance, the smoking-depression marginal table of the ML estimates of the expected cell frequencies under (SD, DG, SG) is

```
> MLE.SD <- xtabs(hom.assoc$fitted.values ~ S + D)
```

and, as expected, coincides with the corresponding marginal table of observed frequencies, which for arrays is obtained by

```
> margin.table(depsmok3, c(1,2))
```

or

```
> apply(depsmok3, c(1,2), sum)
```

However, were the data available only in the data frame format (`depres.fr`), with `obs` the vector of observed frequencies, then the smoking-depression observed marginal table would be

```
> MLE.SD <- xtabs(obs ~ S + D)
```

5.5 Independence for Incomplete Tables

In case of structural zeros existence (see also Sect. 4.9.1), the corresponding cells are of zero probability and must be excluded from the analysis . Thus, any model assumed will not apply on all cells of the contingency table under consideration but only on the subset of its nonstructural zero cells. Hence, structural zeros affect the assumed model in substance. A table with structural zeros is known as an *incomplete* or *truncated table*.

As an illustration, we will consider the independence model for an $I \times J$ table. Independence is considered not for all IJ cells but only for the subset of the nonstructural zero cells $S = \{(i, j) : \pi_{ij} > 0\}$. The model of independence applied on an incomplete table is known as the *quasi-independence* (QI) model, term introduced by Goodman (1968).

QI is defined naturally in the log-linear models framework, as the classical model of independence (4.1), applied on a subset S of the table

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y \ , \qquad (i, j) \in S \ . \tag{5.24}$$

The main effect parameters satisfy the identifiability constraints (4.4), and the associated df are $df = (I-1)(J-1) - s$, where $s = IJ - |S|$ is the number of structural zeros, i.e., the cardinality of the set of structural zeros S^c .

The restriction $(i,j) \in S$ can be incorporated in the model by introducing s additional parameters in (4.1), one for each structural zero. Hence, (5.24) is equivalent to

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + q_{ij} \mathrm{I}_{ij}^{S^c} \ , \ \ i = 1, \ldots, I, \ j = 1, \ldots, J \ , \tag{5.25}$$

where $\mathrm{I}_{ij}^{S^c}$ is the indicator function for structural zeros

$$\mathrm{I}_{ij}^{S^c} = \begin{cases} 1 \ , \ (i,j) \notin S \\ 0 \ , \ (i,j) \in S \end{cases} .$$

This way, the structural zero cells equal the observed counts ($n_{ij} = m_{ij} = 0$ for $(i,j) \notin S$), sacrificing thus $s \ df$. Structural zeros have no contribution to the value of the X^2 or G^2 test statistic.

QI is expressed directly on the cell probabilities, as

$$\pi_{ij} = \alpha_i \beta_j \ , \qquad (i,j) \in S \ ,$$

where the marginal parameters are no more the marginal probabilities.

Additionally, structural zeros serve as a powerful tool in contingency table analysis, since they can be activated by the needs of the analysis to exclude a specific cell or region of the table that is nonzero but exhibits "special behavior" and exacerbates the fit of the assumed model. This is often the case for mobility tables or panel studies, where the tables are square with augmented diagonal entries, corresponding to non-change. It is natural thus to exclude the diagonal from the analysis by considering $S = \{(i,j) : \ i \neq j\}$. Other incomplete square tables that received special attention are triangular tables. We will return to special QI models for square tables in Sect. 9.3. References on conditions for existence of ML estimates for truncated tables are provided in Sect. 5.7.1.

Structural zeros are incorporated easily in log-linear models analysis in the GLM framework. A cell (i,j) is excluded from the model, by the inclusion in the log-linear model (5.25) of the additional parameters q_{ij} that is responsible for fixing it to its observed frequency ($e_{ij} = 0$). In practice, this is achieved in standard software by adding in the log-linear model the index variable of (5.25) as an explanatory variable. In the presence of more structural zeros, additional index variables are added in the model, one for each structural zero. Alternatively, in the GLM context, all structural zeros can be indicated in one single variable that will be used to determine the subset of cells on which model (5.24) will be applied. SPSS handles structural zeros in the "general log-linear analysis" straightforward. An index variable has to be added in the data file, taking values 0 for structural zero cells and 1 otherwise. This index variable has to be declared in the "Cell Structure" field of the window:

Analyze > Loglinear > General...

QI will be illustrated in R, using Example 5.1 below.

When interaction is significant, model (4.5) is expressed for two-way incomplete tables as

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} , \qquad (i,j) \in S . \tag{5.26}$$

The main effect parameters satisfy constraints (4.4) while the sum to zero constraints in (4.6) for the interaction parameters are corrected to

$$\sum_{i=1}^{I} \mathrm{I}_{ij}^{S^c} \lambda_{ij}^{XY} = \sum_{j=1}^{J} \mathrm{I}_{ij}^{S^c} \lambda_{ij}^{XY} = 0 .$$

Log-linear models for multi-way incomplete contingency tables can be defined and fitted in an analogous manner.

5.5.1 Example 5.1

A typical example of contingency table with structural zeros is a survey on teenagers' health concerns. Teenagers are cross-classified according to their health concerns (in four categories), gender, and age (in two categories: 12–15, 16–17) in a $4 \times 2 \times 2$ table. The table has two structural zeros, since the health concerns category "menstrual problems" cannot refer to boys. This example is analyzed by Grizzle and Williams (1972) and Fienberg (2007, pp. 148–150). Ignoring age, i.e., merging over the age, the data are provided in Table 5.3, and there exists 1 structural zero; thus, the test of QI will be based on $2\,df$. QI is rejected, since $G^2(\mathrm{QI}) = 12.60$ (p-value = 0.0018) and $X^2(\mathrm{QI}) = 12.39$ (p-value = 0.0020). The ML estimates of the expected under QI cell frequencies along with the standardized residuals are provided in Table 5.3 in parentheses. Observing them, we conclude that the greatest difference between genders lies on the category "how healthy I am," for which girls are significantly less concerned and boys more than under independence, followed by "sex, reproduction" for which boys are significantly less interested while girls more, though not as significant. Finally, boys are more health concerns-free than expected under independence and girls less, but these differences are at the limit of 5 % significance.

This model was fitted in R by the function `fit.QI()`, provided in web appendix (see Sect. A.3.4). This function fits the QI model by (5.24), excluding the structural zero cells from the analysis. It needs to read the numbers of rows I and columns J of the table, the cell frequencies in a vector (by rows) of length IJ, where 0 are put in places of structural zeros, and an index vector of length IJ with entries the $\mathrm{I}_{ij}^{S^c}$ indices, given by rows. Thus for our example, the analysis is carried out by the commands

```
> freq<-c(6,16,0,12,49,29,77,102)
> zer<- c(0,0,1,0,0,0,0,0)
> fit.QI(freq,zer,4,2)
```

Table 5.3 Teenagers' cross-classification by gender and their health concerns (Brunswick 1971)

	Gender	
Health concerns	Male	Female
Sex, reproduction	6 (10.41, −2.13)	16 (11.59, 1.85)
Menstrual problems	–	12 (12.00, 0.00)
How healthy I am	49 (36.90, 3.08)	29 (41.10, −3.41)
Nothing	77 (84.69, −1.95)	102 (94.31, 1.90)

In parenthesis are provided the ML estimates under the QI model and the standardized residuals

The output of `fit.QI()`, beyond the results presented above, includes the overview of the fit provided by `summary()` and the estimates of the log-linear model parameters in vector forms for possible further use.

Alternatively, without restricting the cells on which the model applies, the QI model can be fitted by (5.25), including s extra parameters in the model, one for each structural zero. For this example, $s = 1$ and would have

```
> NI <- 4
> NJ <- 2
> row<-gl(NI,NJ,length=NI*NJ)
> col<-gl(NJ,1,length=NI*NJ)
> example <- data.frame(row, col, freq, zer)
> QI.model <- glm(freq ~ row+col+zer, poisson)
```

Under this approach, in the presence of $s > 1$ structural zeros, the index vector `zer` used in `glm()` above, needs to be replaced by a *factor* of $s + 1$ levels. Level 0 is assigned to the non-structural zero cells and a different level (from 1 to s) is assigned to every structural zero cell.

In case of existence of sampling zeros as well, they will not differ from the structural zeros in the frequency vector but in their index vector entry.

5.6 Models for Joint and Marginal Distributions

Model (5.6) applies directly on the cell entries of the table. In certain frameworks, it is of interest to model or test hypotheses about linear functions of the cell entries. For this, (5.6) is extended to

$$\log(\mathbf{Mm}) = \mathbf{X}\boldsymbol{\beta} \ , \tag{5.27}$$

with $\mathbf{M}$ a matrix suitably defined in order to form the desired functions of the expected cell entries when applied on $\mathbf{m}$.

The most famous models of this type are those modeling the marginals of a table, since some structures can easier be expressed in terms of marginal distributions, leading to the *marginal models*. Marginal models for contingency tables impose

structural restrictions on certain marginals of the classification variables and are usually of log-linear type. A characteristic example is the *marginal homogeneity* model for a square $I \times I$ table, presented in Sect. 9.2.2. For higher dimensional tables, modeling the marginal distributions is important for clustered and longitudinal categorical data (see Sects. 5.7.2 and 9.7.4).

However, if we would like to model the local odds ratios of an $I \times J$ table, model (5.27) is not appropriate; a further extension is needed. A brighter family of models is the generalized log-linear model (GLLM)

$$\mathbf{C}\log(\mathbf{Mm}) = \mathbf{X}\boldsymbol{\beta} \ . \tag{5.28}$$

Matrix $\mathbf{C}$ provides more flexibility and allows an even brighter variety of models to be included in this class. GLLM is introduced by Lang and Agresti (1994) and opened new origins in the analysis of multivariate categorical data, providing a powerful and flexible framework to model structures of associations. Model (5.28) is suitable for modeling, among others, the log of local or global odds ratios (see Sect. 2.2.5). Recall the matrix definition of the generalized odds ratios, given by (2.54) and (2.55), which correspond to the left-hand side of (5.28).

GLLM is itself a member of the broader *multinomial-Poisson homogeneous* (MPH) model, which is of the very general form

$$\mathrm{L}(\mathbf{m}) = \mathbf{X}\boldsymbol{\beta} \ , \tag{5.29}$$

where L is a link function. Details on inference for the MPH model are beyond the scope of this book and can be found in Lang (2004, 2005). Setting $\mathrm{L}(\mathbf{m}) = \mathbf{C}\log(\mathbf{Mm})$, (5.29) reduces to (5.28).

Another special case of the MPH model (5.29) is the

$$\mathrm{h}(\mathbf{m}) = \mathbf{0} \ , \tag{5.30}$$

where h() is a smooth constraint function with the constraints in (5.30) being nonredundant. With the adequate choice of the constraint function h(), model (5.30) reduces to the independence model (2.52), expressed in terms of the local odds ratios.

Though inference for the MPH model is not straightforward, it can be implemented in `R` by the `mph` function of Lang or the package `hmmm` of Colombi et al. (2013). We will illustrate `mph`, which is a powerful and flexible function that fits a big variety of general models via maximum likelihood. We limit its use only to GLLM models (5.28) and to model (5.30), both considered for the local odds ratios and the global odds ratios of a contingency table.

Function `mph` is available on request. The file "mph.Rcode.txt" is then sent and the routine `mph` is activated in `R` by

```
> source("c://...//mph.Rcode.txt")
```

The data are read in vector form that has to be defined as matrix. Thus, the $I \times J$ table of observed frequencies is expanded (by rows) in a $IJ \times 1$ vector `freq` and this vector finally forms the $IJ \times 1$ data matrix

```
> y <- matrix(freq)
```

The derived vector of expected cell frequencies **m** is also a matrix of size $IJ \times 1$.

The typical expression of the `mph` function for fitting (5.29) is

```
> mph.out <- mph.fit(y=y,L.fct=L.fct,X=X, strata=1)
```

where `L.fct` is the link function and `X` the design matrix of the MPH model (5.29) under consideration. The link for the GLLM model (5.28) is defined by

```
> L.fct <- function(m)C%*%log(M%*%m)
```

with `C` and `M` appropriate defined matrices. In the sequel, command

```
> mph.summary(mph.out,cell.stats=T,model.info=T)
```

produces summary output of the model, i.e., goodness-of-fit statistics, parameter estimates, expected cell frequency estimates under the assumed model, and information on the model applied and its convergence.

Model (5.30) is fitted by

```
mph.constr <- mph.fit(y, constraint=h.fct, strata=1)
```

where `h.fct` is the constraints function. For example, in order to fit the independence model (2.52), it should be

```
> h.fct <- function(m) {C%*%log(m)}
```

with `C` an appropriate $(I-1)(J-1) \times IJ$ matrix.

Examples of fitting the GLLM model through the `L.fct` option will be discussed in Sects. 6.6.4 and 7.1, for the local and the global odds ratios, respectively. The standard expression of `mph.fit()` assumes one single multinomial sample (`strata=1`). The extra option for defining more strata of data will be discussed in Sect. 5.6.2. At this point we will use `mph` to fit model (2.52) for our familiar Example 2.3, illustrating the use of `h.fct`.

5.6.1 *Example 2.4 by* `mph`

The function `local.odds.DM()` in the web appendix (see Sect. A.3.2) produces the matrix `C` needed to derive the logs of the local odds ratios when multiplied to `log(m)`, for tables of any size $I \times J$.

Hence, after actualizing `mph` in `R`, model (2.52) is fitted for our example by

```
> NI <- 3;  NJ <- 5
> freq <- c(45,116,19,48,23,40,167,33,68,41,47,185,34,63,26)
> C<-local.odds.DM(NI,NJ)
> h.fct <- function(m) {C%*%log(m)}
> ind.odds <- mph.fit(y, constraint=h.fct, strata=1)
```

The corresponding output is derived by

```
> mph.summary(ind.odds,cell.stats=T,model.info=T)
```

Part of this output is provided in Table 5.4.

Table 5.4 Output of the `mph` function, fitting the independence model on the local odds ratios of Example 2.4

```
MODEL GOODNESS OF FIT: Test of Ho: h(p)=0 vs. Ha: not Ho...

Likelihood Ratio Stat (df=8):  Gsq=10.36287 (pval=0.2405)
Pearson's  Score Stat (df=8):  Xsq=10.52048 (pval=0.2304)
Generalized Wald Stat (df=8):  Wsq=10.40275 (pval=0.2379)

Adj Resids: -1.709 -1.604 ...1.865 2.195,
Number |Adj Resid| > 2: 1

SAMPLING PLAN INFORMATION...
Number of strata: 1
Strata identifiers: 1
Strata with fixed sample sizes: all
Observed strata sample sizes: 955
CELL-SPECIFIC STATISTICS...
    strata OBS       FV StdErr.FV  PROB  StdErr.PROB ADJ.RESIDS
y1     1    45  34.6932   3.3753  0.0363    0.0035     2.1954
y2     1   116 123.0031   7.8052  0.1288    0.0082    -1.0299
y3     1    19  22.6031   2.6280  0.0237    0.0028    -0.9253
y4     1    48  47.0461   4.0679  0.0493    0.0043     0.1797
y5     1    23  23.6545   2.6971  0.0248    0.0028    -0.1647
y6     1    40  48.2387   4.4071  0.0505    0.0046    -1.6041
y7     1   167 171.0283   9.2226  0.1791    0.0097    -0.5415
y8     1    33  31.4283   3.4996  0.0329    0.0037     0.3690
y9     1    68  65.4147   5.2158  0.0685    0.0055     0.4452
y10    1    41  32.8901   3.5852  0.0344    0.0038     1.8653
y11    1    47  49.0681   4.4699  0.0514    0.0047    -0.4012
y12    1   185 173.9686   9.3027  0.1822    0.0097     1.4776
y13    1    34  31.9686   3.5528  0.0335    0.0037     0.4752
y14    1    63  66.5393   5.2853  0.0697    0.0055    -0.6073
y15    1    26  33.4555   3.6394  0.0350    0.0038    -1.7087
CONVERGENCE INFORMATION...
    Original counts used.
    iterations = 5 , time elapsed = 0.18
    norm.diff = 1.80924e-09 = dist between last and second
    last iterates.
           Norm diff convergence criterion [1e-06] was met.
    norm.score = 1.61128e-09 = norm of score at last iteration.
           Norm score convergence criterion [1e-06] was met.
```

If we wanted to express the independence model in terms of the global odds ratios, then $\mathrm{h}(\mathbf{m})$ in (5.30) equals $\mathrm{h}(\mathbf{m}) = \mathbf{C}\log(\mathbf{M}\mathbf{m})$, with matrices $\mathbf{C}$ and $\mathbf{M}$ appropriately defined. Function `global.odds.DM()` of the web appendix (see Sect. A.3.2) returns these two matrices for tables of size $I \times J$. The procedure above had to be adjusted as follows:

```
> C <- global.odds.DM(NI,NJ)$C;  M <- global.odds.DM(NI,NJ)$M
> h.fct <- function(m) {C%*%log(M%*%m)}
> ind.glob <- mph.fit(y, constraint=h.fct, strata=1)
```

5.6.2 *Example 3.3 by* `mph`

The hypothesis of homogeneous association (3.7) in $2 \times 2 \times K$ tables can be treated also in the GLLM framework, expressed by (5.28) with $\mathbf{m}$ the expected cell frequencies under the homogeneous association hypothesis expanded in a $4K \times 1$ matrix form, $\mathbf{X} = (1)_{K\times 1}$, and $\mathbf{C}$ the $K \times 4K$ block-diagonal matrix $\mathbf{C} = diag(\mathbf{C}_1, \ldots, \mathbf{C}_K)$ matrix with $\mathbf{C}_k = \mathbf{C}_0 = (1, -1, -1, 1)$, for $k = 1, \ldots, K$. $\mathbf{C}_0$ is the matrix for constructing the log odds ratios when applied on $\mathbf{m}$. It has this form, provided that the expected frequency table is expanded by columns. In this case the parameter is scalar and is equal to the assumed log odds ratio for all partial 2×2 tables under the homogeneous association hypothesis, i.e., $\beta = \log \theta$.

This approach is illustrated in `mph` for Example 3.3, as follows. Function `bdiag()` of library `Matrix` is applied to produce the block-diagonal matrix $\mathbf{C}$.

```
> source("c://Program Files//R//mph.Rcode.txt");
freq <- c(79,68,5,17,89,221,4,46,141,77,6,18,45,26,29,21,81,112,
          3,11,168,51,13,12);
y<- matrix(freq);  K <- 6;  X1 <- matrix(rep(1, K));
library(Matrix);  C0<-c(1, -1, -1, 1);
C <- t(bdiag(C0,C0,C0,C0,C0,C0)); # 6×6 block-diagonal matrix
L.fct <- function(m){C%*%log(m)};
mph.out <- mph.fit(y=y,strata=K,L.fct=L.fct,X=X1);
mph.summary(mph.out,cell.stats=T,model.info=T)
```

From the observed output we have that $G^2 = 7.950$ (p-value=0.159, df=5) and $X^2 = 7.896$ (p-value=0.162, df=5) while the ML estimate of the common under homogeneous association log odds ratio is $\hat{\beta} = 1.0759$, i.e., $\hat{\theta} = 2.9326$. This model is equivalent to the homogeneous association log-linear model applied on the cell frequencies (see Sect. 4.6.1.1). Recall from Sect. 3.3.4 that the Mantel–Haenszel estimate was $\hat{\theta}_{MH} = 2.96$.

5.7 Overview and Further Reading

The classical reference for GLMs is McCullagh and Nelder (1989). Additionally, a comprehensive reference is Fahrmeir and Tutz (2001). For application of GLMs in S-Plus and `R`, we refer to Venables and Ripley (2002, Chap. 7). Dobson and Barnett (2008) provide an easy to follow introduction to GLMs, with theoretical counterpart

but focusing on the analysis of particular types of data and their implementation in standard software, categorical data included. They consider also Bayesian analysis and Markov chain Monte Carlo (MCMC) methods to fit GLMs. A formulation and presentation of models for categorical data through the GLM family can be found in Agresti (2007, 2013).

GLMs have been extended in various directions, like for incorporating nonconstant variance, modeling dispersion, or generalizing the link function (McCullagh and Nelder 1989). In categorical data context, characteristic cases are, for example, the consideration of a negative binomial instead of a Poisson response or the introduction of dispersion effect in the cumulative link model (McCullagh 1980).

The Fisher information matrix plays an important role in statistics in many different aspects, the two most characteristic being in determining the variance of an estimator and the "noninformative" priors determination in the Bayesian setup. Spall (2005) reviews basic principles associated with the information matrix and presents a resampling-based method for computing the information matrix.

When the n_i's are small, the residuals are not approximately normal distributed. For such cases the transformed *Anscombe residuals* have been proposed (see McCullagh and Nelder 1989). For a survey on residuals for GLMs, we refer to Pierce and Schafer (1986). For goodness-of-fit testing of GLMs for sparse data, see Farrington (1996).

5.7.1 Incomplete Contingency Tables

Incomplete tables attracted researchers' attention very early. Stigler (1992), in an interesting and enlightening historical review, points out that in 1913, Karl Pearson was the first to consider the independence model for two-way incomplete tables. The historical fingerprint data set in Waite (1915) contains structural and sampling zeros while Harris and Treloar (1927) and Harris and Tu (1929) face for incomplete tables the problems occurring in the applicability of the contingency coefficient.

The existence of ML estimates for models considered on incomplete tables became a central issue in the late 1960s and 1970s. The most well-known model for incomplete tables is the QI model, presented in Sect. 5.5. Very popular, especially in the context of rater agreement and mobility tables, is the QI model for square tables having the main-diagonal entries missing or excluded. The key reference for the QI model is Goodman (1968), though the QI model for diagonal truncated square tables had been considered earlier by Savage and Deutsch (1960) and Goodman (1963a) in transaction flows analysis and White (1963) and Goodman (1965) in mobility table analysis. Fundamental papers in developing inference for QI in the log-linear model setup were Bishop and Fienberg (1969), Fienberg (1970a), and Haberman (1973a), with the last two providing conditions for existence of unique nonzero ML estimates. The QI model is discussed in detail in Bishop et al. (1975).

Interesting is the approach of Fienberg (1969) that locates the cells exhibiting interaction, when the number of such cells is relatively small compared with the total number of cells in the table, and applies then the QI model, excluding these cells. Mantel (1970) focused on determining the appropriate degrees of freedom and considered, beyond independence, also symmetry testing for incomplete square tables. Goodman (1971a) proposed a test procedure for testing the hypothesis of QI simultaneously for several different subsets of the cells of a table. Enke (1977) considered incomplete two-way tables of special structures that are decomposed to separable tables and lead to closed form MLEs. For the ML estimation of the diagonal truncated independence model, Morgan and Titterington (1977) compared the performance of the EM, Newton–Raphson, and iterative scaling algorithms, concluding empirically that the last is the least efficient method.

Another special type of incomplete square tables are the triangular tables. Such form of incomplete tables occurred already in Waite (1915), while is special referred in Goodman (1968) and Bishop and Fienberg (1969). Special on triangular QI are Goodman (1979a, 1994) and Altham (1975), who considered also the Bayesian analysis with conjugate prior. For ordinal triangular tables, Sarkar (1989) interpreted QI in terms of likelihood ratio dependence and Tsai and Sen (1995) provided an alternative test of QI. We considered in Sect. 5.5 the problem of incorporating structural zeros in the simple independence model for two-way tables. The diagonal and triangular truncated tables will be presented in Sect. 9.3.

Colombo and Ihm (1988) applied the QI model in an unusual context to estimate failure rates of components classified by two qualitative covariates. QI allows for different operating times in the various cells, zero operating time included.

Incomplete tables may occur in tables of higher dimension and of more complex association structures. Klimova et al. (2012) introduce a general family of models for contingency tables, the rational models, which provide a unified framework for analysis of complete and incomplete tables by log-linear models and others, like association models (Chap. 6) and rater agreement models (Sect. 9.5.2). They provide sufficient conditions for the existence of the ML estimates under this general model and prove the classical equivalence between the Poisson and multinomial likelihoods.

A nice review of the literature on the sensitivity analysis of overparameterized models for incomplete categorical data, Bayesian and frequentist, is provided by Poleto et al. (2011).

5.7.2 Marginal Distributions Modeling

Marginal models have been mainly developed by Lang and Agresti (1994), Lang (1996a), Lang et al. (1999), and Bergsma and Rudas (2002a,b). Their approach is based on earlier work by Haber (1985) and Haber and Brown (1986). Bartolucci et al. (2007) generalized the model of Bergsma and Rudas (2002a) to allow for global and continuation type logits, which may be more adequate for ordinal

data analysis. Rudas et al. (2010) formed conditional independence models in a marginal log-linear parameterization. Becker et al. (1998) explored similarities and differences between standard log-linear and marginal models with special emphasis on square tables and reference to multi-way tables as well in the social sciences framework. For a detailed presentation of marginal models and their features, we refer to the book by Bergsma et al. (2009).

Marginal models are applied for modeling repeated (or clustered) categorical data (see also Sect. 9.7.4).

Chapter 6
Association Models

Abstract The association models, appropriate for the analysis of ordinal contingency tables, are presented for two-way and multi-way contingency tables. Their features, properties, and the associated graphs are discussed. The models of uniform association (U), row effect (R), column effect (C), multiplicative row–column effect (RC), and the more general RC(M) model are illustrated with examples in terms of fit, presentation, and interpretation. They are all worked out in `R`, through functions provided for their fit and the construction of their scores' plots.

Keywords Association models: U, R, C, RC(M) • Graphs for the RC(2) model • Association models for multi-way tables

6.1 Basic Association Models for Two-way Tables

We realized so far that in the context of classical log-linear models there are just two options for modeling two-way contingency tables: the parsimonious but restrictive model of independence (4.1) and the saturated. Association models fill the gap between these two extreme cases by imposing a special structure on the association and reducing the number of interaction parameters, providing thus intermediate models of dependence. For ease in understanding but also for interpretation purposes, it is convenient to think in terms of local associations in the table and first define the models on local odds ratios rather than cell frequencies. Recall that for models applied on an $I \times J$ contingency table there always exists an equivalent expression defining them on the $(I-1) \times (J-1)$ table of the corresponding set of local odds ratios.

In most of the cases association models apply to ordinal classification variables and are thus usually introduced as *models for ordinal data*. However, some of them do not require ordinality, as we shall see later on in this chapter.

M. Kateri, *Contingency Table Analysis: Methods and Implementation Using R*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4811-4_6,

6.1.1 Linear-by-Linear Association Model

We have seen in Sect. 2.2.5 that for an $I \times J$ contingency table and under the model of independence, all the local odds ratios are equal to 1, i.e., $\theta_{ij}^L = 1$, $i = 1 \ldots, I-1$; $j = 1 \ldots, J-1$. Whenever the model of independence is of poor fit, the only alternative in the framework of classical log-linear models is the saturated model (4.5), which assumes all θ_{ij}^L's to be free parameters and is noneffective in summarizing the underlying significant association. A natural way to proceed is to assume a pattern for this underlying association. This way, the number of parameters to be estimated is reduced and, most important, we can provide a meaningful interpretation. The easiest pattern to think of, which is meanwhile of clear and strong interpretational power, is that of constant θ_{ij}^L's, as under independence, but different than 1. That is, to introduce the model

$$\theta_{ij}^L = c, \qquad i = 1, \ldots, I-1 \ , \ \ j = 1, \ldots, J-1, \tag{6.1}$$

for some $c > 0$, to be estimated. This model allows for interaction while remains parsimonious, since it has just one parameter more than the independence model, the parameter c. Under the independence model, all possible odds ratios $\theta_{ij}^{k\ell}$ of the table are equal to 1. Under (6.1), local association is uniform, since all the local odds ratios are equal to c. This property characterizes model (6.1), which is therefore called *uniform* association model, denoted as U. When it comes to the odds ratio $\theta_{ij}^{k\ell}$ of any 2×2 subtable of our initial table, through (2.46), (6.1) takes the form

$$\theta_{ij}^{k\ell} = c^{(k-i)(\ell-j)}, \ i = 1, \ldots, I-1 \ , \ j = 1, \ldots, J-1 \ , \ i < k \leq I \ , \ j < \ell \leq J \ ,$$

and in log-scale

$$\log \theta_{ij}^{k\ell} = (k-i)(\ell-j) \log c, \quad i = 1, \ldots, I-1 \ , \ j = 1, \ldots, J-1 \ , \tag{6.2}$$

for $i < k \leq I$, $j < \ell \leq J$. Under the U model, the general $\theta_{ij}^{k\ell}$ odds ratio is influenced by the categories of each classification variable but only through their distances. Hence, odds ratios formed by cells further apart will exhibit stronger association. Measuring thus how far apart are two categories of a classification variable is crucial. Distances between categories are meaningful only when the corresponding classification variable is ordinal. Hence the U model makes sense to be considered only for tables with both classification variables ordinal or with one ordinal and the other binary.

The U model assumes that all successive categories of a classification variable are equidistant. However, there can arise ordinal variables of non-equidistant successive categories. A typical example of this type is a categorized income variable, which is actually interval scaled with categories corresponding to intervals of unequal length. A flexible way to handle such situations is to assign scores to the categories of the classification variables and express their distances by the corresponding scores'

differences. Thus, let $\{\alpha_1, \alpha_2, \ldots, \alpha_I\}$ and $\{\nu_1, \nu_2, \ldots, \nu_J\}$ be the scores assigned to the row and column categories, respectively. The simplest and most natural choice for the scores is $\alpha_i = i$ $(i = 1, \ldots, I)$ and $\nu_j = j$ $(j = 1, \ldots, J)$, which corresponds to model (6.2). Allowing the scores to take other values as well and setting $\varphi = \log c$, we are led to model

$$\log \theta_{ij}^{k\ell} = \varphi(\alpha_k - \alpha_i)(\nu_\ell - \nu_j), \quad i = 1, \ldots, I-1\ ,\ j = 1, \ldots, J-1\ , \tag{6.3}$$

with $i < k \leq I\ ,\ j < \ell \leq J$, for which scores of successive row or column categories are not necessarily equidistant. Their distance is meant in terms of their similarity as they interact with the other classification variable. Thus different scores may be assigned to the same levels of a classification variable X when interacting with different variables Y or Z. This will be illustrated in a three-way contingency table example in Sect. 6.7.1. Regarding the scores' assignment, refer also to the related discussion in Sect. 2.3.1.

For non-equidistant scores for successive categories, the local odds ratios under (6.3) are no more all equal but proportional (in log-scale) to the distance between the enrolled categories of each classification variable. Due to this linear dependence on each of the classification variables, model (6.3) is called the *linear-by-linear* association model (LL).

Though the interpretation of these models is clear and natural when formulated in terms of the local odds ratios, the development of inferential aspects and model fitting is more straightforward for their equivalent formulation in terms of expected cell frequencies. Recalling that the saturated log-linear model in terms of θ_{ij} is provided by (4.7) and equating (4.7) to (6.3) for $k = i+1$ and $\ell = j+1$, we conclude that the (i, j)th interaction term under the LL model has the form $\lambda_{ij}^{XY} = \varphi \alpha_i \nu_j$. Hence, the equivalent expression of LL model (6.3) in terms of expected cell frequencies is

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \varphi \alpha_i \nu_j, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J, \tag{6.4}$$

where the overall mean and the main effects parameters are those of the classical log-linear model.

Model (6.4) reduces to the U model and is thus equivalent to (6.2), not just for $\alpha_i = i$, $(i = 1, \ldots, I)$ and $\nu_j = j$, $(j = 1, \ldots, J)$ but for any choice of row and column scores $\{\alpha_1, \alpha_2, \ldots, \alpha_I\}$ and $\{\nu_1, \nu_2, \ldots, \nu_J\}$, as long as they are both equidistant for successive categories. This is due to model's LL property of being invariant in linear transformation of the scores, as we shall see in Sect. 6.4. Thus, for identifiability purposes, usually the scores are set to satisfy the sum-to-zero and the sum of squares-to-one constraints

$$\sum_{i=1}^{I} \alpha_i = 0 \qquad \text{and} \qquad \sum_{i=1}^{I} \alpha_i^2 = 1, \tag{6.5}$$

$$\sum_{j=1}^{J} \nu_j = 0 \qquad \text{and} \qquad \sum_{j=1}^{J} \nu_j^2 = 1. \tag{6.6}$$

For scores satisfying the (6.5) and (6.6) constraints, multiplying (6.4) by $\propto_i \nu_j$ and adding over i and j leads to

$$\varphi = \sum_{i,j} \propto_i \nu_j \log m_{ij} \,, \tag{6.7}$$

i.e., φ measures the correlation between row and column scores, fact that justifies its characterization as *intrinsic association* parameter.

6.1.2 Example 6.1

We shall demonstrate the utility and interpretation power of this parsimonious association model with just one degree of freedom less than complete independence by an example. We shall first focus on explaining the nature and use of such a model and we will provide inferential details and application in software at a later stage. The data used are from a survey on the use of cannabis among students, conducted at the University of Ioannina (Greece) in 1995 and published in Marselos et al. (1997). The students' frequency of alcohol consumption is measured on a four-level scale ranging from at most once per month up to more frequent than twice per week while their trial of cannabis through a three-level variable (never tried–tried once or twice–more often). These two ordinal variables are cross-classified leading to a 4×3 table provided in Table 6.1.

These data provide strong evidence against the independence model (4.1), since the corresponding LR test statistic is $G^2(I) = 152.793$, which is highly significant with an asymptotic p-value< 0.00005 ($df = 6$). In the context of classical log-linear models the only alternative is to add the interaction term λ_{ij}^{XY} in the model and end up thus to the saturated model.

Taking advantage of the ordinal nature of the classification variables, we apply the U model to the data of Table 6.1, by fitting model (6.4) with $\propto_i = i$ ($i = 1, \ldots, 4$) and $\nu_j = j$ ($j = 1, 2, 3$). Thus, we introduce just one additional parameter to the independence model, the φ. The LR test statistic for model (6.4) equals $G^2(U) = 1.469$, leading to a reduction of 151.324 from $G^2(I)$ by sacrificing just 1 df. This model is of impressive fit with p-value$= 0.92$. The cell estimates under U are provided in parentheses in Table 6.1.

As already mentioned, under the U model, the local odds ratios θ_{ij}^L are constant all over the table. The corresponding sampling values for the local odds ratios are provided in Table 6.2. In this case, the association parameter φ is estimated as $\hat{\varphi} = 0.803$ and furthermore $\hat{\theta}_{ij}^L = \hat{\theta} = \exp(\hat{\varphi}) = \exp(0.803) = 2.23$, for all $i = 1, 2, 3$ and $j = 1, 2$. This means that the odds of having tried cannabis once or twice vs. never tried is 2.23 times higher for students who drink twice a month than those who drink at most once a month. The same comparison holds for any odds ratio comparing successive row and successive column categories.

Table 6.1 Students' survey about cannabis use at the University of Ioannina, Greece (1995)

	I tried cannabis...			
Alcohol consumption	Never	Once or twice	More often	Total
At most once/month	204 (204.4)	6 (5.7)	1 (0.9)	211
Twice/month	211 (211.4)	13 (13.1)	5 (4.5)	229
Twice/week	357 (352.8)	44 (48.8)	38 (37.4)	439
More often	92 (95.3)	34 (29.4)	49 (50.3)	175
Total	864	97	93	1054

In parentheses are given the maximum likelihood estimates under the model of uniform association (U)

Table 6.2 Sample local odds ratios for the students' survey about cannabis use at the University of Ioannina, Greece (1995)

	I tried cannabis...		
Alcohol consumption	Never	Once or twice	More often
At most once/month	2.10	2.31	
Twice/month	2.00	2.25	
Twice/week	3.00	1.67	
More often			

If we would like to compare any non-successive categories, the results can be adjusted accordingly. For example, for the odds ratio formed by the corner ("extreme") cells of the table, it holds

$$\frac{\hat{\pi}_{11}\hat{\pi}_{43}}{\hat{\pi}_{13}\hat{\pi}_{41}} = \exp\left(\hat{\varphi}(\alpha_4 - \alpha_1)(\nu_3 - \nu_1)\right) = \exp\left(\hat{\varphi} \cdot 3 \cdot 2\right) = 123.387\,,$$

meaning that the odds of using often cannabis instead of never tried is 123 times higher for student who drink more often than twice a week than for students who drink at most once a month.

6.1.3 Row and Column Effect Models

The LL model presented above is a very parsimonious and useful model of strong interpretation power when it is applicable. However, often it is proved insufficient. It can be the case that the structure of model (6.4) is appropriate but there is no obvious way of deciding about the scores of one of the classification variables. It is then natural to broaden model (6.4) to a class of more flexible association models by relaxing the assumptions about known scores. Model (6.4) with unknown row scores $\{\alpha_1, \ldots, \alpha_I\}$ and thus parameters to be estimated is the *row effect* association model, to be denoted as R. Under this model, the odds defined over the column classification variable vary from row to row, i.e., the effect of the row classification

variable on the column odds is significant but unknown. This effect is reflected in the row scores and more precisely in the unknown (and unequal) distances between successive row categories. Model R has $I-2$ additional parameters than model LL, corresponding to the row scores. The number of parameters is reduced by two, due to the identifiability constraints (6.5) that hold. Thus, the associated df of model R equal $(I-1)(J-2)$. Analogously, the *column effect* association model C is defined by expression (6.4) for known row scores and unknown column scores $\{\nu_1,\ldots,\nu_J\}$. It models the effect of the column classification variable on the row odds. The associated df are $df(\mathrm{C}) = (I-2)(J-1)$.

We have seen in the context of the U association model that its definition in terms of odds ratios (6.1) is more natural with respect to interpretation. The R model is equivalently defined in terms of local odds ratios as

$$\theta_{ij}^{L} = c_{1i}\ ,\ \ i=1,\ldots,I-1\ ,\ j=1,\ldots,J-1, \tag{6.8}$$

and the C model as

$$\theta_{ij}^{L} = c_{2j}\ ,\ \ i=1,\ldots,I-1\ ,\ j=1,\ldots,J-1. \tag{6.9}$$

Expression (6.8) reveals the dependence of the column odds on the row category while the analogue statement is true for the C model (6.9).

In terms of local odds ratios and categories' scores, model R is expressed as

$$\log\left(\theta_{ij}^{L}\right) = \varphi(\propto_{i+1}-\propto_i)(\nu_{j+1}-\nu_j),\ \ i=1,\ldots,I-1\ ,\ j=1,\ldots,J-1, \tag{6.10}$$

with parametric row scores $\{\propto_i,\ i=1,\ldots,I\}$ and known (equidistant) column scores $\{\nu_j,\ \ j=1,\ldots,J\}$. Analogously, model C is (6.10) with parametric column scores and known (equidistant) row scores.

6.1.4 Row by Column Effect Model

The LL, U, R, and C models considered so far are special types of log-linear models. The LL model is applicable on two-way tables when both the classification variables are ordinal. The R and C models are less restrictive about the nature of the underlying classification variables and thus also less parsimonious. They allow the row or column classification variable, respectively, to be ordinal but with unknown distances between the scores assigned to its successive categories or even nominal. This is achieved by considering the row or the column scores as unknown parameters to be estimated. Furthermore, a more flexible model can be defined by (6.4), considering the row and the column score vectors to be both unknown parameters. Thus, we model a multiplicative *row by column* association. This model, denoted by RC, is no more linear in its parameters and their estimation is not straightforward. The estimation problem will be faced in Sect. 6.2.

Table 6.3 Association models and related df. The U model is a special LL model

Association model	$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_I)$	$\boldsymbol{\nu} = (\nu_1, \ldots, \nu_J)$
Linear × linear (LL)	Known	Known
Row effect (R)	Unknown	Known
Column effect (C)	Known	Unknown
Multiplicative row–column (RC)	Unknown	Unknown

Model	Parameters additional to independence	d.f.
LL	1	$(I-1)(J-1)-1$
R	$1+(I-2)$	$(I-1)(J-2)$
C	$1+(J-2)$	$(I-2)(J-1)$
RC	$1+(I-2)+(J-2)$	$(I-2)(J-2)$

In terms of local odds ratios, the RC model is defined by

$$\theta_{ij}^{L} = c_{1i}c_{2j}\,, \quad i = 1,\ldots,I-1\,, \quad j = 1,\ldots,J-1\,,$$

allowing the effect of each classification variable on the odds defined by the other one to vary from category to category. In log-scale, the RC model is given by (6.10) with parametric (unknown) row and column scores. The RC model does not require ordinality for any of the classification variables. Thus it can be applied in tables of nominal variables as well. Of course, scores' assignment is more natural for ordinal variables.

The association models considered so far are all defined in terms of expected cell frequencies by

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \varphi\alpha_i\nu_j, \quad i = 1,\ldots,I, \quad j = 1,\ldots,J, \tag{6.11}$$

i.e., by expression (6.4). Thus, all association models considered so far are defined by the same expression (6.4) and differentiated by the assumptions made for the nature of the scores, known or unknown parameters. They are summarized in Table 6.3. The U model is a special LL model and is not listed in the table.

6.1.5 Example 6.1 (Revisited)

Revisiting the cannabis example, we fit in Table 6.1 the R, C, and RC models. The test statistic values along with their corresponding significance are provided in Table 6.4. The estimates of parametric scores as well as the values of the fixed scores for these models are provided in Table 6.5. The estimated score parameters for the rows and the columns are close to be equidistant for two successive score parameters. Thus it seems not to be worth to adopt a more complex model than U.

Table 6.4 LR goodness-of-fit tests for the independence and the association models applied in Table 6.1

Model	G^2	d.f.	p-value
I	152.7933	6	0.0000
U	1.4687	5	0.9167
C	1.1004	4	0.8942
R	1.2964	3	0.7230
RC	0.6044	2	0.7392

Table 6.5 ML estimates for parameters and fixed scores values for the U, R, C, and RC models applied in Table 6.1. Values in *italics* correspond to fixed scores

	U	R	C	RC
$\hat{\varphi}$	2.5382	2.4776	2.4638	2.3191
α_1	*−0.6708*	−0.6640	*−0.6708*	−0.6494
α_2	*−0.2236*	−0.2238	*−0.2236*	−0.2365
α_3	*0.2236*	0.2043	*0.2236*	0.1880
α_4	*0.6708*	0.6836	*0.6708*	0.6979
ν_1	*−0.7071*	*−0.7071*	−0.7331	−0.7447
ν_2	*0.0000*	*0.0000*	0.0553	0.0825
ν_3	*0.7071*	*0.7071*	0.6779	0.6622

This is verified also from the corresponding goodness-of-fit tests, where we see that moving from the simple U model to less parsimonious association models, the fit improvement is very minor. A more detailed discussion on association model selection will be carried out in Sect. 6.3.

Focusing on the C and RC models, the estimated local odds ratios under these models are provided in Table 6.6. Recall that under the U model, the common local odds ratios estimate is 2.23. We can verify that the expected local odds ratios under the C model are column dependent, i.e., the value is common in each column but differs from column to column while under the more general RC model they are row and column dependent, thus all different to each other. However, the estimated local odds ratios are not that different to justify the use of more complicated models than the simple U model, which was of impressive fit. Under the C model, the odds of having tried cannabis once or twice vs. never tried is 2.38 times higher for those who are one level higher in the alcohol consumption scale, no matter what this level is. The odds in the second column of Table 6.6 can be interpreted similarly. In this example we did not refer at all at model R since it is less parsimonious of C and of worse fit.

Table 6.6 Estimated local odds ratios under the RC model and under the C model (in parentheses) for the students' survey about cannabis use at the University of Ioannina, Greece (1995)

	I tried cannabis...		
Alcohol consumption	Never	Once or twice	More often
At most once/month	2.21 (2,38)	1,74 (1,99)	
Twice/month	2,26 (2,38)	1,77 (1,99)	
Twice/week	2,66 (2,38)	1,98 (1,99)	
More often			

6.2 Maximum Likelihood Estimation for Association Models

For any association model, the maximum likelihood estimation approach is that described in Sect. 4.2 for log-linear models. Thus, independent of the underlying sampling scheme, ML estimates of an association model's parameters and eventually of its expected cell frequencies m_{ij} are achieved by maximizing the *Poisson log-likelihood kernel* (4.13) with respect to the parameters of the model. Substituting m_{ij} by the association model expression and equating the partial derivative of (4.13) with respect to a parameter of the model to zero, one is led to the likelihood equation corresponding to this parameter.

The likelihood equations with respect to the main effect parameters of model (6.4) are the same as the corresponding of the two-way standard log-linear model, given in (4.14). For the more general RC model where both set of scores are parameters, the likelihood equations for the row scores $(\alpha_1, \ldots, \alpha_I)$ are derived as

$$\sum_j \nu_j(\hat{m}_{ij} - n_{ij}) = 0, \qquad i = 1, \ldots, I \tag{6.12}$$

while for the column scores $(\nu_1, \ldots, \nu_J)$ as

$$\sum_i \alpha_i(\hat{m}_{ij} - n_{ij}) = 0, \qquad j = 1, \ldots, J\,. \tag{6.13}$$

Finally, the likelihood equation corresponding to the intrinsic association parameter φ is

$$\sum_{i,j} \alpha_i \nu_j(\hat{m}_{ij} - n_{ij}) = 0\,. \tag{6.14}$$

For the rest of the association models defined by (6.11) by considering the row or column scores or both of them as fixed, the likelihood equations are derived from the above set by eliminating the equations corresponding to known scores. Thus, the likelihood equations for the U model are (4.14) and (6.14) while for the R model (4.14), (6.12), and (6.14). Analogously, the likelihood equations for model C

are (4.14), (6.13), and (6.14). Note that the likelihood equation (6.14) is redundant given (6.12) or (6.13). This means that parameter φ is redundant whenever at least one set of scores is parametric and can thus be eliminated. At this point it is worth mentioning that the RC model was introduced by Goodman (1979b, model (4.1b)) in terms of the non-redundant parameters, as

$$m_{ij} = \alpha_i \beta_j \exp(\propto_i \nu_j), \quad i = 1, \ldots, I; \quad j = 1, \ldots, J \,, \tag{6.15}$$

with unconstrained row and column scores. Introducing the intrinsic association parameter φ with the cost of imposing constraints (6.5) and (6.6) on the scores, Goodman (1979b, model (4.5b)) proposed the equivalent expression

$$m_{ij} = \alpha_i \beta_j \exp(\varphi \propto_i \nu_j), \quad i = 1, \ldots, I; \quad j = 1, \ldots, J \,. \tag{6.16}$$

This way the RC model and its scores are comparable to other standard models (Chap. 7). The multiplicative form (6.16) is also equivalent to the log-form (6.4).

The ML estimates of the parameters of any association model cannot be derived in closed-form expression and the corresponding likelihood equations have to be solved iteratively. The simplest iterative procedure for association models ML estimation is based on the *Newton's unidimensional method*. The updating equations (at the tth iteration) for the RC model parameters' estimation, based on expression (6.16), are

$$\begin{aligned}
\alpha_i^{(t)} &= \alpha_i^{(t-1)} \frac{n_{i+}}{\tilde{m}_{i+}} \,, && i = 1, \ldots, I, \\
\beta_j^{(t)} &= \beta_j^{(t-1)} \frac{n_{+j}}{\tilde{m}_{+j}} \,, && j = 1, \ldots, J, \\
\propto_i^{(t)} &= \propto_i^{(t-1)} + \frac{\sum_j \nu_j^{(t-1)} (n_{ij} - \tilde{m}_{ij})}{\tilde{\varphi}^{(t-1)} \sum_j \left(\nu_j^{(t-1)} \right)^2 \tilde{m}_{ij}} \,, && i = 1, \ldots, I, \\
\nu_j^{(t)} &= \nu_j^{(t-1)} + \frac{\sum_i \propto_i^{(t-1)} (n_{ij} - \tilde{m}_{ij})}{\tilde{\varphi}^{(t-1)} \sum_i \left(\propto_i^{(t-1)} \right)^2 \tilde{m}_{ij}} \,, && j = 1, \ldots, J, \\
\varphi^{(t)} &= \varphi^{(t-1)} + \frac{\sum_i \propto_i^{(t-1)} \nu_j^{(t-1)} (n_{ij} - \tilde{m}_{ij})}{\tilde{\varphi}^{(t-1)} \sum_{i,j} \left(\propto_i^{(t-1)} \nu_j^{(t-1)} \right)^2 \tilde{m}_{ij}} \,,
\end{aligned}$$

where $\tilde{m}_{ij}$ stands for the ML estimate of m_{ij}, recalculated at each step of the iterations (Goodman 1979b).

As in every iterative procedure the assignment of initial values to the parameters' estimates is crucial. In this setup, a reasonable choice for the main effects is

$$\alpha_i^{(0)} = \exp\left(\frac{\ell_{i+}}{J} - \frac{\bar{\ell}}{2}\right) \text{ and } \beta_j^{(0)} = \exp\left(\frac{\ell_{+j}}{I} - \frac{\bar{\ell}}{2}\right),$$

where $\ell_{ij} = \log(n_{ij})$ and $\bar{\ell} = \frac{\ell_{++}}{IJ}$. A natural choice for the initial estimates of the parametric scores is to consider them equidistant for successive categories, i.e., as if the U model was applied. In this case, starting by considering the scores equal to the corresponding category index and rescaling them linearly so that constraints (6.5) and (6.6) are satisfied, we conclude to

$$\propto_i^{(0)} = \sqrt{\frac{3}{I(I^2-1)}}(2i - I - 1) \text{ and } \nu_j^{(0)} = \sqrt{\frac{3}{J(J^2-1)}}(2j - J - 1).$$

A compatible then choice for $\varphi^{(0)}$ would be $\varphi^{(0)} = \sum_{i,j} \propto_i \nu_j \log n_{ij}$; see (6.7). The algorithm convergence is checked through the change in the log-likelihood value (4.13), calculated after each parameters' estimates updating circle.

The standard algorithms normally applied are the *Newton–Raphson's* or the *Fisher's scoring* algorithm (see Sect. 5.3.1). In this context, the parameters of the under consideration association model has to be written in a vector form. For example, for the U model the parameter vector is $\boldsymbol{\beta} = (\lambda, \lambda_1^X, \ldots, \lambda_{I-1}^X, \lambda_1^Y, \ldots, \lambda_{J-1}^Y, \varphi)$. The Newton's unidimensional method is simpler, since it does not require matrix inversion but for this with the drawback that it does not estimate the s.e. of the parameters.

Information on available software and special programs for estimation of association models based on each of these algorithms will be provided in Sect. 6.6.

6.3 Association Model Selection

We have already faced in the context of the cannabis example the problem of selecting the appropriate association model when more than one of them is of adequate fit. The problem of model selection in the framework of association models is connected to the analysis of association (ANOAS) in a contingency table and is based on the interconnection between the models. In particular, it holds

$$\text{I} \subset \text{U (or (LL)} \subset \text{R (or C)} \subset \text{RC}.$$

Indeed, the I model is the U (or LL) model with $\varphi = 0$, while the C model, for example, is the RC model for a specific choice for the row scores. This means that

$$G^2(\text{I}) > G^2(\text{U}) > G^2(\text{C}) > G^2(\text{RC}) \,,$$

for example, with the analogous results for the LL or the R model. The crucial question at this point is whether the reduction in G^2 value as we move to less parsimonious models is worth, justifying the loss in df and simplicity. The answer is provided through the conditional testing procedure (see Sects. 4.6 and 5.3.4). As soon as we detect the simplest association model $\mathscr{M}_1$ of adequate fit, we abandon it in favor of a more complicated $\mathscr{M}_2$ ($\mathscr{M}_1 \subset \mathscr{M}_2$) only if the reduction in G^2 is statistically significant. Hence, we proceed testing the fit of $\mathscr{M}_1$ conditional on the fact that $\mathscr{M}_2$ holds by (4.34).

Thus, for example, given that the U, R, or C models hold, one could propose the conditional tests of independence $G^2(\mathrm{I}|\mathrm{U})$, $G^2(\mathrm{I}|\mathrm{R})$, or $G^2(\mathrm{I}|\mathrm{C})$, being asymptotically distributed as $\mathscr{X}^2$ with df equal to 1, $I-1$, or $J-1$, respectively. These conditional tests of independence, given that model U, R, or C holds, are of greater asymptotic power, compared to the traditional unconditional test of independence (Gross 1981; Agresti 1983a). The tests I|U and I|LL are special mentioned since they are most powerful as 1 df tests. In this context it is important to note that the conditional test I|RC is not that straightforward since $G^2(\mathrm{I}|\mathrm{RC}) = G^2(\mathrm{I}) - G^2(\mathrm{RC})$ is *not* asymptotically $\mathscr{X}^2$ distributed with $df = df(\mathrm{I}) - df(\mathrm{RC})$ as probably expected. The asymptotic null distribution of $G^2(\mathrm{I}|\mathrm{RC})$ for testing independence is that of the largest eigenvalue from a Wishart distributed matrix (Haberman 1981). Gradually conditional testing from the RC to I, such as I|U, U|R, and R|RC, is possible and provides an analysis of association (ANOAS) table, throwing light on the underlying association structure of the table and analyzing deviance from independence in terms of source (overall, row, interaction) in a manner analogous to the ANOVA table (Goodman 1981a).

6.3.1 Model Selection for Example 6.1

We have already seen that for the cannabis data set all association models provide an acceptable fit. It seems natural to favor the C model over the R, due to parsimony and better fit. Thus, the choice lies between the U, C, and RC models. By the conditional testing procedure one has $G^2(\mathrm{C}|\mathrm{RC}) = G^2(\mathrm{C}) - G^2(\mathrm{RC}) = 0.496$, which is non-significant based on the $\mathscr{X}^2_2$ distribution (p-value=0.7804). Thus, there is no point in adopting the RC model since it does not provide a significant improvement of the fit over the C model. Further on, since $G^2(\mathrm{U}|\mathrm{C}) = G^2(\mathrm{U}) - G^2(\mathrm{C}) = 0.3683$ is again non-significant (p-value=0.5439, $df = 1$), the model that seems to be appropriate for this data set is the simple U model, with just 1 df less than the independence and a straightforward interpretation of constant local association all over the table. This sequence of conditional testing is summarized in the ANOAS table, provided for this example at the end of Sect. 6.6.1.

6.4 Features of Association Models

We have mentioned that the LL model (U as well) is invariant under linear transformations of its scores. Actually, this property holds for all association models considered so far. Let $\propto_i^* = a_1 \propto_i + b_1$ and $v_j^* = a_2 v_j + b_2$ be any choice of linear rescaling for the row and column scores, respectively. Then in terms of the local odds ratios and the new scores, the association model would be defined as

$$\log \theta_{ij}^{k\ell} = \varphi^*(\propto_k^* - \propto_i^*)(v_\ell^* - v_j^*)\ , \quad i = 1,\ldots,I-1\ , \quad j = 1,\ldots,J-1\ ,$$

for $i < k \leq I$ and $j < \ell \leq J$. This is further transformed to

$$\log \theta_{ij}^{k\ell} = a_1 a_2 \varphi^* (\alpha_k - \alpha_i)(\nu_\ell - \nu_j) \ ,$$

which for $\varphi^* = \frac{\varphi}{a_1 a_2}$ is equivalent to (6.3). Thus, without affecting the expected cell frequencies, their estimates, and consequently the fit of the model, we can replace the normalizing constraints on the scores by the *weighted normalizing constraints*:

$$\sum_i w_{1i} \alpha_i = \sum_j w_{2j} \nu_j = 0 \quad \text{and} \quad \sum_i w_{1i} \alpha_i^2 = \sum_j w_{2j} \nu_j^2 = 1. \tag{6.17}$$

Although the choice of weights does not affect the model fit, it has an impact on the scores' values and thus issues related to or depending on them. The most common choices for weights are uniform ($w_{1i} = w_{2j} = 1, \ i = 1, \ldots, I \ \ j = 1, \ldots, J$) or *marginal* ($w_{1i} = \pi_{i+}, \ i = 1, \ldots, I$ and $w_{2j} = \pi_{+j}, \ j = 1, \ldots, J$). Uniform weights are preferred when the marginal distributions are not fixed and interest lies on comparing tables with unequal marginal distributions. The marginal weights are the choice when scores of association models have to be compared to correspondence analysis results (see Sect. 7.2) or when merging rows and/or columns of a table is the issue (see Sect. 7.5). For a more detailed discussion on the choice of the weighting system, please see Goodman (1985, 1991) or Becker and Clogg (1989).

Replacing the standard constraints (6.5) and (6.6) by the more general (6.17) and working analogously as for deriving (6.7), the intrinsic association parameter φ satisfies

$$\varphi = \sum_{i,j} w_{1i} w_{2j} \alpha_i \nu_j \log \pi_{ij} \ ,$$

i.e., it is a weighted measure of correlation between the row and columns of the table. However, as already stated, parameter φ is redundant in models R, C, and RC.

Models LL, U, R, and C are log-linear while RC is log-multiplicative (not linear in its parameters). As already mentioned, models LL and U require that both classification variables of the contingency table are ordinal and thus are sensitive in re-ordering of rows or columns. Similarly, model R (C) is invariant in re-ordering of the rows (columns) of the table and the corresponding classification variable needs not necessarily be ordinal. Ordinality is required only for columns (rows). Finally, the RC model is invariant in re-ordering of columns or rows. Hence, it can also be applied to tables with nominal classification variables. Overall, parametric scores in an association model can correspond either to nominal underlying classification variable or to ordinal with unknown distances between successive categories. Thus, the parametric scores of models R, C, and RC need not necessarily be monotone. Lack of monotonicity implies non-monotone association, in the sense that local association will be positive in some areas of the table and negative in others.

Thus, monotonicity of the row and column scores is naturally connected to positive dependence and stochastic ordering of the conditional distributions in rows or columns of the table. In particular, Goodman (1981a) showed that under the RC model, association is isotropic and tables possessing this property are TP_2, i.e.,

$\theta_{ij} \geqslant 1$ for all $i = 1,\ldots,I-1,\ j = 1,\ldots,J-1$ with at least one strict inequality (see also Sect. 2.5.5). As indicated by (6.10), in case the row and column scores are both ordered and of the same ordering (i.e., both increasing or both decreasing), $\varphi > 0$ is equivalent to positive dependence and consequently the conditional row or column probabilities are stochastically ordered. This means that if $\mathbf{X_i}$ and $\mathbf{X_{i'}}$ are the conditional row distributions of rows i and i' with $i < i'$, then the positive dependence implies $\mathbf{X_i} \leqslant_{\mathbf{st}} \mathbf{X_{i'}}$, e.g., $\mathbf{X_i}$ is stochastically smaller than $\mathbf{X_{i'}}$. The distribution of $\mathbf{X_i}$ is said to be stochastically smaller than that of $\mathbf{X_{i'}}$, if $F_{\mathbf{X_i}}(t) \geqslant F_{\mathbf{X_{i'}}}(t)$, for all $t = 1,\ldots,J$, where $F_{\mathbf{X}}$ is the cumulative distribution function of $\mathbf{X}$.

In general, it is not ensured that the ML estimates of monotone parametric scores will be monotone as well. If the ML estimates are non-monotonic, then one can proceed in order-restricted estimation of the corresponding association model (see Sect. 6.8.2).

Another nice property of association models is their connection to the bivariate normal distribution. In fact, association models lead to very good approximations of the discretized bivariate normal distribution (Goodman 1981b, 1985; Wang 1987; Becker 1989a; Rom and Sarkar 1990). To see this, consider the bivariate normal density

$$f(x,y;\propto_x,\propto_y,\sigma_x,\sigma_y,\rho) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times \exp\left(-\frac{1}{2(1-\rho^2)}\left[(\frac{x-\propto_x}{\sigma_x})^2 - 2\rho(\frac{x-\propto_x}{\sigma_x})(\frac{y-\propto_y}{\sigma_y}) + (\frac{y-\propto_y}{\sigma_y})^2\right]\right)$$

and partition the $\Re^2$ surface in small rectangular regions $(a_{i-1} \times a_i) \times (b_{j-1} \times b_j)$, where $i = 1,\ldots,I,\ j = 1,\ldots,J$, $a_0 = b_0 = -\infty$, and $a_I = b_J = +\infty$. Then, the U model, or more precisely the symmetric U model (with $I = J$ and $\propto_i = \nu_i$), applied in the table formed by this partition approximated well the discretization of the above density. For standardized scores, parameter φ is analogue to $\frac{\rho}{1-\rho^2}$ of the normal density.

Finally, we would like to emphasize that beyond the sophisticated insight in the structure of the underlying association, if significant, one of the major strong points of the association models is the ability for conditional testing of independence, as already discussed in Sect. 6.3.

6.5 Association Models of Higher Order: The RC(M) Model

The RC model, though the less parsimonious association model considered so far and in spite of its impressive abilities and often impressive fit, is not always adequate. RC itself imposes a restrictive structure which can sometimes be insufficient to model the underlying association. It leaves $(I-2)(J-2)\ df$, enough space for more in-between models for building up the interaction until the saturated model is reached.

Indeed, one could consider to add more multiplicative terms of the RC-type. For example, the next model to consider would be

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \varphi_1 \alpha_{i1} \nu_{j1} + \varphi_2 \alpha_{i2} \nu_{j2} \ , \ \ i = 1, \ldots, I \ , \ j = 1, \ldots, J \ .$$

In fact, this idea can be extended further, as long as I and J are large enough, since in the saturated model, there are $(I-1)(J-1)$ association parameters. Thus, considering M association terms for the general case, we are led to the model

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \sum_{m=1}^{M} \varphi_m \alpha_{im} \nu_{jm} \ , \ i = 1, \ldots, I \ , \ j = 1, \ldots, J \ , \qquad (6.18)$$

denoted by RC(M).

How large can M be? To answer this question one must see what M represents. The concept behind the RC(M) general association model is that of dimensionality of the underlying association and its decomposition in axes. The idea is the same as in other well-known methods of reduction of dimensionality, such as factor analysis and principal component analysis. As in these methods, for identifiability purposes as well as for convenience of interpretation, the axes to which the association is decomposed are considered to be orthogonal. In our framework, the key for this decomposition is the *singular value decomposition* (SVD) of the interaction parameters matrix $\boldsymbol{\Lambda} = \left(\lambda_{ij}^{XY}\right)_{I \times J}$ of the saturated log-linear expression. Thus, M is the rank of matrix $\boldsymbol{\Lambda}$, the parameters φ_m ($m = 1, \ldots, M$) are the associated eigenvalues while the row and column scores for a certain m are the components of the mth corresponding eigenvector. In particular, the SVD of the interaction matrix $\boldsymbol{\Lambda}$ gives

$$\boldsymbol{\Lambda} = \mathbf{M} \boldsymbol{\varphi} \mathbf{N}'$$

where $\boldsymbol{\varphi} = \text{diag}(\varphi_1, \ldots, \varphi_M)$ with $\varphi_1 \geqslant \ldots \geqslant \varphi_M > 0$ are the eigenvalues while the eigenvectors $\boldsymbol{\alpha}_m = (\alpha_{1m}, \ldots, \alpha_{Im})$ and $\boldsymbol{\nu}_m = (\nu_{1m}, \ldots, \nu_{Jm})$, associated to the mth eigenvalue, form the matrices $\mathbf{M}_{I \times M} = (\alpha_{im})$ and $\mathbf{N}_{J \times M} = (\nu_{jm})$, respectively. $\mathbf{M}$ and $\mathbf{N}$ are orthonormal, e.g., they satisfy

$$\mathbf{M}'\mathbf{M} = \mathbf{N}'\mathbf{N} = \mathbf{I}_M$$

where $\mathbf{I}_M$ is the Mth order identity matrix. The maximum possible value for the dimension of the decomposition M is $M^* = \min(I, J) - 1$. Thus, model (6.18) can be considered for $0 \leq M \leq M^*$. The associated degrees of freedom equal $df[\text{RC}(M)] = (I - M - 1)(J - M - 1)$. Model RC(0) is the independence model, RC(1) is the RC while RC(M^*) is the saturated model. The orthonormality of the eigenvectors is equivalently expressed as

$$\sum_i \alpha_{im} = \sum_j \nu_{jm} = 0 \ ,$$

$$\sum_i \alpha_{im}^2 = \sum_j \nu_{jm}^2 = 1, \quad m, \ell = 1, \ldots, M,$$

$$\sum_i \alpha_{im}\alpha_{i\ell} = \sum_j \nu_{jm}\nu_{j\ell} = 0, \quad m \neq \ell \ .$$

Note that the first two restrictions are the identifiability constraints we have already imposed on the row and column scores of the RC model for uniform weights while the last one corresponds to the orthogonality of the dimensions. In order to generalize the above constraints and allow the use of weights, the *generalized singular value decomposition* (GSVD) of the interaction matrix Λ has to be applied instead of the SVD. By GSVD, $\mathbf{M}$ and $\mathbf{N}$ are orthonormalized with respect to the weights

$$\mathbf{W}_1 = \text{diag}(w_{11}, \ldots, w_{1I}) \text{ and } \mathbf{W}_2 = \text{diag}(w_{21}, \ldots, w_{2J}),$$

e.g., they satisfy

$$\mathbf{M}'\mathbf{W}_1\mathbf{M} = \mathbf{N}'\mathbf{W}_2\mathbf{N} = \mathbf{I}_M,$$

or equivalently, the row and column scores satisfy the constraints:

$$\sum_i w_{1i}\alpha_{im} = \sum_j w_{2j}\nu_{jm} = 0, \quad m = 1, \ldots, M \ , \tag{6.19}$$

$$\sum_i w_{1i}\alpha_{im}\alpha_{i\ell} = \sum_j w_{2j}\nu_{jm}\nu_{j\ell} = \delta_{m\ell}, \quad m, \ell = 1, \ldots, M,$$

where $\delta_{m\ell}$ is Kronecker's delta.

Analogously to the RC, the RC(M) model can alternatively be expressed by the multiplicative form, used by Goodman:

$$m_{ij} = \alpha_i \beta_j \exp\left(\sum_{m=1}^{M} \varphi_m \alpha_{im} \nu_{jm} \right), \ \ i = 1, \ldots, I \ , \ j = 1, \ldots, J.$$

However, the most convenient expression for physical interpretation is in terms of the local odds ratios

$$\log \theta_{ij}^L = \sum_{m=1}^{M} \varphi_m (\alpha_{im} - \alpha_{i+1,m})(\nu_{jm} - \nu_{j+1,m}) \ , \tag{6.20}$$

for $i = 1, \ldots, I-1$, $j = 1, \ldots, J-1$.

6.5.1 Maximum Likelihood Estimation of the RC(M) Model

The estimation procedure for the RC(M) follows the lines of the procedure described in Sect. 6.2 for the simple RC model. The extension for the RC(M) model is straightforward. Thus it can be proved that the likelihood equations for the RC(M) model are the (4.14) for the main effects while the likelihood equations corresponding to the row and column scores and the association parameters φ_m's are

$$\sum_j \nu_{jm}(\hat{m}_{ij} - n_{ij}) = 0, \qquad i = 1,\ldots,I, \;\; m = 1,\ldots,M\,, \tag{6.21}$$

$$\sum_i \propto_{im}(\hat{m}_{ij} - n_{ij}) = 0, \qquad j = 1,\ldots,J, \;\; m = 1,\ldots,M\,, \tag{6.22}$$

$$\sum_{i,j} \propto_{im} \nu_{jm}(\hat{m}_{ij} - n_{ij}) = 0, \qquad m = 1,\ldots,M\,, \tag{6.23}$$

i.e., straightforward extensions of (6.12), (6.13), and (6.14), respectively.

In practice, the updating equations of the simple Newton's unidimensional method for the interaction parameters of RC(M) are direct extensions of the corresponding updating equations for the RC model, presented in Sect. 6.2, while the updating equations for the main effects remain the same. The orthonormal constraints that must be satisfied by the scores of RC(M) need not to be enrolled in the iterative procedure. Since it is only a matter of parameters' identifiability and rescaling that does not affect the cell estimates, it is sufficient if they are fulfilled by the initial values and if the final estimated interaction parameters are rescaled by SVD at the final stage, after the convergence of the algorithm is achieved. The initial values $\varphi_m^{(0)}$, $\propto_{im}^{(0)}$, and $\nu_{jm}^{(0)}$ $(m = 1,\ldots,M)$ can easily be obtained as the corresponding values of the first M terms of the SVD of the observed interaction matrix, e.g., matrix Γ with entries $\gamma_{ij} = \frac{n_{ij}}{\alpha_i^{(0)}\beta_j^{(0)}}$. The extension of the Newton–Raphson algorithm, presented in Sect. 6.2, is also straightforward.

6.5.2 Example 6.2

The data considered in Table 6.7 are from Wermuth and Cox (1998) and cross-classify people in West Germany (Central archive, 1993) according to their type of schooling completed and their age in a 5×5 table. As can be observed in Table 6.8, there exists a highly significant association between age and type of schooling which is not captured by the RC model. Hence, the consideration of an association model RC(M) with $M > 1$ is necessary. The RC(2) model is of very good fit and is the model we propose for this data set and base inference on.

In the context of the RC model, we have seen that the important information lies not on the values of the row and column scores themselves but on their distances

Table 6.7 Cross-classification of 3,673 subjects according to their age and type of school attended, West Germany 1991/92

	Age group				
Type of schooling	18–29	30–44	45–59	60–74	>74
Basic, incomplete	12	13	12	20	7
	(11.943)	(13.068)	(12.008)	(20.994)	(5.987)
Basic, complete	215	507	493	460	137
	(215.823)	(504.677	(495.243)	(458.267)	(137.990)
Medium	277	300	192	126	38
	(273.739)	(309.576)	(182.462)	(129.431)	(37.792)
Upper medium	52	91	47	15	6
	(51.859)	(91.307)	(46.776)	(16.149)	(4.909)
Intensive	233	225	102	74	19
	(235.637)	(217.372)	(109.510)	(70.158)	(20.323)

In parentheses are given the ML estimates under the RC(2) model

Table 6.8 G^2 statistics for the fit of independence and association models applied in Table 6.7

Model	G^2	d.f.	p-value
I	357.146	16	0.000
RC	24.275	9	0.039
RC(2)	2.599	4	0.627

for successive categories. Distances are the quantities that are interpreted in terms of closeness of the effect of the underlying categories on odds formed by categories of the other classification variable. For the RC(M) model with $M > 1$, the logic of interpretation is the same, as can easily be verified by definition (6.20). However, distances between rows (or columns) are now defined by the Euclidean distance in the M dimensional space. For $M = 2$, this is easily visualized on the two-dimensional space, with the i-th row ($i = 1, \ldots, I$) and the j-th column ($j = 1, \ldots, J$) being represented by the points $(\hat{\alpha}_{i1}, \hat{\alpha}_{i2})$ and $(\hat{\nu}_{j1}, \hat{\nu}_{j2})$, respectively. For our example, Fig. 6.1 presents such graphs for scores satisfying constraints (6.19) subject to the uniform (left) or marginal (right) weights. The MLEs of the scores (under uniform weights) are provided in Table 6.9.

These two graphs, though they obviously refer to different scores' values, they correspond to equivalent expressions of the RC(2) estimates just differently scaled through the choice of the weights. It is evident from the plots that the 2nd dimension captures the differentiation of row 1 from 2 (incomplete from complete basic education) and 4 from 3 and 5 (upper medium from medium and intensive education). The closeness of columns 4 and 5 (ages 60–74 or > 74) is remarkable, especially in the marginal weights plot, where they are almost indistinguishable. This observation motivates Sect. 7.5 on merging categories, where this example is revisited (Sect. 7.5.1). Though the marginal weighted scores are preferred for comparisons in rows (or in columns), the uniform weighted are more appropriate for investigating the row–column combinations of strong association. We can observe,

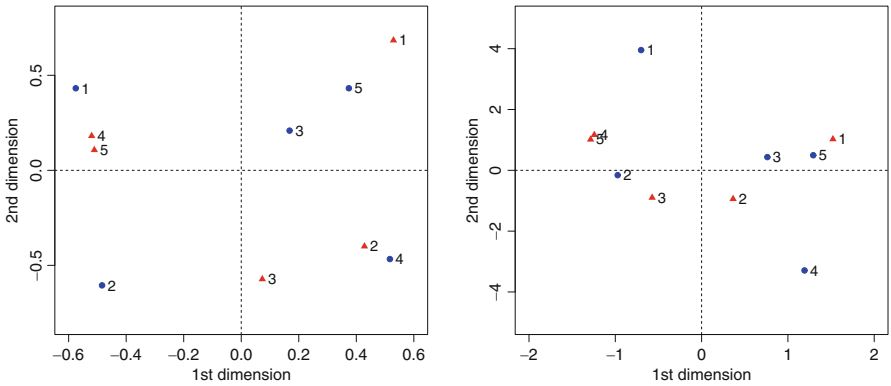


Fig. 6.1 Plots of the estimated row (*bullets*) scores $(\hat{\alpha}_{i1}, \hat{\alpha}_{i2})$, $i = 1, \ldots, 5$, and column (*triangles*) scores $(\hat{\nu}_{j1}, \hat{\nu}_{j2})$, $j = 1, \ldots, 5$, under the RC(2) model applied to Table 6.7, with respect to uniform (*left*) and marginal (*right*) weights

Table 6.9 Association parameters' ML estimates for the RC(2) model applied in Table 6.7

	$\varphi_1 = 1.7830$		$\varphi_2 = 0.6904$	
	α-scores		ν-scores	
	$m = 1$	$m = 2$	$m = 1$	$m = 2$
1	−0.575	0.431	0.529	0.684
2	−0.484	−0.605	0.428	−0.400
3	0.168	0.209	0.073	−0.573
4	0.517	−0.467	−0.520	0.181
5	0.374	0.432	−0.511	0.107

The estimates are subject to the orthonormal constraints with uniform weights

for example, that upper medium education (row 4) is stronger associated with people aged 30–44 (column 2). Also, as expected basic incomplete education (row 1) is more often among elder people (columns 4 and 5). Alternatively, one could apply Correspondence Analysis (CA) and conclude to very similar results. The CA of this data set is provided in Sect. 7.2.2.

6.6 Software Applications for Association Models

Association models, though so powerful tools in modeling the association in contingency tables, did not receive the attention one would expect. The major reason for that is the fact that their fit is not provided as a standard option in statistical software. They can be fitted in statistical packages, but some extra programming is required. Additionally, a Fortran algorithm for ML estimation of the RC(M) model by the Newton's unidimensional method has been provided by Becker (1990a) while the Newton–Raphson algorithm has been implemented in Fortran by Haberman

(1995) and a Fisher's scoring type algorithm using the weighted least squares as initial estimates by Ait-Sidi-Allal et al. (2004). As already mentioned in Sect. 6.2, the above algorithms are appropriate for the estimation and fit of models linear in their parameters, i.e., models U, R, and C. In case of RC(M), $M \geq 1$, we still apply these methods by considering at each step of estimation for the row (columns) scores that the column (row) scores are fixed at the estimated value of the previous step. This procedure is continued until convergence is achieved.

Association models which are linear in their parameters, e.g., the models U (and LL), R, and C, are log-linear and can be fitted as GLM by any available software, adopting the procedure described next for `R`. In the web appendix (see Sect. A.4) are also available syntax codes for automatized fitting of all the association models in SPSS, including the RC(M).

6.6.1 Association Models in R: Example 6.1

The simple association models U (or LL), R, and C can be fitted in `R` straightforward, in the generalized linear models framework by the `glm()` function of `R`, as described below.

First of all, the data have to be in the standard format for fitting classical log-linear models. Thus, let `freq`, `row`, and `col` be the usual variables of a data frame corresponding to the vectors of observed frequencies and row and column classification variables, respectively. Then, we have to construct the variables of row and column scores, `mu = row` and `nu = col`, respectively. This way, the row and column scores are set equal to $\propto_i = i$, $i = 1, \ldots, I$ and $\nu_j = j$, $j = 1, \ldots, J$. In the sequel, `row` and `col` have to be defined as factors and then the U, R, and C models are the log-linear models with terms in the model `row + col + mu:nu`, `row + col + row:nu` and `row + col + mu:col`, respectively. The LL model can be fitted as the U model with the only difference that the score variables `mu` and `nu` will now contain the values of the prefixed, not equidistant scores for the corresponding row and column categories.

To illustrate, let us consider the cannabis example. The data are saved under the data frame `cannabis.fr`. Models U, R, and C are fitted by `glm()` as follows:

```
> freq <- c(204,6,1,211,13,5,357,44,38,92,34,49)
> row <- rep(1:4, each=3);  col <- rep(1:3,4)
> mu <- row;  nu <- col
> row <- factor(row);  col <- factor(col)
> cannabis.fr <- data.frame(freq, row, col, mu, nu)
> model.U <- glm(freq~row+col+mu:nu, poisson, data=cannabis.fr)
> model.R <- glm(freq~row+col+row:nu, poisson, data=cannabis.fr)
> model.C <- glm(freq~row+col+mu:co„ poisson, data=cannabis.fr)
```

Table 6.10 Output of the U model fit in R for the cannabis data (Table 6.1)

```
Call:
glm(formula = freq~row + col + mu:nu, family = poisson, data = cannabis.fr)

Deviance Residuals:
      1       2       3       4       5       6       7
-0.03133 0.13352 0.13252 -0.02757 -0.02850 0.23335 0.22138
      8       9      10      11      12
-0.69889 0.10336 -0.34184 0.82418 -0.17906
Coefficients:
            Estimate Std. Error z value    Pr(>|z|)
(Intercept)  4.51766    0.10017  45.098    < 2e-16 ***
row2        -0.76921    0.12181  -6.315    2.70e-10 ***
row3        -1.05962    0.17968  -5.897    3.70e-09 ***
row4        -3.17104    0.30478 -10.404    < 2e-16 ***
col2        -4.38621    0.25357 -17.298    < 2e-16 ***
col3        -7.06112    0.53471 -13.205    < 2e-16 ***
mu:nu        0.80265    0.07827  10.255    < 2e-16 ***
--
Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1357.7001 on 11 degrees of freedom

Residual deviance: 1.4687 on 5 degrees of freedom

AIC: 79.655

Number of Fisher Scoring iterations: 4
```

for models U, R, and C, respectively. The output is then derived, for the U model, for example, through the command

```
> summary(model.U)
```

and is provided in Table 6.10.

Note that by the above procedure, the scores involved in the interaction term are not standardized, they are equal to the corresponding category index and thus they do not satisfy constraints (6.5) and (6.6). For these scores, we have $\hat{\varphi} = 0.80265$, as given in Sect. 6.1.2. This, however, does not affect the ML estimates of the common value of all expected under U local odds ratios or of the expected cell frequencies, which can be obtained by

```
> MLE.U <- xtabs(model.U$fitted.values~row+col)
```

verifying the corresponding entries of Table 6.1.

The R and C models fitted above are defined by (6.15), the parameterization without the intrinsic association parameter φ. If we want the models to be in the form (6.4) and the scores to be standardized, then `mu` and `nu` have to be rescaled appropriately before applying the model, while the estimates of the parametric scores have to be rescaled at a final stage as well. To simplify this procedure, we conducted for each association model the corresponding R function, namely

the `fit.U()`, `fit.R()` and `fit.C()`, to be found in the web appendix (see Sect. A.3.5). They fit the corresponding model subject to the general constraints (6.17), controlling the weights used by the parameter `iflag`, with the option of uniform (=0) or marginal (=1) weights. Hence, the U model on the cannabis example with marginal weights could be fitted by this function as

```
> U <- fit.U(freq, NI=4, NJ=3, iflag=1)
```

where `NI`=I and `NJ`=J. Under `U`, additional to the standard `glm` output, the standardized scores are saved under `U$mu` and `U$nu`, respectively, as well as $\hat{\varphi}$ (`U$phi`), the G^2 value (`U$G2`), the degrees of freedom (`U$df`), the p-value (`U$p.value`), and the ML estimates of the expected cell frequencies (`U$fit.freq`). Functions `fit.R()` and `fit.C()` are called analogously and give output of the same format.

The RC and more generally the RC(M) models, $M \geq 1$, cannot be fitted by `glm`, since they are not linear in their parameters and thus not in the GLM family. Thus, these models need special treatment. They can be fitted through functions available in special packages developed for nonlinear models, such as `gnm`, developed by Turner and Firth. An overview of version 1.0–6 is provided by Turner and Firth (2012a). An alternative choice is the `VGAM` package, which deals with Vector Generalized Additive Models (Yee and Wild 1996). For a short presentation of the package, see Yee (2008).

We will illustrate association models by the `gnm` package, based on Turner and Firth (2007). It is designed for models multiplicative in their parameters and defines the product of parameters, corresponding to factors f1 and f2, respectively, through `Mult(f1,f2)`. Thus, the RC model is fitted on our cannabis example by

```
> library(gnm)
> RC.model<-gnm(freq~row+col+Mult(row,col),family=poisson)
```

Recall that `row` and `col` have to be defined as factors before calling the model. Output is printed on the screen by typing

```
> RC.model
```

The output is provided in Table 6.11.

The ML estimates of the expected cell frequencies under the RC model are provided by

```
> predict(RC.model, type="response", se.fit=TRUE)
```

```
$fit
         1          2          3          4          5          6
204.249905   5.393465   1.356630 211.279836  12.320820   5.399344
         7          8          9         10         11         12
355.826805  46.847419  36.325776  92.643454  32.438296  49.918250

$se.fit
         1          2          3          4          5          6
14.2819358  1.9407176  0.8911325 14.5093040  2.8007752  1.9625848
         7          8          9         10         11         12
18.7968741  5.6652758  5.5879960  9.5879954  5.3135278  6.9616519

$residual.scale
[1] 1
```

The ML estimates of the parameters of the model are printed on screen by typing:

```
> coefficients(RC.model)
```

Table 6.11 Output for the RC model applied on the cannabis example (data in Table 6.1) by `gnm`

```
   Call:
gnm(formula=freq~row+col+Mult(row,col),family=poisson)

      (Intercept)                row2                row3
          4.97406             0.20900             0.91031
             row4                col2                col3
         -0.21912            -2.07242            -2.35822
 Mult(.,col).row1   Mult(.,col).row2   Mult(.,col).row3
         -0.67224            -0.33122             0.01931
 Mult(.,col).row4   Mult(row,.).col1   Mult(row,.).col2
          0.44034            -0.51364             1.80955
 Mult(row,.).col3
          3.43750

  Deviance: 0.5888162
  Pearson chi-squared: 0.5802588
  Residual df: 2
```

Furthermore, the command `coef()` gives the ability to save the ML estimates of a parameter in a separate vector in order to be handy for further use. For example, the row main effects estimates can be saved under the vector `a`:

```
> a<-c(0,coef(model.RC) [2:4])
> a
```

```
                    row2        row3         row4
   0.0000000   0.2090018   0.9103069   -0.2191229
```

Note that the model is fitted through (6.15) and the scores' estimates are not with respect to the constraints (6.17). They can be rescaled linearly though, in order to satisfy them. The `getContrasts()` command of the `gnm` package provides this facility. Thus, for uniform weights, the rescaling is achieved as

```
mu<-getContrasts(model.RC, pickCoef(model.RC,"[.]row"),
+                          ref="mean", scaleWeights="unit")
```

and

```
> nu<-getContrasts(model.RC, pickCoef(model.RC,"[.]col"),
+                          ref="mean", scaleWeights="unit")
```

for the row and column scores, respectively, leading to

```
> mu
```

```
                            Estimate        Std. Error
   Mult(., col).row1       -0.6494141       0.07259224
   Mult(., col).row2       -0.2364548       0.10092333
   Mult(., col).row3        0.1880143       0.05142579
   Mult(., col).row4        0.6978546       0.04442136
```

```
> nu
```

	Estimate	Std. Error
Mult(row, .).col1	-0.74474733	0.04043242
Mult(row, .).col2	0.08252274	0.09813196
Mult(row, .).col3	0.66222459	0.05769954

For marginal weights, the vectors of row and column marginal probabilities have to be computed first:

```
> rowProbs<-with(cannabis.fr, tapply(freq,row,sum)/sum(freq))
> colProbs<-with(cannabis.fr, tapply(freq,col,sum)/sum(freq))
```

The rescaling follows then analogously:

```
> mu<-getContrasts(model.RC, pickCoef(model.RC,"[.]row"),
+                          ref=rowProbs, scaleWeights=rowProbs)
> nu<-getContrasts(model.RC, pickCoef(model.RC,"[.]col"),
+                          ref=colProbs, scaleWeights=colProbs)
```

For our example, this leads to

```
>mu
```

	Estimate	Std. Error
Mult(., col).row1	-1.5106642	0.17070781
Mult(., col).row2	-0.5686042	0.22575503
Mult(., col).row3	0.3997124	0.08382567
Mult(., col).row4	1.5627815	0.14715755

and

```
>nu
```

	Estimate	Std. Error
Mult(row, .).col1	-0.2849555	0.003252048
Mult(row, .).col2	0.8920811	0.141339476
Mult(row, .).col3	1.7168786	0.117206021

Alternatively, the RC model can be fitted by the function `fit.RC()`, provided in the web appendix (see Sect. A.3.5), with the option of selecting marginal or uniform weights for the constraints (6.17) on the scores. The function is called exactly as `fit.U()` and provides the same type of output. However, this function does not provide the standard errors of the parametric scores. For this, the `getContrasts()` function described above is needed.

The conditional testing between nested association models, when allowed, can be performed by function `anova()`. Thus, for our cannabis example, the ANOAS table based on the conditional tests $G^2(\mathrm{I}|\mathrm{U})$, $G^2(\mathrm{U}|\mathrm{C})$, and $G^2(\mathrm{C}|\mathrm{RC})$ (see Sect. 6.3.1) is produced by

```
> I<-glm(freq~row+col, family=poisson)
> m1<- fit.U(freq,4,3,1)
> m2<- fit.C(freq,4,3,1)
> m3<- fit.RC(freq,4,3,1)
> anova(I,m1$model,m2$model,m3$model,test="Chisq")
```

```
Analysis of Deviance Table

Model 1: freq ~ row + col
Model 2: freq ~ X + Y + mu:nu
Model 3: freq ~ X + Y + Y:mu
Model 4: freq ~ X + Y + Mult(X, Y)

   Resid. Df  Resid. Dev  Df  Deviance   P(>|Chi|)
1       6       152.793
2       5         1.469     1   151.325   <2e-16 ***
3       4         1.100     1     0.368     0.5439
4       2         0.589     2     0.512     0.7743
--
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.6.2 The RC(M) Model in R: Example 6.2

Association models of order M, $M > 1$, can be fitted in the `gnm` package applying the `instances` argument for the multiplicative term of the model. Thus, for our Example 6.2 (Table 6.7), the RC(2) model can be fitted as

```
> RC2.model <- gnm(freq ~ row+col+instances(Mult(row,col),2),
+                          family=poisson)
```

where `freq` is the vector of cell frequencies while `row` and `col` are the factors corresponding to the rows (type of schooling) and columns (age group) of the table, respectively. The `fit.RCm()` function in the web appendix (see Sect. A.3.5) fits the RC(M) model ($M \geq 1$) on a contingency table, read in vector form, and rescales the row and column score vectors through singular value decomposition of the appropriate table, so that constraints (6.19) hold for uniform or marginal weights.

Thus, for Example 6.2, the 5×5 data table is provided in vector form (by rows) as

```
> WCox <- c(12,13,12,20,7,215,  507,493,460,137,
+           277,300,192,126,38,52,91,47,15,6,233,225,102,74,19)
```

and the RC(2) model is fitted by

```
> m <- 2
> RC.m <- fit.RCm(freq=WCox, NI=5, NJ=5, m=2, iflag=1)
```

where the parameter `m` specifies the order of the association model. The derived score vectors are subject to constraints (6.19) with marginal weights. Changing the last argument of `fit.RCm()` from 1 to 0, the uniform weights are applied.

One can save the scores' estimates in order to proceed with the presentation of the results, for example, through appropriate graphs. For Example 6.2, the vectors of row and column scores subject to marginal weights can be saved in vectors `mu1` and `nu1` as

```
> mu1 <- RC.m$mu ;  nu1 <- RC.m$nu
```

while subject to uniform weights in `mu0` and `nu0` as

```
> RC.m0 <- fit.RCm(freq=WCox, NI=5, NJ=5, m=2, iflag=1)
> mu0 <- RC.m0$mu ;  nu0 <- RC.m0$nu
```

respectively.

The plot, for example, of the row and column scores' coordinates under uniform weights (Fig. 5.1 (left)) can easily be obtained through the standard `plot()` command, applied on `mu0` and `nu0`. This plot is produced by function `plot_2dim()`, provided in the web appendix (see Sect. A.3.5). The plot in Fig. 5.1 (left) is obtained by calling this function as

```
> plot_2dim(mu0, nu0, -0.6, 0.6, -0.8, 0.8, -0.7, 1.1, 1.2)
```

The parameters of this function following `nu0` control the plot appearance. Thus, $(-0.6, 0.6)$ and $(-0.8, 0.8)$ define the range of values of the first and second axis, respectively. The value set -0.7 leaves a gap of 70% of the text width between the category label and the corresponding plotted symbol. According to the case, it can be adjusted each time for the better appearance of the graph. The size of text characters in axes and labels is set to be 1.1 times the default text size while the size of the symbols and their categories' labels are 1.2 times the default text size. Analogously, the plot in Fig. 5.1 (right) is obtained through

```
> plot_2dim(mu1, nu1, -2, 2, -5, 5, -0.7, 1.1, 1.2)
```

6.6.3 Example 2.4 (Revisited)

Recall the data set on varicella disease in Table 2.5, where 170 children are cross-classified by complication occurrence and age (in a 2×4 table). Independence was rejected (p-value=0.040) and the linear trend test suggested that the linear association is non-significant (p-value=0.104). Fitting association models on this example, we confirm the inappropriateness of linear association since the U model is rejected with $G^2 = 7.093$ (p-value=0.029, $df = 2$). Note that because $I = 2$, the R model is equivalent to the U while the C model is saturated ($G^2 = 0$, $df = 0$). However, derivation of the column scores of the C model is very informative on comparing the different age groups in terms of their association to the complication response. The C model is fitted by function `fit.C` of the web appendix (see Sect. A.3.5). From the corresponding output, the coefficients along with their standard errors and significances are provided below.

```
Coefficients: (1 not defined because of singularities)
             Estimate   Std. Error   z value   Pr(>|z|)
(Intercept)  2.1503     0.2759       7.794     6.47e-15 ***
X2           -0.2063    0.1944       -1.062    0.2885
Y2           0.3980     0.3399       1.171     0.2416
Y3           0.2939     0.3395       0.866     0.3867
Y4           1.9272     0.2759       6.986     2.84e-12 ***
Y1:mu        -0.1522    0.2759       -0.552    0.5811
Y2:mu        0.6024     0.2415       2.495     0.0126 *
Y3:mu        0.2470     0.2409       1.025     0.3053
Y4:mu        NA         NA           NA        NA
--
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the estimation procedure above, the parametric score ν_4 is redundant and is the reference category (coefficient for `Y4:mu`, shown as not defined). The fixed row scores used are $\propto_1 = -1$ and $\propto_2 = 1$. The estimated column scores are rescaled to satisfy the marginal weighted constraints (6.17). The rescaled scores and the interaction parameter are also part of the output. In particular, $\hat{\varphi} = 0.231$, $\hat{\nu}_1 = -1.127$ $\hat{\nu}_2 = 2.136$, $\hat{\nu}_3 = 0.560$, and $\hat{\nu}_4 = -0.468$.

Observing that only the column score $\hat{\nu}_2$ is significantly different from the others, we conclude that only the category "1–2 years old" relates differently to complications than all other age categories. Thus, we proceed by applying the LL model with the constraint $\nu_1 = \nu_3 = \nu_4$. In `R` this is easily achieved, as shown next. Before fitting the LL model, we rescale the simple raw scores 1,2, assigned initially to the rows and column categories, through the function `rescale` of the web appendix (see Sect. A.3.5), so that the $\hat{\varphi}$, derived by `glm()`, corresponds to the marginally weighted scores:

```
> NI <- 2
> NJ <- 4
> freq <- c(10,7,9,59,6,19,12,48)
> row<-gl(NI,NJ,length=NI*NJ)
> col<-gl(NJ,1,length=NI*NJ)
> dtable <- data.frame(freq,row,col)
> mu0<-c(1,2)
> nu0<-c(1,2,1,1)
> mu<-rep(rescale(mu0, dtable, 1, 1)$score,each=NJ)
> nu<-rep(rescale(nu0, dtable, 1, 0)$score, NI)
> LL.model <- glm(freq~row+col+mu:nu,poisson)
```

From the summary output, obtained by `summary(LL.model)`, we see that the model is acceptable, since $G^2 = 1.572$ (p-value=0.456, $df = 2$). Commands

```
MLEs <- xtabs(LL.model$fitted.values ~ row + col)
stdres <- xtabs(rstandard(LL.model) ~ row + col)
```

express in table form the ML estimates of the expected frequencies and the corresponding standardized residuals. None of the standardized residuals exceeds 1.96; thus all cells are fitted satisfactorily by the model.

```
> MLEs
        col
row            1           2            3           4
1       8.666667    7.000000    11.375000    57.958333
2       7.333333   19.000000     9.625000    49.041667
> stdres
        col
row            1           2            3           4
1       0.6924565   0.0000000   -1.1684753   0.3975276
2      -0.7328918   0.0000000    1.0833608  -0.4001372
```

The interaction parameter is estimated as $\hat{\varphi} = 0.210$ (coefficient for `mu:nu` in the output) while the row and column scores are $\alpha_1 = -1$, $\alpha_2 = 1$, $\nu_1 = \nu_3 = \nu_4 = -0.4249$ and $\alpha_2 = 2.3534$ (saved in vectors `mu` and `nu`).

This model provides a clear and strong interpretation. The equality restrictions among the column scores impose on the expected local odds ratios the restrictions $\theta^L_{13} = 1$ and $\theta^L_{11} = 1/\theta^L_{12}$. The odds ratios θ^{23}_{11} and θ^{24}_{11}, opposing age categories 1 to 3 and 1 to 4, respectively, are also equal to 1 and $\theta^{24}_{12} = \theta^L_{12}$. Thus we conclude that the odds of complication occurrence for children 1–2 years old is $\hat{\theta}^L_{11} = e^{\hat{\varphi}(\alpha_2-\alpha_1)(\nu_2-\nu_1)} = e^{1.1669} = 3.2$ times higher than for children of any other age. The $\hat{\theta}^L_{1j}$, $j = 1,2,3$, could also have been computed by the `local.odds.DM()` function (see Sect. A.3.2), implemented as follows:

```
> NI <- 2;  NJ <- 4;  C <- local.odds.DM(NI, NJ)
> LO <- as.vector(C%*%log(LL.model$fitted.values))
> exp(t(matrix(LO, NJ-1)))
```

```
          [,1]      [,2]  [,3]
[1,]  3.207792  0.3117409     1
```

6.6.4 *Association Models Fitted on the Local Odds Ratios*

Association models can also be fitted directly on the local odds ratios through the generalized log-linear model GLLM (5.28) and implemented in `R` by Lang's `mph` package. The GLLM turns to a model on local odds ratios by eliminating matrix $\mathbf{M}$ and appropriately defining matrix $\mathbf{C}$, so that $\mathbf{C}\log(\mathbf{m})$ becomes the vector of the expected local odds ratios under the assumed model, where $\mathbf{m}$ is the vector of expected cell frequencies. For an $I \times J$ table, $\mathbf{C}$ and $\mathbf{m}$ are of size $(I-1)(J-1) \times IJ$ and $IJ \times 1$, respectively. This matrix $\mathbf{C}$ for an $I \times J$ table is produced by function `local.odds.DM()` of the web appendix (see Sect. A.3.2). The design matrix $\mathbf{X}$ specifies the restrictions imposed on the local odds ratios by the model under consideration and is of size $(I-1)(J-1) \times s$, where s is the number of parameters in the model.

We illustrate this option fitting the U model on our cannabis example. Under the U model a common value is assumed for all local odds ratios; thus the parameter β is scalar and the design matrix $\mathbf{X}$ is the $(I-1)(J-1)\times 1$ vector of 1's. Recall (see Sect. 5.6) that `mph` needs to be actualized in `R` and that the data are read as a vector saved in matrix form. The way we define the $\mathbf{C}$ matrix requires the data to be read by rows. For the cannabis example,

```
> y <- c(204,6,1,211,13,5,357,44,38,92,34,49)
> y <- matrix(y);  NI <- 4;  NJ <- 3;  dim1<-(NI-1)*(NJ-1)
> X<-matrix(rep(1,dim1))
```

In our context, matrix $\mathbf{C}$ is

```
> C <- local.odds.DM(NI,NJ)
```

and the link of the GLLM model is defined by the function

```
> L.fct <- function(m){C%*%log(m)}
```

Finally, the U model is fitted by

```
> mph.out <- mph.fit(y=y,L.fct=L.fct,X=X)
> mph.summary(mph.out,cell.stats=T,model.info=T)
```

The derived output provides goodness-of-fit statistics for the model; estimate of the β (the log of the common local odds ratio under U), and estimates of the expected cell frequencies and the associated residuals. Further informations, as for example on the algorithm's convergence, are also provided.

In case of one or more sampling zeros, when working with odds ratios and for ensuring their existence, we set

```
> z <- y+0.000001
> mph.out <- mph.fit(y=z,L.fct=L.fct,X=X)
```

6.7 Association Models for Multi-way Tables

Association models can also be applied on contingency tables of higher dimension. Consider a $I\times J\times K$ contingency table with classification variables X, Y, and Z, respectively. Association models can be derived by replacing one or more of the interaction terms of any hierarchical log-linear model by multiplicative terms based on scores, leading thus to more parsimonious models of special structure, in analogy to two-way association models.

For example, consider the model

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \varphi^{XZ}\alpha_i\tau_k + \varphi^{YZ}\nu_j\tau_k\,, \qquad (6.24)$$
$$i=1,\ldots,I,\ j=1,\ldots,J,\ k=1,\ldots,K,$$

with $(\alpha_1,\ldots,\alpha_I)$, $(\nu_1,\ldots,\nu_J)$, and $(\tau_1,\ldots,\tau_K)$ sets of known scores assigned to the categories of the classification variables X, Y, and Z, respectively, all equidistant for successive categories. This model is a special type of conditional XY independence model, derived from the $(XZ,\ YZ)$ log-linear model by replacing the λ_{ik}^{XZ} and λ_{jk}^{YZ} interaction terms by the uniform (U)-type terms $\varphi^{XZ}\alpha_i\tau_k$ and $\varphi^{YZ}\nu_j\tau_k$, respectively.

For this, it will be denoted as $(XZ_U,\ YZ_U)$. It is very parsimonious, having $df = IJK - I - J - K$, just 2 less than the complete independence model $(X,\ Y,\ Z)$.

More options are available by considering some of the scores to be parametric. Assuming thus an R-type interaction only for the term λ_{ik}^{XZ}, the $(\alpha_1, \ldots, \alpha_I)$ scores would be considered as parameters in (6.24) and the model would then be $(XZ_X,\ YZ_U)$. In terms of notation, an interaction term without an index is of log-linear model type, with U of uniform association type, while when it is of row or column effect type, the variable of parametric scores is set as an index. A multiplicative RC-type has also a multiplicative index. Thus $(XZ_{XZ},\ YZ_U)$ is the model defined by (6.24) with parametric α- and τ-scores of the XZ interaction and fixed equidistant scores for the ν- and τ-scores of the YZ interaction term. Were the layer scores parametric in both interaction terms, they could be homogeneous or not. The model for parametric nonhomogeneous τ-scores, $(XZ_Z,\ YZ_Z)$, is

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \varphi^{XZ} \alpha_i \tau_k^{XZ} + \varphi^{YZ} \nu_j \tau_k^{YZ} \ ,$$

while with the additional restrictions $\tau_k^{XZ} = \tau_k^{YZ}$, $k = 1, \ldots, K$, the homogeneous $(XZ_Z,\ YZ_Z)$ is derived.

A flexible form of association model, including three-factor interaction, is

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \varphi^{XY} \alpha_i^{XY} \nu_j^{XY} + \varphi^{XZ} \alpha_i^{XZ} \tau_k^{XZ} + \varphi^{YZ} \nu_j^{YZ} \tau_k^{YZ} + \varphi^{XYZ} \alpha_i^{XYZ} \nu_j^{XYZ} \tau_k^{XYZ} \ , \tag{6.25}$$

which offers a variety of model options, depending on the combinations of assumptions about the scores.

The most general expressions for imposing association structures on the two-factor interaction terms of a three-way log-linear model are

$$\lambda_{ij}^{XY} = \sum_{m=1}^{M_1} \varphi_m^{XY} \alpha_{im}^{XY} \nu_{jm}^{XY} \ , \qquad \lambda_{jk}^{YZ} = \sum_{m=1}^{M_2} \varphi_m^{YZ} \nu_{jm}^{YZ} \tau_{km}^{YZ} \ , \tag{6.26}$$

$$\lambda_{ik}^{XZ} = \sum_{m=1}^{M_3} \varphi_m^{XZ} \alpha_{im}^{XZ} \tau_{km}^{XZ} \ ,$$

with $1 \le M_1 \le \min(I,J) - 1$, $1 \le M_2 \le \min(J,K) - 1$, and $1 \le M_3 \le \min(I,K) - 1$. The three-factor interaction can be decomposed in an analogue manner

$$\lambda_{ijm}^{XYZ} = \sum_{m=1}^{M_4} \varphi_m \alpha_{im} \nu_{jm} \tau_{km} \ . \tag{6.27}$$

The consideration of (6.27) for $M_4 > 1$ as well as other options for decomposing three-way arrays, known as *trilinear decomposition*, is beyond the scopes of this book (see Sect. 6.8.1).

The scores are subject to constraints analogue to (6.19) of the two-way case. When $M_i = 1$ ($i = 1,\ldots,4$), we conclude to model (6.25). Furthermore, the scores can be considered known and lead to terms of U-, R-, C-, or L- (for the layer scores) type.

The idea extends analogously to contingency tables of higher dimension. However, the number of possible association models augments with the dimension of the table. It is difficult to control all possible combinations of assumptions regarding the interaction terms of a multi-way table, so an automated stepwise association model selection procedure is not feasible. In practice we start by selecting the appropriate hierarchical log-linear model by a stepwise procedure and then try to conclude to a more parsimonious model by imposing special structures to some of the interaction terms. In this procedure, conditional tests between nested models are helpful. Finally, we can test whether parametric scores of the same classification variable but on different interaction terms are homogeneous.

Multi-way association models and their physical interpretations will be illustrated with two examples that follow.

6.7.1 Example 6.3

In a study, 16,236 teenagers in Holland are cross-classified in a $6 \times 7 \times 2$ table by their educational level after 4 years of second-level education, their test for intellectual capacity (TIC) score, and their gender (Siciliano and Mooijaart 1997). The data are provided in Table 6.12.

In the framework of hierarchical log-linear model, we do not have another option for this data set than the saturated model, since the three-factor interaction is significant. We can verify that the model of homogeneous association (EI, EG, IG) is rejected with $G^2(GE, GI, EI) = 61.517$ (p-value=0.001, df=30). It is notable however that the highly significant G^2 value is also affected by the large sample size of the table. The corresponding dissimilarity index is $\hat{\Delta} = 0.02$, at the limit for a satisfying data representation by this model (see Sects. 4.2 and 4.2.1 for calculation in `R`). The significance of each term in the log-linear model is summarized in the analysis of deviance table of the saturated model, derived as shown below.

Provided that the data are given in vector `freq`, expanded by rows, followed by columns and layers, we program in `R`

```
> G<-factor(rep(1:2,each=42));    I<-factor(rep(1:7,12))
> E<-factor(rep(1:6,2,each=7));   educ.fr<-data.frame(freq,E,I,G)
> sat.glm <- glm(freq ~ E*I*G, family=poisson, data=educ.fr)
> anova(sat.glm, test="Chisq")
```

and get the output of Table 6.13.

Since all interaction terms are significant, the basis for selecting the appropriate association model will be the saturated. The simplest model expression of this type is the

$$log(m_{ijk}) = \lambda + \lambda_i^E + \lambda_j^I + \lambda_k^G + \varphi^{EI} \alpha_i \nu_j + \varphi^{IG} \nu_j \tau_k + \varphi^{EG} \alpha_i \tau_k + \varphi^{EIG} \alpha_i \nu_j \tau_k, \tag{6.28}$$

Table 6.12 Cross-classification of 16,236 teenagers in Holland by their educational level after 4 years of second-level education, their test for intellectual capacity (TIC) score, and their gender (Siciliano and Mooijaart 1997)

Gender		TIC							Total
		1	2	3	4	5	6	7	
Boys	Education								
	DO	75	77	105	125	89	38	17	526
		(57.24)	(79.28)	(127.13)	(128.12)	(87.64)	(37.99)	(8.60)	
	LBO	216	305	495	522	389	168	34	2129
		(212.43)	(304.47)	(505.33)	(527.06)	(373.10)	(167.38)	(39.23)	
	MAVO	67	144	267	368	339	194	54	1433
		(71.70)	(131.16)	(277.81)	(369.79)	(334.07)	(191.27)	(57.20)	
	MBO	51	84	239	345	301	208	65	1293
		(49.32)	(98.25)	(226.62)	(328.50)	(323.18)	(201.50)	(65.63)	
	HAVO	26	65	200	332	383	258	98	1362
		(31.04)	(71.86)	(192.65)	(324.57)	(371.13)	(268.95)	(101.81)	
	VWO	12	27	104	216	325	321	178	1183
		(8.81)	(27.28)	(97.81)	(220.36)	(336.94)	(326.51)	(165.28)	
	Total	447	702	1410	1908	1826	1187	446	7926
Girls	Education								
	DO	51	60	115	123	78	56	9	492
		(44.21)	(65.96)	(113.96)	(123.73)	(91.18)	(42.58)	(10.39)	
	LBO	144	223	382	370	290	107	26	1542
		(154.72)	(221.27)	(366.45)	(381.38)	(269.39)	(120.59)	(28.20)	
	MAVO	60	134	288	424	442	266	72	1686
		(64.61)	(128.52)	(296.03)	(428.51)	(420.98)	(262.10)	(85.24)	
	MBO	75	167	320	458	428	258	72	1778
		(80.81)	(152.43)	(332.96)	(457.05)	(425.81)	(251.41)	(77.54)	
	HAVO	23	68	211	373	450	402	169	1696
		(25.43)	(65.95)	(198.08)	(373.88)	(478.94)	(388.83)	(164.89)	
	VWO	5	9	77	183	307	326	209	1116
		(4.67)	(16.59)	(68.17)	(176.07)	(308.64)	(342.88)	(198.98)	
	Total	358	661	1393	1931	1995	1415	557	8310

In parentheses are given the fitted values under the association model (6.29). Educational-level scale: (1) DO, dropped out; (2) LBO, junior level of education for professions; (3) MAVO, medium level of general education; (4) MBO, senior level of education for professions; (5) HAVO, high level of general education; and (6) VWO, general education preparing for university

with all the involved set of scores known. Considering the scores in each set equidistant for successive categories, model (6.28), denoted by (EIG_U), is the most parsimonious three-way association model in the class of models with up to three-factor interaction, having just 4 parameters more than the model of complete independence (E, I, G).

Table 6.13 Decomposition of the deviance for Table 6.12

```
Analysis of Deviance Table
Model: poisson, link: log
Response: freq
Terms added sequentially (first to last)

       Df   Deviance   Resid. Df   Resid. Dev    P(>|Chi|)
NULL                        83       9063.2
E       5     1910.5         78       7152.7   < 2.2e-16 ***
I       6     4508.9         72       2643.8   < 2.2e-16 ***
G       1        9.1         71       2634.7     0.002580 **
E:I    30     2330.2         41        304.5   < 2.2e-16 ***
E:G     5      222.3         36         82.2   < 2.2e-16 ***
I:G     6       20.6         30         61.6     0.002154 **
E:I:G  30       61.6          0          0.0    0.000584 ***
--
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In order to fit association models in R, we create the known score vectors for the classification variables. For simplicity, we set each score equal to the index of the category it corresponds to. We compute

```
> mu<-rep(1:6,2,each=7);  nu<-rep(1:7,12);  tau<-rep(1:2,each=42)
```

and extend the data frame

```
> educ.fr<-data.frame(freq,E,I,G,mu,nu,tau)
```

Model (EIG_U) is then fitted as

```
> EIG.U <- glm(freq~E+I+G+mu:nu+mu:tau+nu:tau+mu:nu:tau,
+              poisson, data=educ.fr)
```

It is of very bad fit, with $G^2(EIG_U) = 450.179$ (p-value< 0.0005, df=67), but reduces the G^2 statistic drastically, compared to complete independence ($G^2(E, I, G) = 2634,719, df$=71).

This means that some of the row and/or column scores in (6.28) have to be considered parametric. Since G is binary, the corresponding scores (τ_1, τ_2) cannot be parametric and their choice does not affect the model fit.

Considering that only the TIC effect is parametric on all the interaction terms, model (6.28) extends to

$$log(m_{ijk}) = \lambda + \lambda_i^G + \lambda_j^E + \lambda_k^I + \propto_i \nu_j^{EI} + \nu_j^{IG}\tau_k + \varphi^{EG} \propto_i \tau_k + \propto_i \nu_j^{EIG} \tau_k \,,$$

denoted by $(EI_I,\ EG_U,\ IG_I,\ EIG_I)$. This last model expression employs non-standardized parametric scores and therefore the redundant intrinsic association φ-parameters are absorbed. It is fitted in R by

```
> EIG.I <- glm(freq~E+I+G+mu:I+mu:tau+I:tau+mu:I:tau,
+              poisson, data=educ.fr)
```

Table 6.14 ML estimates of the parametric scores of model (6.29) fitted on the data in Table 6.12 for equidistant scores $v_j = j$ $(j = 1,\ldots,7)$ and $\tau_k = k$ $(k = 1,2)$

$i =$	1	2	3	4	5	6
$\hat{\propto}_i^{EI}$	−0.742	−0.631	−0.473	−0.250	−0.268	0.000
$\hat{\propto}_i^{EG}$	0.438	0.456	0.583	1.319	0.458	0.000
$\hat{\propto}_i^{EIG}$	0.000	−0.077	0.009	−0.129	0.039	0.062

Its bad fit ($G^2(EI_I,\ EG_U,\ IG_I,\ EIG_I) = 426.6$, p-value< 0.0005, df=52) provides evidence that the education scores in some or all interaction terms should be considered as unknown parameters.

Thus, we try next the model

$$log(m_{ijk}) = \lambda + \lambda_i^G + \lambda_j^E + \lambda_k^I + \propto_i^{EI} v_j + \varphi^{IG} v_j \tau_k + \propto_i^{EG} \tau_k + \propto_i^{EIG} v_j \tau_k \ , \quad (6.29)$$

where only the education level (E) effect is parametric for all interaction terms. This is denoted as $(EI_E,\ EG_E,\ IG_U,\ EIG_E)$ and fitted in R by

```
> EIG.E <- glm(freq~E+I+G+E:nu+E:tau+nu:tau+E:nu:tau,
+              poisson, data=educ.fr)
```

exhibiting an adequate fit with $G^2(EI_E,\ EG_E,\ IG_U,\ EIG_E) = 61.074$ (p-value=0.267, df=55).

The fitted cell frequencies under $(EI_E,\ EG_E,\ IG_U,\ EIG_E)$ are provided in Table 6.12 in parentheses. For equidistant scores $v_j = j$ $(j = 1,\ldots,7)$ and $\tau_k = k$ $(k = 1,2)$, the ML estimate of the intrinsic association parameter φ^{IG} is $\hat{\varphi}^{IG} = 0.0745$ while the ML estimates of the parametric scores are given in Table 6.14.

The interpretation of parameters needs caution and has to be done locally due to the non-monotonicity of the parametric scores. For interpretation, the fitted odds ratios under the model have to be considered. Model (6.29) in terms of the conditional EI log local odds ratios and for the choice of known scores given above is expressed as

$$\log\left(\theta_{ij(k)}^{EI}\right) = \log\left(\frac{m_{ijk}\cdot m_{i+1,j+1,k}}{m_{i+1,j,k}\cdot m_{i,j+1,k}}\right) = (\propto_{i+1}^{EG} - \propto_i^{EG}) + (\propto_{i+1}^{EIG} - \propto_i^{EIG})k$$
$$= \log\left(\theta_{i(k)}^{EI}\right) \ , \quad i = 1,\ldots,I-1,\ j = 1,\ldots,J-1,\ k = 1,2,$$

i.e., independent of j, as expected since successive v_j scores are equidistant. This means that under (6.29) the fitted local odds ratios are constant within rows. In our case, the $\hat{\theta}_{i(k)}^{EI}$ row values are given in Table 6.15. Thus, we see that for boys, the strongest association between educational level and the TIC score is between HAVO and VWO. The odds of a boy achieving a category of TIC score vs. the immediate previous one is 1.34 times higher for a boy having general education preparing for university (VWO) than high level of general education (HAVO). The corresponding odds ratio for girls is 1.37. The conditional (within gender) association between TIC

Table 6.15 ML estimates of the $\theta^{EI}_{i(k)}$, $i = 1,\ldots,5$, $k = 1,2$, under model (6.29) for the example in Table 6.12

Education	Boys	Girls
DO	1.035	0.958
LBO	1.276	1.391
MAVO	1.089	0.948
MBO	1.162	1.375
HAVO	1.337	1.368
VWO		

score and educational level is positive for boys although not equally strong for all educational levels while for girls it is negative (though weak) when comparing DO to LBO and MAVO to MBO.

Based on model (6.29), one could further try if more parsimonious models are preferable. More parsimonious models are obtained by imposing homogeneity constraints among the vectors of the unknown education scores or considering one of them as being equidistant. It could also be tested whether less parsimonious non-log-linear models involving interaction terms multiplicative in their parameters (i.e., of RC-type) lead to a significant improvement of the fit.

6.7.2 *Homogeneous Uniform Association*

Consider a $I \times J \times K$ contingency table, consisting of K independent strata. Then the simplest association structure is to consider that for each XY partial table, all local odds ratios are equal, i.e., assume that the U model holds for each stratum. This model is defined by

$$\theta^{XY}_{ij(k)} = \theta^{XY}_k, \;\; i = 1,\ldots,I-1, \;\; j = 1,\ldots,J-1, \;\; k = 1,\ldots,K. \tag{6.30}$$

Local odds ratios however from different strata may vary and (6.30) is the *nonhomogeneous* U model. An even simpler model is the *homogeneous* U model, assuming that all strata have a common local odds ratio

$$\theta^{XY}_{ij(k)} = \theta^{XY}, \;\; i = 1,\ldots,I-1, \;\; j = 1,\ldots,J-1, \;\; k = 1,\ldots,K. \tag{6.31}$$

To illustrate these models, consider the data in Table 6.16. The first stratum of this $4 \times 3 \times 2$ data table is the cannabis example of Table 6.1 while the second stratum corresponds to an analogue survey among students of another university (artificial data).

The U model, fitted on the 4×3 partial table of the second stratum (data in Table 6.16, stratum (2)), with one of the procedures described in Sect. 6.6.1 for Example 6.1, is of good fit with $G^2_2 = 8.284$ (p-value= 0.141, $df = 5$). Under this model, the MLE of the common local odds ratio in log-scale is $\log \hat{\theta}^L_2 = 0.749$, close to the corresponding estimate for the data in the first stratum ($\log \hat{\theta}^L_1 = 0.803$), for which we had $G^2_1 = 1.469$ (p-value= 0.917, $df = 5$).

Table 6.16 Students' survey about cannabis use at two universities

	I tried cannabis ...		
Alcohol consumption	Never	Once or twice	More often
Stratum (1)			
$\leq$ Once/month	204	6	1
Twice/month	211	13	5
Twice/week	357	44	38
More often	92	34	49
Stratum (2)			
$\leq$ Once/month	311	5	4
Twice/month	339	19	12
Twice/week	429	66	57
More often	134	51	74

Stratum (1) is the data of Table 6.1; stratum (2) is artificial

The nonhomogeneous U model (6.30) for data in Table 6.16 can be derived in `mph` by

```
> source("c://Program Files//R//mph.Rcode.txt")
> freq <-c(204,6,1,211,13,5,357,44,38,92,34,49)
> freq2<-c(311,5,4,339,19,12,429,66,57,134,51,74)
> y<- matrix(append(freq,freq2))
> NI<-4;  NJ<-3;  dim1<-(NI-1)*(NJ-1);  dim2<-2*dim1
> zer<-matrix(rep(0,NI*NJ*(NI-1)*(NJ-1)),(NI-1)*(NJ-1))
> C0<-local.odds.DM(NI,NJ);  C1<-cbind(C0,zer)
> C2<-cbind(zer,C0);  C<-rbind(C1,C2)
> L.fct <- function(m){C%*%log(m)}
> X<-matrix(rep(1,dim1)); Z<-matrix(rep(0,dim1))
> X2<-rbind(cbind(X,Z),cbind(Z,X))
> mph.out <- mph.fit(y=y,L.fct=L.fct,X=X2)
> mph.summary(mph.out,cell.stats=T,model.info=T)
```

leading to $G^2 = 9.752$ (p-value=0.462) with corresponding residual df=10. In this case, since the two strata are independent, model (6.30) is equivalent to fitting the U model independently to each of the partial two-way tables. Indeed, we can verify that $G_1^2 + G_2^2 = 9.752 = G^2$ and that the ML estimates of $\log\theta_k^{XY}$, $k = 1,2$ ($\log\hat{\theta}_1^{XY} = \hat{\beta}_1 = 0.803$, $\log\hat{\theta}_2^{XY} = \hat{\beta}_2 = 0.749$) coincide with the corresponding $\log\hat{\theta}_k^L$, $k = 1,2$.

The homogeneous U model (6.31) can be fitted as follows. The `L.fct` function is defined as above but the design matrix `X2` is replaced by `X1`, defining thus a univariate parameter β instead of the bivariate (β_1, β_2) above:

```
> X1<-rbind(X,X)  # homogeneous U model for both layers
> mph.out <- mph.fit(y=y,strata=2,L.fct=L.fct,X=X1)
> mph.summary(mph.out,cell.stats=T,model.info=T)
```

Selected parts of the output are provided in Tables 6.17 and 6.18.

Table 6.17 Output of the `mph` function for the homogeneous U model, applied on the $4 \times 3 \times 2$ data of Table 6.16: the observed local log odds ratios (OBS LINK) are listed by rows, along with the ML estimate of the common under the assumed model local log odds ratio value, its s.e. and the standardized link residuals

```
MODEL GOODNESS OF FIT: Test of Ho: h(p)=0 vs. Ha: not Ho...

Likelihood Ratio Stat (df=11):  Gsq = 10.05081 (pval = 0.5258 )
Pearson's  Score Stat (df=11):  Xsq = 10.28396 (pval = 0.505 )
Generalized Wald Stat (df=11):  Wsq = 9.83676 (pval = 0.5451)

Adj Resids: -1.708 -1.452 ... 1.546 1.705,
Number |Adj Resid| > 2: 0

SAMPLING PLAN INFORMATION...
Number of strata: 1
Strata identifiers: 2
Strata with fixed sample sizes: all
Observed strata sample sizes: 2555
LINEAR PREDICTOR MODEL RESULTS...
      BETA   StdErr(BETA)  Z-ratio p-value
beta1 0.7691     0.0472   16.2809   0
       OBS LINK ML LINK StdErr(L) LINK RESID
link1   0.7395   0.7691    0.0472   -0.0596
link2   0.8362   0.7691    0.0472    0.0563
link3   0.6934   0.7691    0.0472   -0.2391
link4   0.8089   0.7691    0.0472    0.0701
link5   1.0981   0.7691    0.0472    1.2801
link6   0.5121   0.7691    0.0472   -0.8144
link7   1.2488   0.7691    0.0472    1.2460
link8  -0.2364   0.7691    0.0472   -1.1251
link9   1.0098   0.7691    0.0472    0.9707
link10  0.3129   0.7691    0.0472   -1.0692
link11  0.9058   0.7691    0.0472    0.6372
link12  0.5188   0.7691    0.0472   -0.9681
```

The equivalent expression of model (6.30) in terms of expected cell frequencies is

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \varphi^{XY} \propto_i \nu_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \varphi^{XYZ} \propto_i \nu_j \tau_k, \quad (6.32)$$
$$i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K,$$

where the set of scores $(\propto_1, \ldots, \propto_I)$, $(\nu_1, \ldots, \nu_J)$, and $(\tau_1, \ldots, \tau_K)$ are all known and equidistant for successive categories. They can be considered subject to standardization constraints or set equal to the corresponding category index.

The conditional local odds ratios under this model are fixed within partial tables equal to

$$\theta_{(k)}^{XY} = \exp\left((\varphi^{XY} + \varphi^{XYZ}\tau_k)\Delta_1\Delta_2\right), \quad (6.33)$$
$$i = 1, \ldots, I-1,\ j = 1, \ldots, J-1,\ k = 1, \ldots, K,$$

Table 6.18 Output of the `mph` function: observed and ML fitted cell frequencies under the homogeneous U model applied on the $4 \times 3 \times 2$ data of Table 6.16, along with ML estimates of the cell probabilities, standard errors, and standardized residuals

```
CELL-SPECIFIC STATISTICS...
    strata OBS       FV StdErr.FV  PROB  StdErr.PROB ADJ.RESIDS
y1     2   204 203.9403  13.4845 0.0798    0.0053       0.0247
y2     2     6   6.0413   0.9179 0.0024    0.0004      -0.0181
y3     2     1   1.0183   0.2575 0.0004    0.0001      -0.0188
y4     2   211 210.6392  13.4128 0.0824    0.0052       0.0987
y5     2    13  13.4638   1.6978 0.0053    0.0007      -0.1431
y6     2     5   4.8970   0.8612 0.0019    0.0003       0.0506
y7     2   357 352.2819  16.4381 0.1379    0.0064       0.8152
y8     2    44  48.5868   5.1333 0.0190    0.0020      -0.9936
y9     2    38  38.1313   4.4575 0.0149    0.0017      -0.0312
y10    2    92  97.1386   8.4659 0.0380    0.0033      -1.1012
y11    2    34  28.9081   3.4680 0.0113    0.0014       1.2515
y12    2    49  48.9533   5.7150 0.0192    0.0022       0.0119
y13    2   311 308.3105  16.1909 0.1207    0.0063       0.8984
y14    2     5   9.8678   1.3053 0.0039    0.0005      -1.7076
y15    2     4   1.8218   0.4348 0.0007    0.0002       1.7054
y16    2   339 337.4154  16.4817 0.1321    0.0065       0.3441
y17    2    19  23.3021   2.4133 0.0091    0.0009      -1.0353
y18    2    12   9.2825   1.4482 0.0036    0.0006       1.0162
y19    2   429 432.2293  17.6141 0.1692    0.0069      -0.4620
y20    2    66  64.4085   5.6842 0.0252    0.0022       0.2883
y21    2    57  55.3622   5.2807 0.0217    0.0021       0.3195
y22    2   134 135.0448  10.0966 0.0529    0.0040      -0.2050
y23    2    51  43.4217   4.3209 0.0170    0.0017       1.5465
y24    2    74  80.5335   7.5990 0.0315    0.0030      -1.4519
```

where $\Delta_1 = \alpha_{i+1} - \alpha_i$ and $\Delta_2 = \nu_{j+1} - \nu_j$. Distances Δ_1 and Δ_2 are constant over i and j, respectively, since the corresponding scores are equidistant. In case the scores equal their categories' indexes, (6.33) is simplified to

$$\theta_{(k)}^{XY} = \exp\left(\varphi^{XY} + k\varphi^{XYZ}\right), \quad i = 1,\ldots,I-1,\ j = 1,\ldots,J-1,\ k = 1,\ldots,K.$$

Eliminating the three-factor interaction term in (6.32), the model of homogeneous uniform association (6.31) is derived in its equivalent expression

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \varphi^{XY}\alpha_i\nu_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}, \tag{6.34}$$

and $\theta^{XY} = \exp\left(\varphi^{XY}\Delta_1\Delta_2\right)$.

Replacing λ_{ik}^{XZ} and/or λ_{jk}^{YZ} in (6.32) or (6.34) by $\varphi^{XZ}\alpha_i\tau_k$ and/or $\varphi^{YZ}\nu_j\tau_k$, respectively, more parsimonious models of uniform or homogeneous uniform

association are derived that consider U-type structure also for one at least of the other two-factor interactions. The options for models of this type do not restrict to XZ and/or YZ interactions of U-type. They could be of any other type (R, C, RC, or RC(M)). Such special models of uniform and homogeneous uniform association cannot be captured via the odds ratio formulation (6.30) or (6.31).

Finally, the simplest homogeneous uniform XY association model is obtained when X is jointly independent from X and Y, i.e., the model

$$\log m_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \varphi^{XY} \propto_i \nu_j, \;\; i = 1, \ldots, I-1, \; j = 1, \ldots, J-1 \; ,$$

with $k = 1, \ldots, K$. For the example above, this model is the best option, with $G^2 = 18.165$ (p-value=0.314, $df = 16$), giving $\log \hat{\theta}^{XY} = 0.7688$.

6.8 Overview and Further Reading

Association models, in their dominant form, have been mainly developed by the fundamental and inspiring work of Goodman (1979b, 1981a, 1985, 1986, 1991, 1996) and thus it is common to refer to them also as Goodman's models. For an overview, we refer to the 1986 and 1991 discussion papers and the review Goodman (1982). Significant in the development of association models was continuation work of Haberman (1979, 1995), Clogg (1982a), Becker and Clogg (1988, 1989), and Becker (1989a, 1990a, 1992) as well as the contribution of Anderson and Philips (1981) and Anderson (1984). Association models are presented in the book by Clogg and Shihadeh (1994). An overview of association models with formulation and interpretation based on odds ratios is provided by Breen (2008) along with social sciences-orientated illustrations and references. For their connection to *latent class models*, see Sect. 10.3.1.

To be fair, we must say that the basis for the development of the association models lies back to Tukey's 1 d.f. test (Tukey 1949) and in a different form they have been considered earlier. In particular, Nelder and Wedderburn (1972) consider the U model applied on the popular Boys' Dream Disturbance data set of Maxwell as an illustration of their GLM model on contingency tables. Simon (1974) introduced the R model (his formulation A) as well as the R model for cumulative odds (his formulation B), being the forerunner of the association models for global odds ratios (see Sect. 7.1 below). A similar model for the cumulative odds was earlier considered by Williams and Williams and Grizzle (1972). Also his analysis of information is remarkable, throwing insight into the nature of departure from independence in the direction of the ANOAS, developed later by Goodman. Other multiplicative models modeling triangular or diagonal departures from independence for square tables have been proposed by Goodman (1972). We should mention that methods that analyze contingency tables with ordinal classification variables by applying scores to their categories have been proposed much earlier, even from Yates (1948) and

Armitage (1955), not to forget the linear trend test of Mantel (1963), described in Sect. 2.3. However, these early references, after assigning scores to the categories, treated the corresponding categorical variables applying methods appropriate for continuous variable analysis.

Haberman (1974b) adopted a different approach by generating a class of models through the decomposition of the vector with elements of the expected cell frequencies $\log(m_{ij})$ on an orthonormal basis, formed by orthogonal polynomials. Special members of this class of models are the linear-by-linear association model and the row effect model. Further, he proved standard asymptotic inference results for these models and expressed them in terms of the log odds ratio, noting the importance of the difference between scores. Finally, he was the first to mention that his approach could be extended straightforward to define such models for multi-way tables. Association models for two-way and three-way tables are presented in Wong (2010).

Diagnostics for the RC association model have been discussed in Andersen (1992). De Rooij and Heiser (2005) criticized the classical graphical representation of the RC(M) model and proposed the distance-association model representation, for which the distances between row and column points can be interpreted directly. Marginal association models have been considered by Lapp et al. (1998), Bartolucci et al. (2001), and Bartolucci and Forcina (2002). For ordinal tables with a response variable, Agresti (1986) proposed a regression R^2-type measure of association, based on scores assigned to the classification variables' categories, and used the R association model to estimate these scores. Baccini and Khoudraji (1992) and Baccini et al. (2000) considered least squares estimation of association models. Beh and Farver (2009) discuss on closed-form estimation of the association parameter φ of the U model.

The RC(M) are not the only models with additive multiplicative interaction terms. Goodman (1985) introduced other possible models with interaction of rank M but simpler than the RC(M). For example, for $M = 2$, the R+C model is defined by the same formula as RC(2) but assumes that the column scores of the first term and the row scores of the second are known; thus it has less parameters than RC(2). Similarly, model U+R+C has just one parameter more than the R+C model, since the third term that is added is of uniform type, having assigned fixed row and column scores. More options of parsimonious models of higher rank for the interaction are obtained through the use of orthogonal polynomials for assigning scores (Kateri et al. 1998). For example, the model $\mathrm{U}^{(1)}$+$\mathrm{U}^{(2)}$ is of $M = 2$, but all the involved scores are fixed, assigned through orthogonal polynomials of first and second order for the (1) and (2) term, respectively, and has thus just 2 parameters more than the independence model.

In the special case of a square table with commensurable classification variables, it makes sense to assume that the row and column scores are homogeneous. Thus, the RC model with the homogeneity restriction on its scores $\alpha_i = \nu_i$, $\forall\ i = 1,\ldots,I$, can be applied, which is more parsimonious than the standard RC and simultaneously of special interpretational value for such tables. On this we shall return and comment more on Chap. 9, specialized on square tables.

We have already mentioned in Chap. 2 that the log-linear models are the discrete analogue of the analysis of variance. It is interesting to note briefly at this point the analogues to association model in the two-way ANOVA framework. Special analysis of variance models that impose a structure on the interaction as that of the association models have been considered as well. Indicatively we mention the early work of Williams (1952), who used the multiplicative term for the interaction, and that by Gollob (1968), who introduced the more general term of the RC(M) type. Goodman and Haberman (1990) proved asymptotic normality for the scores of the RC(M) ANOVA model and provided asymptotic confidence intervals for the estimated scores. Furthermore, they developed the asymptotic conditional tests of the appropriateness of a simpler association model of the type U, R, or C given that the RC holds. Finally, they extended their results for the more general RC(M) ANOVA model. Viele and Srinivasan (2000) proceeded to the Bayesian analysis of the RC(M) ANOVA model. Speaking about analogies to the continuous case, Jones (1998) noted that the constant local dependence for continuous bivariate random variables is the continuous analogue of the U model.

We have seen in Sect. 6.3 that for ordinal contingency tables, conditional tests of independence given that the U, R, or C model holds, i.e., testing independence against a directed alternative, are more asymptotically powerful. Alternative approaches for strengthening the power of the classical Pearson's X^2 test of independence are based on the decomposition of Pearson's X^2 into orthogonal components in terms of assigned scores to the categories of the ordinal classification variables. For example, Best and Rayner (1996) and Rayner and Best (2000) considered scores based on orthogonal polynomials while Nair (1986, 1987), proportional to the midrank scores. Beh (1998) studied the use of different types of scores in the correspondence analysis framework. Nair's procedure partitions the X^2 statistic value for testing independence into location, dispersion, and residual effects. It is related to the location-dispersion model of McCullagh (1980), as is also pointed out by McCullagh's and Agresti's comments in the discussion of Nair (1986). Agresti's comment related Nair's statistics also to the statistics of Koch et al. (1982) with fixed or rank-based scores, to the measure of Agresti (1986), and to association models and the models in Semenya et al. (1983). Koshimizu and Tsujitani (1998) consider association models with location and dispersion scores for singly ordered contingency tables. Their model is actually analogue to the $R^{(1)}+R^{(2)}$ model of Kateri et al. (1998) with the column scores of the first dimension being the Nair's scores instead of equidistant for successive categories.

6.8.1 Multi-way Association Models

Conditional and partial associations in multi-way tables are discussed in Clogg (1982b). Becker (1989b) introduced the no three-factor interaction model with all two-way interaction terms replaced by the general terms (6.26). Becker and Clogg (1989) considered three-way association models for the analysis of stratified

two-way tables, with and without homogeneity constraints on the scores across the strata. Association models for stratified tables focusing on detecting layer differences were developed by Goodman and Hout (1998).

On the decomposition of the three-factor interaction term (6.27) focused Goodman (1983, 1986), Agresti and Kezouh (1983), Choulakian (1988), Anderson (1996), and Siciliano and Mooijaart (1997). A review is provided by Wong (2001). Methods of decomposing three-way arrays are reviewed in Ten Berge (2011).

6.8.2 Order-Restricted Inference

In case of association models with parametric scores, the monotonicity of the scores is not ensured by the standard estimation procedures. Since their monotonicity is related to stochastic ordering of the corresponding classification variable (Goodman 1981a), it is usually natural to expect the scores for ordinal classification variables to be monotonic. Estimation procedures subject to order constraints for the parametric scores have been developed for the R (or C) model by Agresti et al. (1987), based on isotonic regression. The RC model with order-restricted row and column scores has been considered by Ritov and Gilula (1991). A test of independence, conditional on the order-restricted RC model, is discussed in Kuriki (2005). Alternative algorithms for fitting the order-restricted RC model have been proposed and compared by Galindo-Garre and Vermunt (2004). Order restrictions yield also for an extended RC model, introduced by Bartolucci and Forcina (2002).

Ordinary or order-restricted inferences for these models rely on large-sample asymptotic methods. As it is stated in Galindo-Garre and Vermunt (2004), these methods do not work well for sparse tables or small sample sizes, common in social and biomedical applications, where the usual asymptotic chi-squared p-values are known to be inaccurate. A promising alternative is the Bayesian approach (see Sect. 10.5).

6.8.3 Comparison of Two Ordinal Responses

The problem of comparing two ordinal responses is very old and of special interest in many fields, especially in biomedical applications. The need to compare the response to a treatment of two independent groups of patients, defined, for example, by the presence of a prognostic factor, is obvious. Another common situation is to compare two different treatments applied on two independent samples with the corresponding responses measured on a common scale. The ordinality of the response scale has to be taken into consideration in handling the problem and answering to the question "Which group of patients benefits more from the treatment?" or "Which treatment is superior?." The underlying sampling scheme can be multinomial or product multinomial. The first is the case whenever a

sample of n subjects is cross-classified with respect to an ordinal response Y and a binary variable X indicating the two groups, while the second when two independent multinomial samples of the same ordinal response and of sizes n_1 and n_2 are available. If the ordinal response has J categories, then the above described data form a $2 \times J$ contingency table. For the multinomial sampling scheme the corresponding joint distribution is $\pi = (\pi_{ij}) = P(X = i,\ Y = j),\ i = 1,2,\quad j = 1,\ldots,J$. In case of two independent multinomials, the row marginals are also fixed, $n_{i+} = \sum_{j=1}^{J} n_{ij} = n_i\ (i = 1,2)$.

The problem of comparing two response profiles is equivalent to the stochastic comparison of the two row distributions of the abovementioned $2 \times J$ contingency table and as such has been faced by a variety of methods. The related bibliography is very rich and an extended critical review of the available methods can be found in Agresti and Coull (2002).

The hypothesis that two multinomial distributions are identical against an ordered alternative is mainly tested through LR, Wald, and score tests or through linear rank tests. It is well known that restricting the alternative hypothesis leads to more powerful tests than the standard chi-squared test of independence. These approaches are all asymptotic, while the linear rank tests depend on the choice of the scores assigned to the ordered categories. Characteristic references of LR tests are Grove (1980, 1984) and Robertson and Wright (1981), while the approaches of Emerson and Moses (1985), Graubard and Korn (1987), and Gautam (1997) are based on linear rank tests. To deal with the sensitivity of the linear rank tests on the scores, Kimeldorf et al. (1992) proposed the min–max scoring and Gautam et al. (2001) the iso-chi-square approach. Nonlinear rank tests have also been proposed. For example, Hilton et al. (1994) and Nikiforov (1994) applied the Smirnov test while Berger (1998) proposed the convex hull methodology that leads to admissible tests. Properties and power of the convex hull test applied on $2 \times J$ tables are further studied in Berger et al. (1998) (see also Cohen and Sackrowitz 1998; Cohen et al. 2000). An interesting approach is provided by Permutt and Berger (2000), who reviewed various rank tests, classified them as Smirnov-like or Wilcoxon-like, and compared them. However, the nonlinear tests are not easy to compute for $J > 3$. It is important to note that "with few exceptions there is no optimal test for this problem," as stated by Berger and Ivanova (2002). Tests based on log-linear models were developed by Agresti and Coull (1998).

The connection of association models to the stochastic ordering of the conditional row (or column) distributions of the contingency table has been discussed in Sect. 6.4. In case of the $2 \times J$ tables, the RC model coincides with the C model, which is saturated. For $2 \times J$ contingency tables with $\alpha_1 < \alpha_2$ and $\varphi > 0$, positive dependence is equivalent to $\nu_j \leqslant \nu_{j+1}\ (j = 1,\ldots,J-1)$ with $\nu_1 < \nu_J$. Thus, *monotonicity* of the column scores $\{\nu_j : j = 1,\ldots,J\}$ implies *stochastic ordering* of the probabilities

$$\pi_i = \left\{ \frac{\pi_{ij}}{\pi_{i+}}, j = 1,\ldots,J \right\},\ i = 1,2.$$

Thus the distribution of the response Y for the second group (row) is stochastically larger than the one of the first group (row).

The comparison of the two row distributions can be further enriched with the option of umbrella ordering as an alternative when stochastic ordering is rejected (Kateri 2011). Umbrella ordering means that the distribution in the first row is stochastically smaller than the one in the second up to a level of the ordinal scale that defines the column categories and stochastically larger after this level (or the opposite). In terms of physical interpretation, when comparing two alternative treatments, umbrella ordering of their response distributions corresponds to cases where one treatment is better over the other up to a certain level of the response scale while the situation changes after this point. In a retrospective study context cross-classifying the "cured"– "not-cured" groups with the J levels of a prognostic factor, this could mean higher risk for the very low and very high levels of the prognostic factor. Umbrella ordering essentially reveals a dispersion effect for the group comparison. Dispersion effects for ordinal responses have been handled by the generalized cumulative link model, introduced by McCullagh (1980). Umbrella ordering can be captured by the C model, with adequately constrained column scores.

6.8.4 Cell Frequencies vs. Local Odds Ratios Modeling

We have seen so far that models applied on contingency tables can be expressed in terms of expected cell frequencies or equivalently in terms of local odds ratios. The choice depends on issues of interpretation and on convenience of model formulation. For example, the association models are easier interpreted through the local odds ratios. On the other hand, the quasi-independence model can be expressed in terms of local odds ratios but is too complicated to compete with (5.24).

A clarifying and inspiring insight into the possible different views of log-linear models is provided by Goodman (1981d), who considers three alternative views, depending on the purpose of the analysis. The model imposed on the cell frequencies is preferred whenever the purpose is the examination of the joint distribution of the contingency table. Local odds ratio formulation of the model is employed when interest lies on the association between the two variables that are cross-classified. In both cases, the classification variables of the table are treated symmetrically. If there exists a response variable, then modeling the possible dependence of the response variable on the explanatory one is more adequate than the symmetric approaches and leads to more direct interpretations. This constitutes the third view and corresponds to modeling the odds for the response variable, given that the explanatory variable is at a fixed, prespecified level. Such models are presented in Chap. 8. Goodman (1981d) discussed the connections between these different approaches of log-linear modeling and illustrated them on characteristic examples. These comments apply also to the special models for square tables in Chap. 9.

Chapter 7
More on Association Models and Related Methods

Abstract Advanced issues on association models are discussed in this chapter. These include exploring the rows and/or columns heterogeneity in a contingency table, the issue of merging categories of a classification variable, and the consideration of association models for generalized odds ratios other than the local odds ratios. The uniform association model for the global odds ratios is illustrated with an example in R. Correspondence analysis (CA) is also presented and connected to association models. For comparison purposes, CA is applied in R on one of the examples analyzed in Chap. 6 by association models.

Keywords Association models for the global odds ratios • Correspondence analysis • Generalized association models • Merging categories

7.1 Association Models for Global Odds Ratios

The association models considered so far were defined on the local odds ratios and interpreted in terms of the local associations of the $(I-1)(J-1)$ odds ratios defined for successive row and column categories. If interest lies on modeling global association in a contingency table, association models can be defined analogously in terms of the global odds ratios (defined in Sect. 2.2.5). For example, we have seen that the U model defined on the local odds ratios assumes constant local odds ratios across the table. Similarly, the uniform global odds ratios model (U^G) is

$$\theta_{ij}^G = \theta^G, \qquad i = 1,\ldots,I-1; \;\; j = 1,\ldots,J-1 \tag{7.1}$$

and assumes a common value θ^G for all global odds ratios of the table. Plackett (1965) was the first who modeled global odds ratios by defining their distribution in terms of their common value and the marginals of the table. Analogously, the R, C, and RC models expressed in terms of local odds ratios by (6.8), (6.9), and (6.10),

M. Kateri, *Contingency Table Analysis: Methods and Implementation Using R*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4811-4_7,

respectively, can be defined for the global odds ratios as well, just by replacing the local odds ratios by the corresponding global ones. Closely related is the follow-up work by Wahrendorf (1980).

The choice between modeling local or global associations is basically interpretation driven. If interest lies in comparing two distinct rows or columns, then local association models are appropriate. On the other hand, models for global associations have to be used if conclusions for dichotomized versions of the classification variables' scales make better sense. Dale (1984) compared local and global associations for $I \times J$ tables with both marginal probabilities ordered, highlighted situations where the one should be preferred over the other, and concluded that they have "complementary roles." Global associations are more stable with respect to category boundaries; thus, they are more appropriate when classification variables have an underlying continuum while local association when the categories are well defined. The U model for the local odds ratios provides a better discretization to the bivariate normal than U for global odds ratios (Goodman 1981c; Dale 1984).

A class of models for bivariate ordinal responses based on global odds ratios that condition over a set of covariates is introduced by Dale (1986). This general model allows for many parameterizations of the marginal cumulative probabilities, the generalized linear model of McCullagh (1980) included. It could be viewed as the precursor of Lang and Agresti's GLLM model (5.28).

The most characteristic model of this type is the constant global odds ratio model, i.e., the U model for the global odds ratios. Before Dale, it has been considered by Pearson and Heron (1913), Plackett (1965), and Mardia (1970) for a bivariate response with specified marginal distributions, discrete or continuous. Models of constant global odds ratios have also been discussed by Molenberghs and Lesaffre (1994) and Heagerty and Zeger (1996) for correlated ordinal data (see Sect. 9.7.4). The continuous analogue of the constant global odds ratio model has been introduced by Clayton (1978) for modeling association in bivariate life tables. Semiparametric global odds ratio models for bivariate censored survival times are discussed in Ghosh (2006).

7.1.1 The U^G Model in R: Example 6.1

For our cannabis example, the sample global odds ratios are provided in Table 7.1 and it is obvious that the U model for the global odds ratios will not be as good as the U model on the local odds ratios. Indeed, the LR statistic for this model is $G^2 = 6.029$ (p-value=0.307) on 5 df, while for the same model considered for the local odds ratios, we derived $G^2 = 1.469$ (p-value=0.917) on the same df (see Table 6.4). This indicates that the underlying association structure in this data table is constant rather locally than globally.

This model can be considered as a special case of the generalized log-linear model (GLLM) (5.28) and implemented in R by Lang's `mph` package. The GLLM becomes the U model on global odds ratios by appropriately choosing matrices $\mathbf{M}$ and $\mathbf{C}$. The $I \times J$ data table is expanded in rows and the vector of expected cell

Table 7.1 Sample global odds ratios $\hat{\theta}_{ij}^{G}$ $(i = 1,2,3; j = 1,2)$, for the students' survey about cannabis use at the University of Ioannina, Greece (1995)

	I tried cannabis...		
Alcohol consumption	Never	Once or twice	More often
At most once/month	8.08	25.73	
Twice/month	6.10	11.94	
Twice/week	6.51	7.38	
More/often			

The ML estimates of the common global odds ratio value under the U model for global odds ratios is $\hat{\theta}^{G} = e^{1.8622} = 6.44$

frequencies $\mathbf{m}$ is of size $IJ \times 1$. Matrix $\mathbf{M}$ is applied on $\mathbf{m}$ to form the sums of cell entries needed for the derivation of the global odds ratios. Thus, it is a table of 1's and 0's. Since an $I \times J$ table forms $(I-1)(J-1)$ global odds ratios and each of them needs 4 sums of cell entries, $\mathbf{M}$ is of size $4(I-1)(J-1) \times IJ$. Finally, $\mathbf{C}$ is defined so that it combines the entries of $\log(\mathbf{M} \cdot \mathbf{m})$ to conclude to the $(I-1)(J-1) \times 1$ vector of all expected global odds ratios $\mathbf{C}\log(\mathbf{M} \cdot \mathbf{m})$. A function that produces the required matrices $\mathbf{M}$ and $\mathbf{C}$ for any $I \times J$ table is provided in the web appendix mentioned in Sect. A.3.2 (function `global.odds.DM()`).

Once the global odds ratios are formed, the rest of the procedure is analogue to the one described for the local odds ratios in Sect. 6.6.4. Thus, for the U model the data and the design matrix are defined as in Sect. 6.6.4. Matrices $\mathbf{M}$ and $\mathbf{C}$ are constructed by

```
> M <- global.odds.DM(NI,NJ)$M
> C <- global.odds.DM(NI,NJ)$C
```

while the link function is specified as

```
> L.fct <- function(m){C%*%log(M%*%m)}
```

Finally,

```
> mph.out <- mph.fit(y=y,L.fct=L.fct,X=X)
> mph.summary(mph.out,cell.stats=T,model.info=T)
```

fits the model and derives the output, as described in Sect. 5.6. The common $\log \theta^{G}$ is estimated by $\hat{\beta} = 1.8622$ and thus $\hat{\theta}^{G} = e^{1.8622} = 6.44$.

If sampling zeros are present and cause a convergence problem, they are treated as in Sect. 6.6.4.

7.2 Correspondence Analysis

A popular method for detecting the pattern of association between the row and column categories of a two-way contingency table is correspondence analysis (CA), which is mainly a descriptive method. CA assigns "optimal" scores to the row and column categories and plots these scores as points in the Euclidean two- or three-dimensional space, providing thus a reduced rank display. "Optimal" is considered in the sense that the reduced rank expression explains the maximum possible

percentage of variation in the data. The relative positions of the points indicate the underlying association between rows (or columns) as well as between rows and columns. CA for two-way tables is known as *simple CA* while for multi-way tables as *multiple CA* (see Sect. 7.6.1).

7.2.1 Simple Correspondence Analysis in Steps

CA is the discrete analogue of principal component analysis (PCA). PCA partitions the total variance while CA partitions the Pearson's X^2 value for testing independence on an $I \times J$ contingency table (n_{ij}) of total sample size $n = \sum_{i,j} n_{ij}$. The partitioning is achieved through the singular value decomposition (SVD) of the matrix $\mathbf{S}$ with elements $s_{ij} = \frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}}$, which correspond to the Pearsonian residuals, since $X^2 = n\sum_{i,j}(s_{ij})^2$. The dimension of the Euclidean space of the saturated model's decomposition is $M^* = \min(I,J) - 1$. The goal is the detection of a subspace of order $M < M^*$ which includes the row and column points in the best possible way. If suitable, the choice $M = 2$ is ideal for the visualization advantage of the points on the two-dimensional space.

The CA procedure can briefly be described in the following steps:

1. Calculation of the SVD of matrix $\mathbf{S}$:
   $\mathbf{S} = \mathbf{U}\Lambda\mathbf{V}'$, where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{M^*})$ is the diagonal matrix of the corresponding eigenvalues $(\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{M^*} \geq 0)$ while the rows of tables $\mathbf{U}_{I\times M^*}$ and $\mathbf{V}_{J\times M^*}$ are the left and right eigenvectors, respectively, with $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}$.
2. Row and column *masses*, *profiles*, and *centroids*:
   The row and column *masses*, *profiles*, and *centroids* are key quantities in CA application and interpretation. In the CA framework, row (or column) masses are called the row (or column) marginal probabilities of the table and they are used as weights associated to the rows (or columns) of the contingency table. Row profiles are called the conditional row probabilities while the column profiles are defined analogously. The rows centroid (or barycenter) is considered to be the average row profile, while the columns centroid is defined analogously. Let $m_r(i) = p_{i+}$ and $m_c(j) = p_{+j}$ be the mass of row i and column j, respectively. Let also $r_{ij} = \frac{p_{ij}}{p_{i+}}$ and $c_{ij} = \frac{p_{ij}}{p_{+j}}$ be the row and column profiles, respectively. Then, the rows centroid is $\bar{\mathbf{r}} = (\bar{r}_1, \ldots, \bar{r}_J)$ with coordinates $\bar{r}_j = \sum_i p_{i+}r_{ij} = p_{+j} = m_c(j)$, $j = 1,\ldots,J$. Analogously, the column centroid is $\bar{\mathbf{c}} = (\bar{c}_1, \ldots, \bar{c}_I)$ with $\bar{c}_i = \sum_j p_{+j}c_{ij} = p_{i+} = m_r(i)$. Thus, the X^2 statistic can be reexpressed as

$$X^2 = n\sum_i p_{i+}\left\{\sum_j \frac{(r_{ij} - \bar{r}_j)^2}{\bar{r}_j}\right\} = n\sum_j p_{+j}\left\{\sum_i \frac{(c_{ij} - \bar{c}_i)^2}{\bar{c}_i}\right\}$$

   and can be interpreted as the weighted average distance of the row profiles (or column profiles) from their centroid.

3. Construction of the *standard coordinates* for rows and columns:
   If $\mathbf{D}_r = \text{diag}(p_{1+}, \ldots, p_{I+})$ and $\mathbf{D}_c = \text{diag}(p_{+1}, \ldots, p_{+J})$ are the diagonal matrices of the row and column masses respectively, then standard coordinates of the rows are

$$\mathbf{X} = \mathbf{D}_r^{-1/2}\mathbf{U}\,,$$

   while of the columns

$$\mathbf{Y} = \mathbf{D}_c^{-1/2}\mathbf{V}\,.$$

   The standard coordinates are weighted orthonormalized, i.e., they satisfy the constraints

$$\mathbf{X}'\mathbf{D}_r\mathbf{X} = \mathbf{Y}'\mathbf{D}_c\mathbf{Y} = \mathbf{I}\,. \tag{7.2}$$

4. Calculation of *inertias*:
   The eigenvalue λ_m ($m = 1, \ldots, M^*$) is a measure of correlation between the vector of scores x_m and y_m, since

$$\lambda_m = \sum_{i,j} p_{ij} x_{im} y_{jm}\,, \quad m = 1, \ldots, M^*, \tag{7.3}$$

   due to (7.2). It can be further shown that

$$X^2 = n \sum_{i,j} (s_{ij})^2 = n \sum_{m=1}^{M^*} \lambda_m^2\,, \tag{7.4}$$

   i.e., the X^2 goodness-of-fit test for independence is partitioned in M^* terms, each of which expresses the part of X^2 explained by the corresponding dimension. The quantities λ_m^2 are called *inertias* and their sum $\sum_{m=1}^{M^*} \lambda_m^2$, known as the *total inertia*, is indicative about the variability in the data, independent of sample size.
5. Construction of the *principal coordinates* for rows and columns:
   The principal coordinates of rows are

$$\tilde{\mathbf{X}} = \mathbf{D}_r^{-1/2}\mathbf{U}\Lambda,$$

   while for columns

$$\tilde{\mathbf{Y}} = \mathbf{D}_c^{-1/2}\mathbf{V}\Lambda\,.$$

   It can be verified by standard matrix calculations that

$$\tilde{\mathbf{X}}'\mathbf{D}_r\tilde{\mathbf{X}} = \tilde{\mathbf{Y}}'\mathbf{D}_c\tilde{\mathbf{Y}} = \Lambda^2,$$

i.e., the weighted sum of squares of the principal row (or column) coordinates on the mth dimension ($1 \leq m \leq M^*$) equals the corresponding inertia. Due to this property, the row and column principal coordinates are crucial quantities in CA and are used for physical interpretation.

6. Graphical representation:
   Whenever feasible, the use of $M = 2$ is engaged so that the corresponding vectors can be represented graphically in the two-dimensional space and easier to visualize and interpret. This is the best bivariate representation of the data.

   The graphical representation of CA offers an overview for the underlying association structure. For example, monotonicity of the association, rows (or columns) not differentiating in one dimension or in total, and row–column combinations of greater probability can be detected. For a set of row points (or column points), the Euclidean distance in the two-dimensional graph corresponds to a statistical distance between pairs of row (or column) profiles. However, there is no direct distance relation between a row and a column point.
7. The role of a specific category:
   Specific conclusions for each category of the row classification variable can be drawn by observing its total and relative contribution to the inertia of the mth axis, defined as $TCT_m^X(i) = p_{i+}\tilde{x}_{im}^2$ and $RCT_m^X(i) = \frac{\tilde{x}_{im}^2}{\sum_m \tilde{x}_{im}^2}$, respectively. The contribution of the ith category to the mth axis is of special interest if $TCT_m^X(i) >> p_{i+}$. Analogously are defined the total and relative contributions $TCT_m^Y(j)$ and $RCT_m^Y(j)$ of the jth column of the table. The definition of the total contributions is justified by the relation $\lambda_m^2 = \sum_i TCT_m^X(i) = \sum_j TCT_m^Y(j)$.

Traditionally, CA is based on the principal scores and the CA model is expressed as

$$p_{ij} = p_{i+}p_{+j}\left(1 + \sum_{m=1}^{M^*} \frac{\tilde{x}_{im}\tilde{y}_{jm}}{\lambda_m}\right), \quad i = 1\ldots,I, \ j = 1,\ldots,J\ .$$

This is actually a reparameterized version of the canonical correlation model

$$p_{ij} = p_{i+}p_{+j}\left(1 + \sum_{m=1}^{M^*} \lambda_m x_{im} y_{jm}\right), \quad i = 1\ldots,I, \ j = 1,\ldots,J\ , \tag{7.5}$$

defined in terms of the standard scores and founded by Fisher (1940), since $\tilde{x}_{im} = \lambda_m x_{im}$ and $\tilde{y}_{jm} = \lambda_m y_{jm}$. This model is saturated and will be denoted by CA(M^*). Replacing M^* with $M < M^*$ in (7.5), more parsimonious models are achieved, namely the CA(M),

$$\pi_{ij} = \pi_{i+}\pi_{+j}\left(1 + \sum_{m=1}^{M} \lambda_m x_{im} y_{jm}\right), \quad i = 1\ldots,I, \ j = 1,\ldots,J\ , \tag{7.6}$$

where π_{ij} is the expected frequency in cell (i, j) under CA(M). The row and column marginal probabilities are equal to the corresponding observed, i.e., $\pi_{i+} = p_{i+}$ ($i = 1,\ldots,I$) and $\pi_{+j} = p_{+j}$ ($j = 1,\ldots,J$). How well the deviation from independence is captured by the subspace of order M is judged by comparing the size of the M largest inertias to that of the rest of them. For this, the index used for the quality of the approximation of model CA(M) to the table of observed sample proportions is the popular percentage

$$\frac{\sum_{m=1}^{M} \lambda_m^2}{\sum_{m=1}^{M^*} \lambda_m^2} .$$

7.2.2 Correspondence Analysis of Example 6.2

We shall illustrate the fit and interpretation of CA on the example of Wermuth and Cox (1998), presented in Sect. 6.5.2 and already analyzed by association models. The fit of CA in practice is straightforward in most statistical or mathematical packages. For example, in SPSS, it can be fitted through

Analyze > Data Reduction > Correspondence Analysis .

To perform CA in `Mathematica`, see Yelland (2010). In `R`, the function `corresp()` of the package `MASS` is appropriate, but we will demonstrate special packages developed for correspondence analysis, the `ca` of Nenadić and Greenacre (2007) and `anacor` of de Leeuw and Mair (2009). The `ca` package fits CA to multi-way tables as well (see Sect. 7.6.1). On the other hand, `anacor` is restricted to two-way contingency tables but provides more options for plots and is applicable also for incomplete tables.

Both packages need the data to be provided in a frequency table format. Thus, for the Wermuth and Cox example, if the data were written in a file named WCox.txt, in a simple matrix form without header

12	13	12	20	7
215	507	493	460	137
277	300	192	126	38
52	91	47	15	6
233	225	102	74	19

then they are read in `R` by the command:

```
> WCox.data <- read.table(file="c:// ... //WCox.txt",header=F)
```

Correspondence analysis in `ca` is performed by the command

```
> ca(WCox.data)
```

which gives the following output

```
 Principal inertias (eigenvalues):
            1          2          3          4
Value       0.088408   0.005845   0.000485   0.000132
Percentage  93.19%     6.16%      0.51%      0.14%

 Rows:
                1          2          3          4          5
Mass      0.017424   0.493330   0.254016   0.057446   0.177784
ChiDist   0.421986   0.286497   0.233038   0.374979   0.406673
Inertia   0.003103   0.040493   0.013795   0.008077   0.029402
Dim. 1   -0.905634  -0.963207   0.776072   0.924481   1.353983
Dim. 2    4.037920  -0.089332   0.166141  -3.293206   0.678866

 Columns:
                1          2          3          4          5
Mass      0.214811   0.309284   0.230329   0.189219   0.056357
ChiDist   0.463855   0.134855   0.195366   0.369847   0.384979
Inertia   0.046219   0.005625   0.008791   0.025883   0.008353
Dim. 1    1.536961   0.359641  -0.616852  -1.206601  -1.259753
Dim. 2    1.031116  -1.035024  -0.761026   1.156251   0.978118
```

The coordinates of the two most important dimensions are reported in this output. To obtain full flexibility in reporting and graphing results, we may want to retrieve the full row and column eigenvector coordinates. This is achieved by the commands

```
> x <- ca( WermCox.data)$rowcoord
```

and

```
> y <- ca( WermCox.data)$colcoord
```

for the row and column coordinates, respectively. For the Wermuth and Cox example, these are

```
> x
```

```
          [,1]         [,2]          [,3]         [,4]
[1,] -0.9056343   4.03791967   3.7885808423 -4.9912234
[2,] -0.9632069  -0.08933166  -0.0009268877  0.3021479
[3,]  0.7760720   0.16614083  -1.2969234500 -0.7904822
[4,]  0.9244814  -3.29320644   1.7135868247 -1.3309127
[5,]  1.3539829   0.67886567   0.9305866854  1.2102433
```

and

```
> y
```

```
          [,1]         [,2]         [,3]         [,4]
[1,]  1.5369608   1.0311159  -0.4633144  -0.1230903
[2,]  0.3596406  -1.0350242   1.0001602   0.1798274
[3,] -0.6168523  -0.7610259  -1.5320267  -0.1866404
[4,] -1.2066013   1.1562510   0.4246632   1.1453166
[5,] -1.2597529   0.9781184   1.1126896  -3.6003042
```

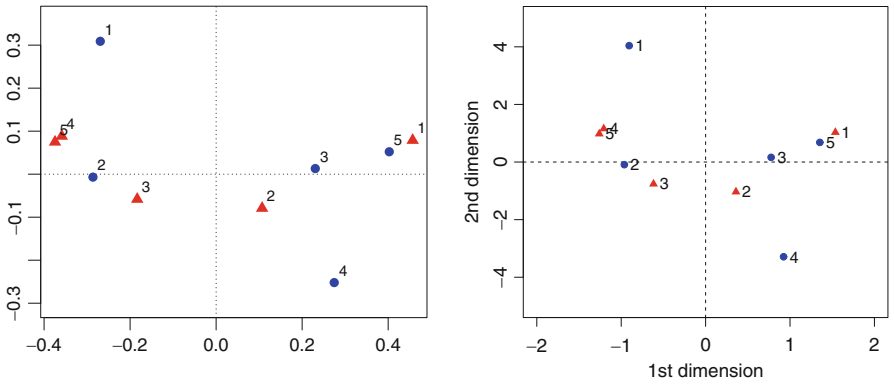


Fig. 7.1 Two-dimensional CA map of the data in Table 6.7, with row (*bullets*) and column (*triangles*) in (i) principal coordinates (*left*) and (ii) standard coordinates (*right*)

The 2-dimensional plots of the row and column coordinates, for which CA is famous, are produced by

```
> plot(ca(WCox.data))
```

This command treats by default the rows and columns symmetric and produces the plot of the principal coordinates for rows and columns. For our example, this plot is provided in Fig. 7.1 (left). The type of coordinates plotted is controlled by the `map` scaling option. For example,

```
> plot(ca(WCox.data),map="rowprincipal")
```

uses for the plot the principal coordinates for the rows and the standard for the columns. A description of possible scale options for the CA plot function is provided in Nenadić and Greenacre (2007).

The coordinates that are comparable to the scores of the RC(2) association model are the standard coordinates, which, for our example, are the first two dimensions of `x` and `y` listed above. The plot of the standard coordinates for both, rows and columns, is not an option in `map`. However, it can be easily obtained through the standard `plot()` command, applied on `x` and `y`. This plot can be produced by function `plot_2dim()` of the web appendix (see Sect. A.3.5).

Thus, the plot in Fig. 5.2 (right) is obtained by calling this function as

```
> plot_2dim(x, y, -2, 2, -5, 5, -0.7, 1.4, 1.2)
```

The arguments of this function, beyond `x` and `y`, control the plot appearance and their use is explained in Sect. 6.6.2. Note the similarity of Fig. 5.2 (right) to Fig. 5.1 with marginal weights.

7.3 Correlation Models

The analogies between CA and the RC(M) model made the comparison unavoidable and the two approaches were developed competitively, borrowing ideas from each other. Thus, graphs were constructed for the scores of the RC(M) model in the spirit of CA while the inferential aspects of the CA model were developed accordingly to RC(M). On the other hand, the estimates applied by classical CA are traditionally least square estimates. The consideration of maximum likelihood estimation for the parameters of the CA(M) model (7.6) led to the development of the RC correlation model of order M (Goodman 1985). MLEs for (7.6) have been considered by Gilula (1986). The scores are subject to constraints (7.2) in terms of matrix notation, which are expressed on the elements as

$$\sum_{i=1}^{I} \pi_{i+} x_{im} = \sum_{j=1}^{J} \pi_{+j} y_{jm} = 0, \qquad m = 1, \ldots, M,$$

$$\sum_{i=1}^{I} \pi_{i+} x_{im} x_{i\ell} = \sum_{j=1}^{J} \pi_{+j} y_{jm} y_{j\ell} = \delta_{m\ell}, \quad m, \ell = 1, \ldots, M.$$

In the special case of $M = 1$, (7.6) reduces to the CA(1) model

$$\pi_{ij} = \pi_{i+}\pi_{+j}(1 + \lambda x_i y_j), \quad i = 1, \ldots, I, \; j = 1, \ldots, J, \tag{7.7}$$

by eliminating the subscript m, while the constraints on the scores are adjusted accordingly to

$$\sum_i \pi_{i+} x_i = \sum_j \pi_{+j} y_j = 0 \quad \text{and} \quad \sum_i \pi_{i+} x_i^2 = \pi_{+j} y_j^2 = 1. \tag{7.8}$$

Following exactly the same consideration and arguments as in the association models framework, for $M = 1$ we could consider that the row or the column scores or both sets of them are known, usually equidistant for successive categories. Then (7.7) turns out to be the column (C_c), row (R_c), or uniform (U_c) correlation model.

The correlation models have the oddity to express the expected cell frequencies in terms of their row and column marginals. This is feasible only when the specific (marginal weighted) constraints are applied on the row and column scores of the model. Furthermore, this means that the main effects of the classification variables are captured only through the corresponding marginals. This lack of flexibility for the main effects has a consequence in the estimation of the special U_c, R_c, and C_c correlation models. Whenever a set of scores is fixed, the related marginals of the ML estimates of the expected probability table are no more equal to the corresponding observed marginals of the sample proportions. This means that

$$\hat{\pi}_{i+} = p_{i+}, \quad i = 1, \ldots, I, \tag{7.9}$$

do not hold for the R_c model and

$$\hat{\pi}_{+j} = p_{+j}, \quad j = 1, \ldots, J, \tag{7.10}$$

do not hold for the C_c model, while none of them holds for the U_c model. Due to the importance of constraints (7.9) and (7.10) and in order for these correlation models to be consistent with association and classical log-linear models, Goodman (1985) suggested also constrained maximum likelihood estimation.

Furthermore, independence (I) could be tested conditional on the U_c, R_c, or C_c models, by the $I|U_c$, $I|R_c$, or $I|C_c$ conditional tests, as in the association models framework (see Sect. 6.3). Also the R_c or the C_c models could be tested conditional on the RC_c, leading thus to an analysis of correlation, in the spirit of the analysis of association (ANOAS) procedure.

Thus, the test statistic $G^2(I|U_c) = G^2(I) - G^2(U_c)$ serves as a 1 d.f. conditional test of independence for a contingency table, provided that the U_c model holds, by testing the null hypothesis $\lambda = 0$. Under I, $G^2(I|U_c)$ is asymptotically $\mathcal{X}_1^2$ distributed. Under the U_c model, (7.3) leads to

$$\hat{\lambda} = \sum_{i,j} \hat{\pi}_{ij} x_i y_j \ ,$$

where $\hat{\pi}_{ij}$ are the constraint MLEs of the cell probabilities under U_c. It is worth to highlight the connection of this test to the linear trend test (2.57), which is also a 1 d.f. test of independence, restricted to the case of linear correlation between the row and column scores (see Sect. 2.3). For scores subject to the marginal constraints (7.8), the correlation ρ is estimated by the sample correlation

$$r = \sum_{i,j} p_{ij} x_i y_j \ .$$

If U_c expresses the association structure of the table, the π_{ij}'s should be close to the p_{ij}'s. Hence, as $\hat{\pi}_{ij}$'s under U_c approach the sampling proportions p_{ij}'s, $\hat{\lambda}$ approaches r and these two tests become similar.

7.4 Generalized Association Models

In the 1980s, association and CA models were developed competitively to each other, until their connection was pointed out in the pioneer paper of Gilula et al. (1988). They stated that under certain conditions, both of them are the closest model to independence. Their difference is on the scale measuring the closeness to independence. Association models are the closest in terms of the *Kullback–Leibler divergence* while correlation models in terms of the *Pearson's divergence*. This gave ground to the development of general classes of dependence models by modeling the departure from independence in terms of generalized measures.

An important generalized measure is the ϕ-divergence (see Sect. 4.9.5). If $\pi = (\pi_{ij})$ and $\mathbf{q} = (q_{ij})$ are two discrete finite bivariate probability distributions, then the ϕ-divergence between π and $\mathbf{q}$, given in (4.38), takes the form

$$\mathrm{I}^{\mathrm{C}}(\pi, \mathbf{q}) = \sum_{i,j} q_{ij} \phi(\pi_{ij}/q_{ij}) \; . \tag{7.11}$$

Under the conditions of Gilula et al. (1988), the closest model to independence, in terms of the ϕ-divergence, is the

$$\pi_{ij} = \pi_{i+}\pi_{+j} F^{-1}\left(\alpha_i + \beta_j + \sum_{m=1}^{M} \varphi_m \propto_{im} \nu_{jm}\right), \quad i = 1, \ldots, I, \; j = 1, \ldots, J \; , \tag{7.12}$$

where F^{-1} is the inverse function of $F(x) = \phi'(x)$ and $1 \leq M \leq M^*$ (Kateri and Papaioannou 1995). This is the ϕ-divergence association model of order M, denoted by $\mathrm{RC}_\phi(M)$, and its scores $\propto_m$ and ν_m satisfy the restrictions (6.17).

The $\mathrm{RC}_\phi(M)$ model defines a family of models. The standard RC(M) model (6.18) is a member of this family and is derived for $\phi(x) = x \log x$, expressed in terms of cell probabilities. For $\phi(x) = (1-x)^2$, setting $\propto_{im} = x_{im}$ $(i = 1, \ldots, I)$, $\nu_{jm} = y_{jm}$ $(j = 1, \ldots, J)$, and $\varphi_m = \lambda_m$ $(m = 1, \ldots, M)$, model (7.12) reduces to the correlation model (7.6). The association models proposed by Rom and Sarkar (1992) correspond to the power divergence of Cressie–Read (4.37) and are included in the $\mathrm{RC}_\phi(M)$ family.

It is interesting to note that the idea of viewing a model as a departure from a parsimonious reference model with the property of being the closest to this reference model under certain conditions in terms of the Kullback–Leibler divergence can be extended to other types of models as well, such as the logistic regression (see Sect. 8.4) and the quasi-symmetry model (see Sect. 9.7). This leads to the generation of classes of generalized models, based on ϕ-divergence, which includes these known models as special cases.

In terms of measures of dependence, the concept of ϕ-divergence was applied by Joe (1989). He proposed measures of multivariate or conditional dependence based on relative entropies, studied their properties, found connections to classical measures of association, and extended his results for the class of the ϕ-divergence measures.

7.5 The Role of Scores in Merging Categories

The issue of merging categories and thus reducing the size of a contingency table is almost as old as contingency table analysis itself and drew the interest even in the very early literature on contingency tables (Yates 1948). At this point we have to distinguish between motives and criteria of merging and where they meet. The basic

motive to consider merging of classification categories has to do with sparseness and the desire to avoid small cell entries. Thus, we tend to merge categories of small entries, usually at the edges of the classification scale. Motivation could also be the observation that the merged table can be described by a simpler and of easier interpretation model. But when is it right to proceed to a merging? The oldest and most popular criterion for merging is *homogeneity* of the corresponding rows (or columns) (Benzécri 1973; Hirotsu 1983; Gilula 1986; Gilula and Krieger 1989; Weller and Romney 1990). Another basic criterion is that of *structure* (Goodman 1981c, 1985; Wermuth and Cox 1998). Greenacre (1988a) clusters homogeneous rows or columns, adopting the procedure by Hirotsu (1983) and decomposing the X^2 statistic of independence with respect to the nodes of the binary tree associated with either the row or column clustering, displaying thus graphically the heterogeneity. In case of ordinal classification variable, *order violation* of certain estimated parametric scores of an association model for a classification variable is a reason to merge the corresponding categories in order to ensure the known order (Goodman 1985, 1986; Agresti et al. 1987; Ritov and Gilula 1991, 1993).

The rows r and s of an $I \times J$ contingency table $\Pi = (\pi_{ij})$ are set to be *homogeneous* if the corresponding conditional column probabilities are equal, e.g., if

$$\frac{\pi_{rj}}{\pi_{r+}} = \frac{\pi_{sj}}{\pi_{s+}}, \qquad \forall\ j = 1, \ldots, J\ .$$

The definition is straightforwardly extended to more than two rows while homogeneous columns are defined analogously.

The basic property that characterizes homogeneous rows or columns is that independence holds for every subtable formed from homogeneous rows or columns. It was thus natural that homogeneity was initially related to the basic model of independence. Consequently, if we denote by $\mathbf{I}$ and $\tilde{\mathbf{I}}$ the models of independence for the initial $I \times J$ and the merged $\tilde{I} \times \tilde{J}$ tables, respectively ($\tilde{I} \leqslant I, \tilde{J} \leqslant J$), then, provided the merging has been done among homogeneous rows and columns, the difference of the LR statistics for the fit of models $\mathbf{I}$ and $\tilde{\mathbf{I}}$, $G^2(\mathbf{I}) - G^2(\tilde{\mathbf{I}})$, should not be statistically significant (see Williams 1952 and Goodman 1985). Following similar arguments and in correspondence analysis framework, Benzécri (1973) introduced the principle of distributional equivalence. In this setup, homogeneity is expressed as equality of the corresponding row profiles. Recall that $\frac{\pi_{rj}}{\pi_{r+}}$ has been named by Benzécri (1973) as the sth row profile.

Goodman questioned whether the above criterion is valid even in the case when independence is rejected for the initial table. In this case merging categories that seem homogeneous according to the independence of the subtable criterion can affect the underlying association structure, although the fit of independence remains very bad. This led him to introduce the *structural criterion*, according to which, two (or more) homogeneous categories can be merged only if the association structure remains unchanged (Goodman 1981a,b). He connected detection of categories' homogeneity to association models by stating that equality of the scores of two

rows (or columns) under the RC model (or R or C), if it is adequate for describing the underlying association structure, implies homogeneity of the corresponding categories (Goodman 1981a). Later, he generalized this result by showing that equality of the scores of two rows (or columns) on all dimensions of the saturated association model RC(M^*) implies homogeneity of the corresponding categories (Goodman 1986). Analogously, homogeneity of categories can be decided upon the observed similarities in the row and/or column canonical coordinates, as discussed in Gilula (1986) and Gilula and Haberman (1986).

This result was further extended by Kateri and Iliopoulos (2003) for the general ϕ-divergence association model $\mathrm{RC}_\phi(M)$, defined by (7.11). They also showed that the predominant criteria for merging categories, e.g., homogeneity and structure, are always in agreement. In their context, the structure is described in terms of a generalized association model based on an information theoretic setup, which includes the models used by Goodman and Gilula as special cases. Furthermore, they proved that the scores of the $\mathrm{RC}_\phi(M)$ model applied on the merged table for the new merged category are equal to the corresponding common scores' values of the $\mathrm{RC}_\phi(M)$ of the initial table, provided the weights in constraints (6.17) for the merged table are the same as those of the initial table for the non-merged categories while for the new categories derived by merging, their weights are the sum of the weights of the corresponding merged categories in the initial table. This weights' condition is satisfied by the marginal weights but not by the uniform. Thus, when considering merging, the marginal weights should be used.

To summarize, merging between homogeneous classification categories ensures the preservation of the underlying structure of the probability table π. As a consequence, no simpler model should be appropriate for the merged table. Nevertheless, in practice we have to be cautious if, after merging categories, the merged table satisfies a simpler model. Either the assumption of categories' homogeneity does not hold or the association structure adopted for one of the two tables (initial and merged) is false. It cannot be the case that all these assumptions are correct but not in agreement.

Finally, keep the remark that deciding about merging categories based on the classical and simple independence criterion of the corresponding subtable is not always safe when independence is rejected for the initial table and its association structure is complicated. It is always better to draw conclusions about homogeneities and merging categories based on a model consistent with the underlying association structure.

7.5.1 Example 6.2 (Continued)

To illustrate this last remark, let us reconsider the example in Sect. 6.5.2. Wermuth and Cox (1998), by the homogeneity criterion, suggested the merging of columns 4 and 5 as well as rows 1 and 2. Indeed, independence holds for the 5×2 subtable formed by the two last columns ($G^2 = 0.835$, $df = 4$, p-value=0.9336).

Table 7.2 G^2 statistics for the fit of independence and association models applied on the table produced by merging the last two columns of Table 6.7

Model	G^2	d.f.	p-value
I	356.310	12	0.000
RC	23.487	6	0.001
RC(2)	1.809	2	0.405

The evidence for merging rows 1 and 2 is much weaker, since the test of independence applied on the corresponding subtable gives $G^2 = 6.949$ ($df = 4$, p-value=0.1386). Obviously the columns 4 and 5 are merged first. Testing then for homogeneity of the first two rows of the merged 5×4 table, we get $G^2 = 6.823$ ($df = 3$, p-value=0.078), which is not rejected at the 5 % level and is also consistent with the natural motivation of merging due to the low frequencies of the first row.

We have seen that the appropriate model for this data table is the RC(2). Recall that for deciding about merging categories we work with the marginal weights as explained in the previous section. Observing the scores' estimates for marginal weights in Table 6.9 we realize that the estimates for the last two columns are very close on both dimensions while for the first two rows their scores' estimates are close on the first dimension of the association ($m = 1$) but they are much different for the second ($m = 2$). This is directly visualized in Fig. 6.1. Scores for columns 4 and 5 are almost indistinguishable for marginal weights while for rows 1 and 2 are far apart, basically due to their second axis coordinate. Hence we have a strong indication that columns 4 and 5 are homogeneous but rows 1 and 2 not. This is verified by the hypothesis testing of homogeneity for the corresponding scores. For this example it has been tested asymptotically by computing the appropriate Mahalanobis distances and checking their significance (see Kateri and Iliopoulos 2003). Thus, we merge only the last two columns.

As expected (and verified in Table 7.2), the structure of association remains the same in the merged table, e.g., RC(2). If we checked homogeneity looking only on the scores of the first dimension (e.g., consider the RC model), then we would decide to merge also the first two rows. But in this case it is wrong to work with the RC model, since it does not describe the underlying structure of the association.

7.6 Overview and Further Reading

Correspondence analysis was originally developed primarily in France by Benzécri in the early 1970s (Benzécri 1973), though the algebraic derivation of the technique is due to Hirschfeld (1935). For the early history of CA, see de Leeuw (1983a,b). Important in the development and spreading of CA is the contribution, among others, of Hill (1974), Escoufier (1984), van der Heijden and de Leeuw (1985), Carroll et al. (1986, 1987, 1989), Greenacre and Hastie (1987), Heiser (1987),

Choulakian (1988), Greenacre (1989), van der Heijden et al. (1989), de Leeuw and van der Heijden (1991), Kim (1992), Kroonenberg and Lombardo (1999), and Gabriel (2002). Gilula (1984) connected CA to *latent class models* (Sect. 10.3.1). van de Velden and Kiers (2005) considered rotation in CA. Classic reference texts on CA are Greenacre (1984) and Lebart et al. (1984), while of practical importance with R implementations is Greenacre (2007). For a literature review on correspondence analysis see Beh (2004) and for recent trends in research, the Computational Statistics & Data Analysis special issue on CA (vol. 53(8), 2009), edited by Blasius, Greenacre, Groenen, and van de Velden. On nonsymmetric CA, we refer to Sect. 8.4.4 while *multiple CA* for multi-way tables is discussed in Sect. 7.6.1 below.

Goodman (1985, 1986, 1996, 2002a) related CA to correlation models and association models. Gilula et al. (1988) clarified the nature of their connection further while Kateri and Balakrishnan (2008) proved that association and correlation models are not equivalent in terms of statistical evidence (Royall 2000; Royall and Tsou 2003). The association models are bounded by the maximum of the bump function while the correlation models are not.

The ordered restricted CA has been considered by Schriever (1983), Parsa and Smith (1993), and Ritov and Gilula (1993). Beh (1997) considered CA for ordinal contingency tables with preassigned scores through orthogonal polynomials. Tests of independence, conditional on order-restricted CA or order-restricted RC models, are discussed in Kuriki (2005).

7.6.1 Homogeneity Analysis

CA has also been considered for higher dimensional tables, leading to *multiple correspondence analysis* (MCA). MCA is based on the simple CA of the *Burt matrix*. Characteristic related references are Greenacre (1988b), Kroonenberg (1989), Heiser and Heiser and Meulmann (1994), and Carlier and Kroonenberg (1996). MCA is equivalent to *homogeneity analysis* (see Gifi 1990; Michailidis and De Leeuw 1998). Tenenhaus and Young (1985) discussed extensively MCA and relevant methods, providing a rich reference list, also of the early developments. An overview of MVA can be found in Greenacre and Blasius (2006). For regularized MCA, motivated by ridge regression, see Hwang et al. (2009) and references cited therein. Park et al. (2007) applied CA in a genetic study.

7.6.2 Canonical Correlation and Correspondence Analysis

The model of correspondence analysis is equivalent to the canonical correlation model of Fisher (1940) while the partition of Pearson's X^2 statistic for testing independence in terms of the eigenvalues (7.4) lies back to the fundamental paper by

Lancaster (1963). The asymptotic distributions of the canonical correlations can be found in O' Neill (1978a,b) and their variances and covariances in O' Neill (1981). A conditional test of independence, based on the canonical correlation, is provided by Haberman (1981). The links of MCA to the canonical correlation are clarified in Gower (1990).

Chapter 8
Response Variable Analysis in Contingency Tables

Abstract Logit models for binary, nominal, and ordinal responses are introduced in Chap. 8. In particular beyond the basic logit model for binary response, the baseline category logit, the cumulative logit, and the proportional odds models are presented. Also logit models for ordinal explanatory variables are considered as well as the logit analysis of stratified 2×2 contingency tables. Logit models are connected to association models and illustrated with examples, worked out in R.

Keywords Logit model • Ordinal explanatory variables • Cumulative logit model • Cox's proportional odds model • Adjacent categories odds logit model • Rasch model • Stereotype model

8.1 Logit Models for Binary Response

Log-linear and association models treat all the classification variables in a symmetric way, modeling the association structure among them. They do not distinguish between dependent and independent variables. In case there exists an explicit *response* variable and interest lies in modeling its dependence on one or more explanatory (or predictor) variables, then *logit* models are applied instead. The response is characterized as dependent variable and the explanatory variables as independent. A logit model is a logistic regression model with all the independent variables categorical.

Consider an $I \times 2$ table $(n_{ij})_{I\times 2}$ with the column classification variable Y being the response (yes–no or success–failure). If $(\pi_{ij})_{I\times 2}$ is the corresponding probability table under the multinomial sampling scheme, the probability of success, conditional on the level i of the explanatory variable X, is defined as

$$\pi_{1|i} = P(Y = 1|X = i) = \frac{\pi_{i1}}{\pi_{i1} + \pi_{i2}} ,$$

M. Kateri, *Contingency Table Analysis: Methods and Implementation Using R*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4811-4_8,

while the odds of success will be $\frac{\pi_{1|i}}{1-\pi_{1|i}} = \frac{\pi_{i1}}{\pi_{i2}}$. The log of the odds of success is known as the *logit* of success

$$\text{logit}(\pi_{1|i}) = \log\left(\frac{\pi_{1|i}}{1-\pi_{1|i}}\right) = \log \pi_{i1} - \log \pi_{i2} \ .$$

Considering further the saturated log-linear model for the contingency table, since $\pi_{i1}/\pi_{i2} = m_{i1}/m_{i2}$, we have

$$\text{logit}(\pi_{1|i}) = \log\left(\frac{m_{i1}}{m_{i2}}\right) = \underbrace{(\lambda_1^Y - \lambda_2^Y)}_{\beta_0} + \underbrace{(\lambda_{i1}^{XY} - \lambda_{i2}^{XY})}_{\beta_i}, \ i = 1, \ldots, I,$$

leading to the standard expression of the logit model

$$\text{logit}(\pi_{1|i}) = \log \frac{\pi_{1|i}}{1-\pi_{1|i}} = \beta_0 + \beta_i, \quad i = 1, \ldots, I \ . \tag{8.1}$$

Parameters β_i express the effect of the explanatory variable X on the response Y, verified by the fact that the local odds ratio opposing the odds of success at predictor's level $i+1$ vs. i is $\theta_{i1}^L = \exp(\beta_{i+1} - \beta_i)$. One of the β_i parameters is redundant; thus, one of them (say, β_1) or their sum is set equal to 0 for identifiability reasons.

Model (8.1) is saturated. The hypothesis of "no effect" of X on Y,

$$H_0: \ \ \beta_1 = \beta_2 = \ldots = \beta_I = 0,$$

is equivalent to the hypothesis of independence between X and Y.

Model (8.1) is equivalently expressed in terms of success probabilities as

$$\pi_{1|i} = \frac{\exp(\beta_0 + \beta_i)}{1 + \exp(\beta_0 + \beta_i)}, \quad i = 1, \ldots, I \ . \tag{8.2}$$

For two-way tables there exists a one-to-one correspondence between the logit models and the class of hierarchical log-linear models. This is not the case for contingency tables of higher order. For example, for an $I_1 \times I_2 \times 2$ table with two explanatory variables X_1 and X_2, model (8.1) extends to

$$\text{logit}(\pi_{1|i_1 i_2}) = \beta_0 + \beta_i^{X_1} + \beta_i^{X_2} \ , \quad i = 1, \ldots, I \ ,$$

which corresponds to the log-linear model $(X_1X_2, \ X_1Y, \ X_2Y)$. However, in logit framework, focus lies on the dependence of the response on the explanatory variable while the association structure among the explanatory variables is not of interest and their interactions are considered in the greatest possible level. Thus, for the three-way table considered above, model $(X_1Y, \ X_2Y)$ has not a logit analogue.

In case of ℓ ($\ell \geq 2$) explanatory variables, the corresponding logit model is

$$\text{logit}(\pi_{1|i_1 i_2 \ldots i_\ell}) = \beta_0 + \sum_{s=1}^{\ell} \beta_{i_s}^{X_s}, \quad i_s = 1, \ldots, I_s \ , \ s = 1, \ldots, \ \ell, \tag{8.3}$$

and is equivalent to the $(X_1 X_2 \cdots X_\ell, \ X_1 Y, \ X_2 Y, \ldots, \ X_\ell Y)$ hierarchical log-linear model. Model (8.3) assumes only main effects of the explanatory variables on the response. Less parsimonious logit models including interaction terms are possible. As in the log-linear models framework, only hierarchical models are considered. The interaction terms are of the type $\beta_{i_k i_q}^{X_k X_q}$ and of higher order, up to the $\beta_{i1 i_2 \ldots i_\ell}^{X_1 X_2 \ldots X_\ell}$, which corresponds to the saturated model.

Summarizing, the point of differentiation between log-linear and logit model is that log-linear models analyze the structure of association among *all* factors, while logit models focus only on the way the response depends on the explanatory variables, conditioning on the explanatory variables cross-classification. The choice between them relies on the type of the problem and the goal of the analysis.

8.1.1 Logit Models for Ordinal Explanatory Variables

In the context of $I \times 2$ tables, when the explanatory variable X is ordinal, one can assign scores $x_1 \leq \ldots \leq x_I$ (with $x_1 < x_I$) to its categories and replace the β_i effect term in (8.1) by βx_i, leading to the model:

$$\text{logit}(\pi_{1|i}) = \log\left(\frac{\pi_{1|i}}{1 - \pi_{1|i}}\right) = \beta_0 + \beta x_i, \quad i = 1, \ldots, I. \tag{8.4}$$

This model has two parameters (just one more than independence) and is the logit expression of model LL in (6.4) with

$$x_i = \propto_i, \ \ i = 1, \ldots, I; \qquad \beta = \varphi(\nu_1 - \nu_2),$$

or of model U, if the x_i scores are equidistant for successive categories.

In case of multi-way tables with a binary response, scores can be assigned to some or all of the ordinal explanatory variables and the corresponding $\beta_i^{X_s}$ terms in model (8.3) can be replaced by the $\beta^{X_s} x_{si}$ terms. For example, for an $I_1 \times I_2 \times I_3 \times 2$ table with explanatory variables X_1 and X_2 ordinal and X_3 nominal or ordinal, the model

$$\text{logit}(\pi_{1|i_1 i_2 i_3}) = \beta_0 + \beta^{X_1} x_{1i} + \beta^{X_2} x_{2j} + \beta_k^{X_3}, \quad i = 1, \ldots, I_1, \ j = 1, \ldots, I_2, \ k = 1, \ldots, I_3,$$

applies parsimonious effect terms for X_1 and X_2 based on one parameter each and known scores and a factor effect term of I_3 levels for X_3.

8.1.2 Inference for Logit Models

The logit models for binary response belong to the family of GLM and thus the inferential results follow from the corresponding general results for GLMs. In particular, model (8.1) can be written in a vector form in terms of its nonredundant parameters, as

$$\text{logit}(\boldsymbol{\pi}_{1|X}) = \mathbf{X}\boldsymbol{\beta}\,, \tag{8.5}$$

where $\boldsymbol{\pi}_{1|X} = (\pi_{1|1},\ \ldots, \pi_{1|I})'$ is the $I \times 1$ vector of success probabilities at every explanatory level $X = i$, $\boldsymbol{\beta} = (\beta_0,\ \beta_1, \ldots,\ \beta_{I-1})'$ the $I \times 1$ vector of nonredundant model parameters, and $\mathbf{X} = (x_{ij})$ the $I \times I$ design matrix $\mathbf{X} = (\mathbf{1}_{I\times 1}\ \ \mathbf{I}^*)$, with $\mathbf{1}_{s\times t}$ the $s \times t$ matrix with all entries equal to 1 and $\mathbf{I}^*$ the $I \times (I-1)$ matrix $\mathbf{I}^* = \begin{pmatrix} \mathbf{I}_{I-1} \\ \mathbf{0}_{1\times(I-1)} \end{pmatrix}$, where $\mathbf{I}_s$ is the $s \times s$ identity matrix and $\mathbf{0}_{s\times t}$ the $s \times t$ matrix with all entries equal to 0.

Analogously, model (8.3) can be equivalently expressed by

$$\text{logit}(\boldsymbol{\pi}_{1|X_1,\ldots X_\ell}) = \mathbf{X}\boldsymbol{\beta}\,, \tag{8.6}$$

where $\boldsymbol{\pi}_{1|X_1,\ldots X_\ell} = (\pi_{1|11\ldots1},\ \pi_{1|11\ldots2}, \ldots \pi_{1|11\ldots I_\ell}, \ldots \pi_{1|I_1 I_2 \ldots I_\ell})'$ is the $\left(\prod_{s=1}^{\ell} I_s\right) \times 1$ vector of success probabilities at every combination of the levels of the ℓ explanatory variables, expanded by their order, $\boldsymbol{\beta} = (\beta_0,\ \beta_1^{X_1}, \ldots,\ \beta_{I_1-1}^{X_1},\ \ldots,\ \beta_1^{X_\ell}, \ldots,\ \beta_{I_\ell-1}^{X_\ell})'$ the $\left(\sum_{s=1}^{\ell} I_s - \ell + 1\right) \times 1$ vector of nonredundant parameters, and $\mathbf{X}$ the design matrix $\mathbf{X} = \left(\mathbf{1}_{\left(\prod_{s=1}^{\ell} I_s\right)\times 1}\ \ \mathbf{X}_1\right)$, with

$$\mathbf{X}_k = \begin{pmatrix} \mathbf{1}_k^{(1)} & \mathbf{X}_{k+1} \\ \mathbf{1}_k^{(2)} & \mathbf{X}_{k+1} \\ \vdots & \vdots \\ \mathbf{1}_k^{(I_k-1)} & \mathbf{X}_{k+1} \\ \mathbf{0}_{\left(\prod_{s=k+1}^{\ell} I_s\right)\times(I_k-1)} & \mathbf{X}_{k+1} \end{pmatrix},\quad k = 1, \ldots, \ell-2,\ \mathbf{X}_{\ell-1} = \begin{pmatrix} \mathbf{1}_{\ell-1}^{(1)} & \mathbf{I}^* \\ \mathbf{1}_{\ell-1}^{(2)} & \mathbf{I}^* \\ \vdots & \vdots \\ \mathbf{1}_{\ell-1}^{(I_{\ell-1}-1)} & \mathbf{I}^* \\ \mathbf{0}_{I_\ell\times(I_k-1)} & \mathbf{I}^* \end{pmatrix},$$

and $\mathbf{I}^* = \begin{pmatrix} \mathbf{I}_{I_\ell-1} \\ \mathbf{0}_{1\times(I_\ell-1)} \end{pmatrix}$, with $\mathbf{1}_k^{(i)}$ the $\left(\prod_{s=k+1}^{\ell} I_s\right) \times (I_k - 1)$ matrix with 1's at the ith column and 0's in all other entries.

Models (8.5) and (8.6) are of the form (5.2). In particular, for model (8.5), the response variable is the random sample success proportion for every level of the explanatory variable X, i.e., the $I \times 1$ vector $\mathbf{Y}$, with $\mathrm{E}(\mathbf{Y}) = \propto = \boldsymbol{\pi}_{1|X}$ and for the *logit link*, i.e., $\eta = g(\propto) = \text{logit}(\propto) = \log\left(\frac{\propto}{\mathbf{1}-\propto}\right)$, where $\mathbf{1}$ is a vector of ones, of the same dimension as $\propto$. For binomial proportions, it holds $\mathsf{Var}(Y_i) = \frac{\propto_i(1-\propto_i)}{n_i}$, $i = 1, \ldots, I$, where n_i is the sample size for level i of X. Substituting in (5.8) the quantities above, the likelihood equations for model (8.5) are calculated as

$$\sum_{i=1}^{I} \frac{n_i(y_i - \propto_i)x_{ij}}{\propto_i(1-\propto_i)} \cdot \frac{e^{\eta_i}}{(1+e^{\eta_i})^2} = 0 \Rightarrow \sum_{i=1}^{I} \frac{n_i(y_i - \propto_i)x_{ij}}{\propto_i(1-\propto_i)} \cdot \frac{\propto_i/(1-\propto_i)}{(1+\propto_i/(1-\propto_i))^2} = 0 \,,$$

for $j = 1,2$. Finally, the likelihood equations for β_0 ($j = 1$) and β ($j = 2$) are derived as

$$\sum_{i=1}^{I} n_i(y_i - \hat{\pi}_{1|i})x_{ij} = 0 \,, \quad j = 1,2 \,. \tag{8.7}$$

Analogously, for model (8.6) the responses are the sample success proportions for each combination of the explanatory variables, i.e., the $\left(\prod_{s=1}^{\ell} I_s\right) \times 1$ vector $\mathbf{y} = \mathbf{p}_{1|X_1,\ldots X_\ell}$ with $\propto = \pi_{1|X_1,\ldots X_\ell}$. The system (5.8) leads to the likelihood equations

$$\sum_{i=1}^{\prod_{s=1}^{\ell} I_s} n_i(y_i - \hat{\pi}_{1|i}x_{ij}) = 0 \,, \quad j = 1,\ldots,\sum_{s=1}^{\ell} I_s - \ell + 1, \tag{8.8}$$

for parameters β_0 ($j = 1$), $\beta_1^{X_1}$ ($j = 2$), …, $\beta_{I_\ell - 1}^{X_\ell}$ ($j = \sum_{s=1}^{\ell} I_s - \ell + 1$), respectively.

For (8.5) and (8.6), the diagonal entries (5.10) of the diagonal matrix $\mathbf{W}$ become $w_i = n_i\pi_{1|i}(1-\pi_{1|i})$, for $i = 1,\ldots, I$, or $i = 1,\ldots, \sum_{s=1}^{\ell} I_s - \ell + 1$, respectively, and the estimated covariance matrix is

$$\widehat{\mathrm{Cov}}(\hat{\beta}) = [\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}]^{-1} \,. \tag{8.9}$$

The ML estimate $\hat{\beta}$ of β is obtained by solving the set of equations (8.7) or (8.8) using an iterative algorithm, like the Newton–Raphson or the Fisher's scoring algorithm. The square roots of the diagonal entries of table (8.9) are the estimated standard errors of $\hat{\beta}$ and are used for testing the asymptotic significance of the terms of β or deriving Wald confidence intervals for β or odds ratios that are function of β. Furthermore, by (8.5), the estimated variance of $\mathrm{logit}(\pi_{1|x})$, for observed $X = x$, is $\mathbf{x}'\widehat{\mathrm{Cov}}(\hat{\beta})\mathbf{x}$. Based on this and (8.2), a corresponding interval for $\pi_{1|X}$ can be derived.

Model fit and model selection among nested logit models follow straightforward in the GLM framework (see Sects. 5.3.2 and 5.3.4). In practice, logit models can easily be fitted as GLMs (with logit link) in any software. We will illustrate it next in R.

8.1.3 Logit Models in R

The data from a quality control study of five production machines are provided in Table 8.1. The binary response is in the row classification variable X.

The probability of a defective product for machine i ($i = 1,\ldots,5$) is $\pi_{1|i}$, and model (8.1) is fitted in R by `glm` as follows.

Table 8.1 Quality control data for a sample of 500 products (hypothetical data) from five production machines, ordered from oldest (A) to newest (E)

	Product. Machine					
Product	A	B	C	D	E	
Defective	24	17	12	10	4	67
Non-defective	79	94	83	89	88	433

Table 8.2 Output by applying model (8.1) to the data in Table 8.1 by `glm`

```
Call:
glm(formula=cbind(def, nodef)~machine, family=binomial(link="logit"))

Deviance Residuals:
[1] 0 0 0 0 0
Coefficients:
             Estimate   Std. Error  z value   Pr(>|z|)
(Intercept)   -1.1914    0.2331      -5.112    3.19e-07 ***
machine2      -0.5187    0.3518      -1.474    0.140417
machine3      -0.7425    0.3869      -1.919    0.054971 .
machine4      -0.9947    0.4069      -2.445    0.014504 *
machine5      -1.8996    0.5619      -3.381    0.000722 ***
--
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

( Dispersion parameter for binomial family taken to be 1 )

   Null deviance: 1.7255e+01 on 4 degrees of freedom

Residual deviance: 9.3258e-15 on 0 degrees of freedom

AIC: 30.747

Number of Fisher Scoring iterations: 3
```

```
> def <- c(24,17,12,10,4);  nodef <- c(79,94,83,89,88)
> machine <- factor(1:5)
> bin.logit <- glm(cbind(def,nodef)~ machine, family=
+                  binomial(link="logit"))
> summary(bin.logit)
```

The output is provided in Table 8.2. We can verify that product defectiveness depends on production machine, since the hypothesis of independence is equivalent to $\beta_1 = \ldots = \beta_5 = 0$, which is rejected ($G^2$(I)=17.26, df=4, p-value=0.0017). In the `glm` output (see Table 8.2), G^2(I) is to be found under "null deviance." Observe that $\beta_1 = 0$, while we notice that the β_i's are decreasing in i, showing that the probability of a defective product decreases as the machine is newer.

In order to test for linear trend, we assign scores $x_i = i$, $i = 1, \ldots, 5$, to the machines and fit model (8.4) by

Table 8.3 Output by applying model (8.4) to the data in Table 8.1 by `glm`

```
Call:
glm(formula=cbind(def, nodef)~machlin, family=binomial(link="logit"))

Deviance Residuals:
1             2           3            4          5
0.08325      -0.39530    0.22304      0.62272   -0.60402
Coefficients:
             Estimate    Std. Error   z value    Pr(>|z|)
Intercept)    -0.8145     0.2792       -2.917     0.003534 **
machlin       -0.3963     0.1027       -3.859     0.000114 ***
--
Signif. codes:   0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

( Dispersion parameter for binomial family taken to be 1 )

   Null deviance: 17.25503 on 4 degrees of freedom

Residual deviance: 0.96556 on 3 degrees of freedom

AIC: 25.713

Number of Fisher Scoring iterations: 4
```

```
> def <- c(24,17,12,10,4);  nodef <- c(79,94,83,89,88)
> machlin <- 1:5
> lin.logit <- glm(cbind(def,nodef)~ machlin, family=
+                    binomial(link="logit"))
> summary(lin.logit)
```

From the derived output (Table 8.3), observe that (8.4) is a parsimonious model of very good fit with G^2=0.966 (p-value=0.8096, df=3) that captures the dependence of the defectiveness on the machine's age by a single parameter β. Since $\hat{\beta} = -0.3963$, the odds of defective product for a machine in age category $i+1$ is $e^{\hat{\beta}(x_i - x_{i+1})} = e^{-\hat{\beta}} = e^{0.3963} = 1.486$ times higher than for a machine in the immediate previous age category i, $i = 1, \ldots, 4$.

The within machine-type probability of defective product is estimated under the last model by

$$\hat{\pi}_{1|i} = \frac{\exp(\hat{\beta}_0 + \hat{\beta} x_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta} x_i)} = \frac{\exp(-0.8145 - 0.3963 x_i)}{1 + \exp(-0.8145 - 0.3963 x_i)}, \quad i = 1, \ldots, I\,.$$

These probabilities are saved in `lin.logit` under

```
> lin.logit$fitted.values
```

```
  1           2           3           4           5
0.22955379  0.16698819  0.11884463  0.08319487  0.05754063
```

8.2 Logit Analysis of Stratified 2×2 Contingency Tables

Reconsider the setup discussed in Sect. 3.3, cross-classifying the binary variables X and Y for the K levels of a third variable Z. Let us assume that Y is a response variable and $\pi_{ik} = \pi_{1|ik} = \mathsf{P}(Y = 1|X = i,\ Z = k)$. Since X is binary, model (8.3), adjusted for this case, becomes

$$\text{logit}(\pi_{1|ik}) = \beta_0 + \beta^X x_i + \beta_k^Z, \quad i = 1,2\ ,\ k = 1,\ldots,\ K, \tag{8.10}$$

with $x_1 = 0$ and $x_2 = 1$, and the β_i^X term in the logit model expression is equivalent to the linear trend term present in the model above. Under (8.10), the XY conditional odds ratio is the same (equal to e^{β^X}) for every level k of Z, i.e., the XY association is homogeneous across the levels of Z. If $\beta^X = 0$, X and Y are conditionally independent and the corresponding model is

$$\text{logit}(\pi_{1|ik}) = \beta_0 + \beta_k^Z, \quad i = 1,2\ ,\ k = 1,\ldots,\ K. \tag{8.11}$$

Model (8.11) is nested in (8.10), and if (8.10) fits the data well, then the fit of (8.11) can be tested conditional on (8.10). In particular, if G_1^2 and G_2^2 are the LR goodness-of-fit statistics of (8.10) and (8.11), respectively, then their difference $G_2^2 - G_1^2$ is asymptotically distributed as $\mathscr{X}_1^2$ (Sect. 5.3.4) and is used for the conditional testing of (8.11), given (8.10).

Model (8.10) is equivalent to the hierarchical log-linear model $(XY,\ XZ,\ YZ)$ while (8.11) to $(XZ,\ YZ)$ and the conditional testing discussed here is equivalent to the conditional test presented in Sect. 4.6.1.

This is illustrated on Example 3.3 of Sect. 3.3.4. The data have to be given in R in a matrix of two response columns, the first containing the successes and the second the failures. Thus, the data are entered as shown below.

```
> suc <- c(79,5,89,4,141,6,45,29,81,3,168,13)
> fail <- c(68,17,221,46,77,18,26,21,112,11,51,12)
> response <- cbind(suc,fail)
```

The factors F (prognostic factor) and C (clinic) are defined and the data are saved in a data frame by

```
> F <- factor(rep(1:2, 6));  C <- factor(rep(1:6, each=2))
> clinstudy.fr <- data.frame(response, F, C)
```

Models (8.10) and (8.11) are then fitted by

```
> bin.logit <- glm(response~F+C, family=binomial(link="logit"),
+                  data=clinstudy.fr)
> bin.logit2 <- glm(response~C, family=binomial(link="logit"),
+                  data=clinstudy.fr)
```

Verify that the models considered and the conditional test performed are identical to the corresponding conditional analysis via log-linear models in Sect. 4.6.1.1.

8.3 Logit Models for Multi-category Response

Consider that the response variable Y has J ($J > 2$) categories. Polytomous logit models describe the log-odds for all possible $\binom{J}{2}$ pairs of responses. A set of $J-1$ pairs is sufficient to produce all possible pair of responses. The way these pairs are defined is dictated by the type of the response variable and the problem-specific interest in forming the pairs to be compared. The possible odds are those used in defining the generalization of the odds ratios for $I \times J$ tables in Sect. 2.2.5.

8.3.1 Nominal Response

The set of the $J-1$ pairs to be opposed in the odds is defined by setting a category of the response variable Y as the baseline or reference category, usually the first or the last one, and then pairing every other category to the reference category. For reference category the last (J) and for nominal explanatory variable X, the logit model, known as the *baseline category logit model*, is defined by a set of $J-1$ equations

$$\text{logit}(\pi_{j|i}) = \log\left(\frac{\pi_{j|i}}{\pi_{J|i}}\right) = \beta_{0j} + \beta_{ij}\,, \quad i = 1,\ldots,I, \;\; j = 1,\ldots,J-1,$$

where $\pi_{j|i} = P(Y = j|X = i) = \frac{\pi_{ij}}{\pi_{i\cdot}}$. It is equivalent to the saturated log-linear model with $\beta_{0j} = \lambda_j^Y - \lambda_J^Y$ and $\beta_{ij} = \lambda_{ij}^{XY} - \lambda_{iJ}^{XY}$.

A non-saturated realistic model is derived by assuming that the explanatory variable effect is the same for all response odds. Then each of the $J-1$ logit equations has its own β_{0j} parameter ($j = 1,\ldots,J-1$), but they all share a common explanatory variable effect, depending only on the level of the explanatory variable i. This is the *proportional baseline category logit model*, given by

$$\text{logit}(\pi_{j|i}) = \beta_{0j} + \beta_i\,, \quad i = 1,\ldots,I, \;\; j = 1,\ldots,J-1. \tag{8.12}$$

The choice of the reference category affects the model parameters estimates but not the estimated cell probabilities.

8.3.2 Ordinal Response: The Cumulative Logit Model

The *cumulative logit model* is based on the cumulative odds for the response and is defined as

$$\begin{aligned}\text{logit}(\pi_{\leq j|i}) &= \log\left(\frac{P(Y \leq j|X=i)}{1-P(Y \leq j|X=i)}\right) = \log\left(\frac{\pi_{i1}+\ldots+\pi_{ij}}{\pi_{i,j+1}+\ldots+\pi_{iJ}}\right) \qquad (8.13)\\ &= \beta_{0j}+\beta_{ij}\,, \qquad\qquad i=1,\ldots,I,\ j=1,\ldots,J-1.\end{aligned}$$

This is the most general form, allowing the explanatory variable effect to vary among the different cumulative odds. When the explanatory variable X is also ordinal, then usually scores x_i, $i=1,\ldots,I$, are employed for its categories and the following model is defined:

$$\text{logit}(\pi_{\leq j|i}) = \beta_{0j}+\beta_j x_i\,, \quad i=1,\ldots,I,\ j=1,\ldots,J-1. \qquad (8.14)$$

In analogy to model (8.12), under the assumption of a common explanatory variable effect for all cumulative odds, model (8.13) leads to

$$\text{logit}(\pi_{\leq j|i}) = \beta_{0j}+\beta_i\,, \quad i=1,\ldots,I,\ j=1,\ldots,J-1, \qquad (8.15)$$

and (8.14) to

$$\text{logit}(\pi_{\leq j|i}) = \beta_{0j}+\beta x_i\,, \quad i=1,\ldots,I,\ j=1,\ldots,J-1, \qquad (8.16)$$

respectively. Model (8.16) is known as Cox's *proportional odds* model.

Under (8.15) the cumulative odds ratios (2.48) of the table are given by

$$\theta_{ij}^{C_Y} = \frac{\sum_{k\leq j}\pi_{ik}/\sum_{k>j}\pi_{ik}}{\sum_{k\leq j}\pi_{i+1,k}/\sum_{k>j}\pi_{i+1,k}} = \exp(\beta_i-\beta_{i+1})\,, \quad i=1,\ldots,I-1;\ j=1\ldots,J-1\,,$$

which means that model (8.15) is equivalent to the row effect association model R for the cumulative odds ratios.

Analogously, model (8.16) with equidistant scores for successive categories is equivalent to the uniform association model U for the cumulative odds ratios with β the common log cumulative odds ratio value, i.e.,

$$\theta_{ij}^{C_Y} = \frac{\sum_{k\leq j}\pi_{ik}/\sum_{k>j}\pi_{ik}}{\sum_{k\leq j}\pi_{i+1,k}/\sum_{k>j}\pi_{i+1,k}} = \exp\left(\beta(x_i-x_{i+1})\right),\ i=1,\ldots,I-1;\ j=1\ldots,J-1\,.$$

Thus, models (8.15) and (8.16) could also be fitted as cumulative odds ratios R and U models, respectively, in an analogue manner to the global odds ratios association models (Sect. 7.1).

The baseline category and cumulative logit models considered above are analogously defined for multi-way contingency tables with explanatory variables $X_1,\ \ldots,\ X_\ell$, with some or all of them ordinal. In matrix notation, the baseline category logit model for more than one explanatory variables is defined as

$$\text{logit}(\pi_{j|X_1,\ldots X_\ell}) = \mathbf{X}\beta\,, \quad j=1,\ldots,J-1\,,$$

with matrices $\mathbf{X}$ and β, defined appropriately, depending on the type of each explanatory variable and in analogy to the binary response case in Sect. 8.1.2. For the cumulative logit, $\pi_{j|X_1,\ldots X_\ell}$ in the equation above is replaced by $\pi_{\leq j|X_1,\ldots X_\ell}$. For example, the proportional odds model for a (ℓ+1)-way table with all ℓ explanatory variables ordinal and scores $x_{s,i_s} = i_s$, assigned to their categories $i_s = 1,\ldots,I_s$ ($s = 1,\ldots,\ell$), would be

$$\text{logit}(\pi_{\leq j|X_1,\ldots X_\ell}) = \beta_{0j} + \sum_{s=1}^{\ell} \beta_s i_s \ , \quad j = 1,\ldots,J-1 \ .$$

8.3.3 Alternative Models for Ordinal Response

The cumulative odds is the most frequent used type of odds for ordinal logit models. The adjacent categories and the continuation odds are possible alternatives. For an $I \times J$ table with the response on the columns, the $J-1$ adjacent column categories odds are defined for $j = 1,\ldots,\ J-1$ by

$$\frac{\pi_{j|j,j+1,i}}{\pi_{j+1|j,j+1,i}} = \frac{P(Y=j|(Y=j \text{ or } Y=j+1),\ X=i)}{P(Y=j+1|(Y=j \text{ or } Y=j+1),\ X=i)} = \frac{\pi_{ij}}{\pi_{i,j+1}} \ , \quad i = 1,\ldots,I \ .$$

Assuming a common explanatory variable effect on all $J-1$ odds, the *adjacent-categories odds logit model* is

$$\text{logit}\left(\pi_{j|j,j+1,i}\right) = \log\left(\frac{\pi_{ij}}{\pi_{i,j+1}}\right) = \beta_{0j} + \beta_i \ , \qquad i = 1,\ldots,I,\ j = 1,\ldots,J-1. \tag{8.17}$$

This model is the equivalent logit expressions of the row association model R, presented in Sect. 6.1.3. Model (8.17) is equivalent to (8.12), easily verified by the fact that $\log\left(\pi_{j|j,j+1,i}\right) = \log\left(\pi_{j|i}\right) - \log\left(\pi_{j|i}\right)$.

For ordinal explanatory variable and fixed scores x_i, $i = 1,\ldots,I$, assigned to its categories, the *proportional adjacent-categories odds logit model* is defined as

$$\text{logit}\left(\pi_{j|j,j+1,i}\right) = \beta_{0j} + \beta x_i \ , \quad i = 1,\ldots,I,\ j = 1,\ldots,J-1, \tag{8.18}$$

which, for equidistant scores for successive categories, is equivalent to the U association model (6.4).

Analogously, for the $J-1$ response continuation odds at explanatory level i

$$\frac{\pi_{j|\geq j,i}}{\pi_{>j|\geq j,i}} = \frac{P(Y=j|\ Y \geq j,\ X=i)}{P(Y>j|\ Y \geq j,\ X=i} = \frac{\pi_{ij}}{\pi_{i,j+1} + \ldots + \pi_{iJ}} \ ,$$

and for common explanatory variable effect on all odds, the *continuation ratio logit model* is defined by

$$\text{logit}\left(\pi_{j|\geq j,i}\right) = \log\left(\frac{\pi_{ij}}{\pi_{i,j+1}+\ldots+\pi_{iJ}}\right) = \beta_{0j}+\beta_i\,. \qquad (8.19)$$
$$i = 1,\ldots,I,\ j = 1,\ldots,J-1.$$

The *proportional continuation ratio logit model* is then

$$\text{logit}\left(\pi_{j|\geq j,i}\right) = \beta_{0j}+\beta x_i\,,\ \ i = 1,\ldots,I,\ j = 1,\ldots,J-1. \qquad (8.20)$$

Obviously models (8.17) to (8.20) extend also to cases of multiple explanatory variables.

The choice of type of logit model for ordinal response relies on the actual problem under consideration and the question addressed. A comparison of alternative options for logit analysis can be found in Cox and Chuang (1984).

8.3.4 Example 6.1 (Continued)

Logit models for ordinal response can be fitted by function `vglm()` of Yee's `VGAM` library. Alternatively, the `polr()` (library `MASS`) can be used.

We illustrate the cumulative logit model (8.14) applied on the cannabis data (Table 6.1). Use of cannabis is the response variable (three levels, ordered) and the explanatory is the alcohol consumption (four categories, ordered). The data have to be written in a matrix format, each column representing the responses vector for a given level of the explanatory variable. For this 4×3 table, the vectors correspond to the columns.

```
> library(VGAM)
> canb1 <- c(204,211,357,92);  canb2 <- c(6,13,44,34)
> canb3 <- c(1,5,38,49);  response <- cbind(canb1,canb2,canb3)
> alcohol <- 1:4
> cum.logit <- vglm(response~alcohol, family=cumulative)
> summary(cum.logit)
```

The output is provided in Table 8.4. The fitted model is

$$\text{logit}(\pi_{\leq 1|i}) = \log\left(\frac{P(Y\leq 1|X=i)}{1-P(Y\leq 1|X=i)}\right) = 4.8122-1.1492x_i\,,\quad i=1,\ldots,4,$$
$$\text{logit}(\pi_{\leq 2|i}) = \log\left(\frac{P(Y\leq 2|X=i)}{1-P(Y\leq 2|X=i)}\right) = 6.4548-1.3673x_i\,,$$

with $x_i = i$, and is of very good fit (p-value=0.5532).

Table 8.4 Output by applying model (8.14) to the data in Table 6.1 in VGAM

```
Call:
vglm(formula=response~alcohol, family=cumulative)

Pearson Residuals:
                 logit(P[Y<=1])   logit(P[Y<=2])
              1         -0.95929         0.427307
              2         -0.32758         0.255103
              3          1.04712        -0.266701
              4         -0.82452         0.042323
Coefficients:
                         Value       Std. Error     t value
(Intercept):1           4.8122          0.36471      13.195
(Intercept):2           6.4548          0.56152      11.495
    alcohol:1          -1.1492          0.11336     -10.137
    alcohol:2          -1.3673          0.16369      -8.353

Number of linear predictors: 2

Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])

( Dispersion Parameter for cumulative family: 1 )

Residual Deviance: 3.02777 on 4 degrees of freedom

Log-likelihood: -18.91251 on 4 degrees of freedom

Number of Iterations: 4
```

The estimated multinomial response probabilities for each of the three cannabis use groups is given by

```
> fitted.values(cum.logit)
```

```
       cannab1      cannab2       cannab3
1   0.9749880   0.01887674   0.006135302
2   0.9251109   0.05123438   0.023654685
3   0.7965296   0.11664030   0.086830108
4   0.5536876   0.17454653   0.271765829
```

Assuming $\beta_1 = \beta_2$, the proportional odds cumulative logit model (8.16) is fitted by

```
> prop.cum <- vglm(response~alcohol,
+           family=cumulative(parallel=TRUE))
> summary(prop.cum)
```

giving the output of Table 8.5. The fitted model is thus

$$\text{logit}(\pi_{\leq 1|i}) = \log\left(\frac{P(Y \leq 1|X=i)}{1-P(Y \leq 1|X=i)}\right) = 4.8914 - 1.1784 \cdot i\,, \qquad i = 1,\ldots,4,$$

$$\text{logit}(\pi_{\leq 2|i}) = \log\left(\frac{P(Y \leq 2|X=i)}{1-P(Y \leq 2|X=i)}\right) = 5.8137 - 1.1784 \cdot i\,,$$

Table 8.5 Output by applying model (8.16) to the data in Table 6.1 in VGAM

```
Call:
vglm(formula=response~alcohol, family=cumulative(parallel = TRUE))

Pearson Residuals:
                logit(P[Y<=1])   logit(P[Y<=2])
             1         -1.48843          1.15241
             2         -0.76749          1.00743
             3          0.93117          0.20792
             4         -0.17846         -0.91400
Coefficients:
                        Value       Std. Error      t value
(Intercept):1          4.8914          0.36524       13.392
(Intercept):2          5.8137          0.38155       15.237
      alcohol         -1.1784          0.11296      -10.432

Number of linear predictors: 2

Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])

( Dispersion Parameter for cumulative family: 1 )

Residual Deviance: 6.33295 on 5 degrees of freedom

Log-likelihood: -20.5651 on 5 degrees of freedom

Number of Iterations: 4
```

with p-value=0.2752. Under this model, all six fitted cumulative odds ratios (2.48) of the table are fixed, i.e.,

$$\hat{\theta}_{ij}^{C_Y} = e^{1.1784[i-(i+1)]} = 3.25\,, \quad i = 1,2,3;\ \ j = 1,2\,,$$

while the estimated multinomial response probabilities are

```
> fitted.values(prop.cum)
```

	cannab1	cannab2	cannab3
1	0.9761771	0.01421219	0.009610758
2	0.9265291	0.04290351	0.030567439
3	0.7951277	0.11193918	0.092933116
4	0.5443026	0.20593812	0.249759244

The fit of this model is comparable to that of the global odds ratios U model (Sect. 7.1.1). We verify that the classical U model of constant local odds ratios, discussed for this example in Sect. 6.1.2, is the most suitable for this data set. The equivalent logit expression for the classical U model (8.18) is fitted in VGAM by

```
> prop.adj <- vglm(response~alcohol, family=acat(parallel=TRUE))
```

8.4 Overview and Further Reading

The logit link was used very early in bioassay applications (Bartlett 1937; Berkson (1944, 1953)). Berkson (1955) proposed a least squares procedure of parameters estimation, alternative to the ML estimation. Odoroff (1970) compared the alternative methods of estimation and discussed small samples properties of tests for interaction in $2 \times 2 \times 2$ and $3 \times 2 \times 2$ contingency tables. The distribution of the estimated parameters for binary explanatory variables and small samples is investigated by Whaley (1991). Other early references on logit models are Cox (1958a,b), Bishop (1969), and Goodman (1971b). The book of Cox (1970a) covered related models and treated their inferential aspects. For a recent reference on models for binary response and issues related to their fit and diagnostics, we address to Collet (2003).

References on testing conditional independence in stratified 2×2 tables in presence of a binary response via the logit consideration include Prentice (1976) and Day and Byar (1979). For updated work, see Agresti and Hartzel (2000) and Cheng et al. (2010).

For ordinal responses, Walker and Dunkan (1967) considered the cumulative logit, while the continuation logit was modeled by Fienberg and Mason (1978). A key reference on regression models for ordinal variables is the discussion paper by Anderson (1984).

Logit models for binary and ordinal response are presented in more detail in Agresti (2010, 2013), Dobson and Barnett (2008), and Hosmer et al. (2013).

8.4.1 Alternative Links to the Logit

An alternative to the logit link is the *probit* (Bliss 1934, 1935). Berkson (1951) argued for the logit link. However, the shapes of the logit and probit models are similar, differentiated at the tails of the distribution (Cox and Snell 1989). Another important link is the *complementary log–log* link. The logit and probit links *probit* are more appropriate for symmetric distributions while the complementary log–log when the distribution is skewed. The optimal choice of the link function is discussed in McCullagh (1980).

Generalized link functions have been proposed by Genter and Farewell (1985) and by Kateri and Agresti (2010), who proposed a log-gamma link and a link based on the ϕ-divergence, respectively. Stukel (1988) proposed a class of logistic models that extends the standard model by introducing two shape parameters, achieving thus improved fit for the noncentral probability regions and applicability of the logistic model to asymmetric probability curves.

8.4.2 Chain Graph Models and Collapsibility

Collapsibility is treated in Sects. 4.8 and 4.9.4 in terms of symmetric associations. Parametric collapsibility criteria in case of categorical explanatory variables and a binary response have been discussed in Wermuth (1987), who considered and compared conditions of collapsibility based on a symmetric measure of association (odds ratio) and a directed one (relative risk).

In the framework of graphical models, whenever there is an a priori division of the classification variables into explanatory and response variables, the corresponding graphical models consist of a combination of directed and undirected edges and are known as *chain graph models* (see Wermuth and Lauritzen 1990). Chain graph models are thus a generalization of undirected graphs and directed acyclic graphs (DAGs). They depend strongly on the explanatory–response partition of the variables made before fitting and selecting the appropriate model with the exception of a set of Markov equivalent models (Roverato 2005). For a discussion on chain graph models, their interpretations, their misuse, and their relation to DAGs and structural equation models (SEMs), see Lauritzen and Richardson (2002). A particular type of chain graph models for contingency tables has been considered by Marchetti and Lupparelli (2011).

Collapsibility of hierarchical log-linear models for multi-way contingency tables was considered in Asmussen and Edwards (1983). Additional to the graph models approach, they used also directed graphs for contingency tables with response variables, noting the corresponding models as "causal chain models." Tunaru (2001) compared the class of graphical models to the class of chain graph models in terms of collapsibility, arguing that these two classes can lead to different results.

8.4.3 The Rasch Model

The Rasch model, introduced by Rasch in 1960 (see Rasch 1980), is a parametric *item response* model, with major areas of application in psychometrics, education, and behavioral sciences. It uses the logistic function to model individual's abilities or attitudes based on the corresponding assessment's data. If Y_{ij} is the ith subject's binary response (correct–false or presence–absence) on item j of the questionnaire or test, then the Rasch model, in its simplest form, is defined by

$$\text{logit}P(Y_{ij}=1)=\beta_0+\beta_{1i}+\beta_{2j}\,,\quad i=1,\ldots,n,\quad j=1,\ldots,J,$$

where β_{1i} and β_{2j} are the ith subject's and jth item's effect on the response, respectively. An important property of the Rasch model is the independence of the item parameters from specific subjects and vice versa. Standard asymptotic results for maximum likelihood fail in this case leading to inconsistent estimators for the item parameters as $n\to\infty$. The Rasch model and associated inferential

results are presented in Andersen (1980). The goodness-of-fit tests for the Rasch model are sensitive in model's assumption. For a related discussion, we refer to Kelderman (1984), who formulated the Rasch model as quasi-independence model in the log-linear models framework. Rasch models for multi-way contingency tables are discussed in Goodman (1990).

The Rasch model has been expanded for multiple-response items by Andersen (1973). For polytomous Rasch models see also Andrich (2010) and references cited there. For the extended class of Rasch models, we refer to Fischer and Molenaar (1995) and von Davier and Carstensen (2007). A family of extended Rasch models can be fitted in `R` by package `eRm` (Mair and Hatzinger 2007). Polytomous Rasch models are fitted by the `R` function `plRasch`. For a presentation of the function but also a connection of the polytomous Rasch model to linear-by-linear association models, an updated bibliography, discussion on Rasch models, and connection to log-linear models, see Anderson et al. (2007).

8.4.4 The Stereotype Model

We have seen that the adjacent-categories odds logit models (8.17) and (8.18) are equivalent logit expressions of the R and U association models, respectively. The logit version of the RC association model with monotone scores for the categories of the response variable is the *stereotype model*, introduced by Anderson (1984). The stereotype model and its properties and estimation are discussed, among others, by Kuss (2006) and Johnson (2007). For its connection to the RC association model, see Douglas and Fienberg (1990).

The logit analogue of correspondence analysis is known as nonsymmetric correspondence analysis (see Lauro and D' Ambra 1984; Lauro and Balbi 1999). Nonsymmetric correspondence analysis for ordinal variables with scores assignment based on orthogonal polynomials is considered by Lombardo et al. (2007).

Chapter 9
Analysis of Square Tables

Abstract Special models for matched pairs of ordinal responses are presented in Chap. 9. Beyond the classical models of symmetry and quasi symmetry, the models of conditional symmetry, diagonal symmetry, and ordinal quasi symmetry are discussed in detail. The model of marginal homogeneity is tested as a generalized log-linear model and not only conditioning on quasi symmetry, as is usually done. Connections to rater agreement models and mobility table analysis are made.

Keywords McNemar's test • Symmetry models • Marginal homogeneity • Quasi independence • Homogeneous association models • Rater agreement • Bradley-Terry model

9.1 Comparison of Two Dependent Proportions

Often it is the case that we are interested in comparing two proportions that are not independent. Thus, the procedure in Sect. 2.1.3 cannot be followed. Dependent proportions occur when a binary response is measured on the same subject at two different time points or occasions, like data in Table 2.1(c), or when a binary response is measured on two different but paired, correlated subjects (e.g., husband and wife). Dependent proportions arise also when two binary response variables are measured on the same subjects and are cross-classified. The n paired observations are then cross-classified, as shown below:

	Occasion B		
Occasion A	Success	Failure	
Success	n_{11}	n_{12}	n_{1+}
Failure	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	n

The success (or the "yes") probabilities of each occasion correspond to the marginal probabilities π_{1+} and π_{+1} and interest lies on testing whether these two

M. Kateri, *Contingency Table Analysis: Methods and Implementation Using R*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4811-4_9,

probabilities are equal. Their equality is equivalent to $\pi_{12} = \pi_{21}$, with the probability of pairs to agree on "yes," π_{11}, playing no role. The corresponding null hypothesis is

$$\mathrm{H}_0 \ : \ \pi_{1+} = \pi_{+1} \ \Leftrightarrow \ \pi_{12} = \pi_{21} \tag{9.1}$$

Conditioning on the sum of the disjoint responses $N_{12} + N_{21} = n^*$, the random number of pairs N_{12} in cell $(1,2)$ follows a binomial distribution

$$N_{12} \sim \mathscr{B}(n^*, \ \pi) \ ,$$

with $\pi = \frac{\pi_{12}}{\pi_{12}+\pi_{21}}$, and the null hypothesis (9.1) is equivalent to

$$\mathrm{H}_0 \ : \ \pi = 1/2 \ . \tag{9.2}$$

Under (9.2), $\mathrm{E}(N_{12}) = \frac{n^*}{2}$ and $\mathrm{Var}(N_{12}) = \frac{n^*}{4}$, leading to the score test statistic

$$Z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} \ . \tag{9.3}$$

The asymptotic distribution for Z under (9.1) is the standard normal. Thus, the asymptotic distribution for

$$X^2 = Z^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \tag{9.4}$$

is $\mathscr{X}^2_1$. Test statistic (9.4) was introduced by McNemar (1947), and the corresponding test of significance, rejecting (9.1) at significance level α if $X^2 \geq X^2_{1,\alpha}$, is known as McNemar's test.

The asymptotic $(1-a)100\%$ Wald confidence interval for the difference of the correlated marginal probabilities $\pi_{1+} - \pi_{+1}$ is

$$\left(p_{1+} - p_{+1} - z_{a/2}\hat{\sigma}, \quad p_{1+} - p_{+1} + z_{a/2}\hat{\sigma}\right) \ , \tag{9.5}$$

where the variance $\sigma^2 = \mathrm{Var}(\hat{\pi}_{1+} - \hat{\pi}_{+1})$ is considered under the general parametric space and is equal to

$$\sigma^2 = \frac{\pi_{1+}(1-\pi_{1+}) + \pi_{+1}(1-\pi_{+1}) - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})}{n} \ . \tag{9.6}$$

It is estimated by

$$\hat{\sigma}^2 = \frac{p_{1+}(1-p_{1+}) + p_{+1}(1-p_{+1}) - 2(p_{11}p_{22} - p_{12}p_{21})}{n} \ , \tag{9.7}$$

where $p_{ij} = n_{ij}/n$ are the observed sample proportions.

Table 9.1 Cross-classification of GSS respondents with college degree aged 20–50 in 2000 by voting in 2000 and 2004 US elections for (a) males and (b) females

	2004		
2000	Yes	No	Total
(a) Males			
Yes	223	11	234
No	25	23	48
Total	248	34	282
(b) Females			
Yes	297	16	313
No	33	50	83
Total	330	66	396

For comparing two success probabilities, the matched pairs design is preferable to the independent samples design, in case it is possible, since it leads to a statistic of smaller variance. Indeed, if π_1 and π_2 are the success probabilities to be compared based on two independent equal-sized samples ($n_1 = n_2 = n$), with sample estimates p_1 and p_2, respectively, then the variance of the difference $\hat{\pi}_1 - \hat{\pi}_2$, $\text{Var}_I(\hat{\pi}_1 - \hat{\pi}_2)$, is given by (2.9). If $\pi_1 = \pi_{1+}$ and $\pi_2 = \pi_{+1}$, the variance (9.6) and $\text{Var}_I(\hat{\pi}_1 - \hat{\pi}_2)$ for $n_1 = n_2 = n$ are related as follows:

$$\sigma^2 = \text{Var}_I(\hat{\pi}_1 - \hat{\pi}_2) - \frac{2}{n}c\,,$$

where $c = \pi_{11}\pi_{22} - \pi_{12}\pi_{21}$. Responses of matched pairs are usually positive associated; thus $\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} > 0$, which implies $c > 0$, and consequently, for positive-associated responses, $\sigma^2 < \text{Var}_I(\hat{\pi}_1 - \hat{\pi}_2)$.

9.1.1 Example 9.1

From the GSS data set, we cross-tabulated respondents born between 1950 and 1980 and with educational level of at least college degree according to whether they voted or not in the 2000 and 2004 US elections by gender. The data are given in Table 9.1.

For our example, π_{1+} and π_{+1} in (9.1) are the probabilities of voting in 2000 and 2004, respectively. The McNemar test (9.4) is evaluated in `R` by the function `mcnemar.test()` of `stats` that reads the data in table form. Thus, for Table 9.1(a), the McNemar test is computed in `R` by

```
> vote.M <- matrix(c(223,11,25,23), nrow=2, byrow=T, dimnames=
+     list("Voted 2000"=c("yes","no"),"Voted 2004"=c("yes","no")))
> mcnemar.test(vote.M)
```

giving the output

```
    McNemar's Chi-squared test with continuity correction
data:   vote.M
McNemar's chi-squared = 4.6944,  df = 1,  p-value = 0.03026
```

Analogously, for females in Table 9.1(b), we get McNemar's $X^2 = 5.2245$ with associated p-value$=0.02227$. Thus, hypothesis (9.1) is rejected, for males and females at significance level $a = 0.05$, meaning that the percentage of voting among the responders is significantly different between the 2000 and 2004 elections. In particular, the percentage of voters has increased in 2004 compared to 2000. For one-sided alternatives, the test statistic (9.3) should be used.

For the males the 95% asymptotic confidence interval (9.5) for the difference $\pi_{1+} - \pi_{+1}$ is $(-0.0909, -0.0083)$, meaning that the increase of the voting percentage for men in 2004 elections lies between 0.8 and 9%. Analogously, the corresponding interval for females is $(-0.0773, -0.0085)$. Their computation is straightforward and the function `McNemar.CI()` of the web appendix (see Sect. A.3.6) can be applied for this.

9.2 Symmetry Models

The special case of square $I \times I$ contingency table with commensurable classification variables occurs often in biomedicine, educational and social sciences applications, in psychology and sports, among other fields. Characteristic such cases refer to treatments' comparison or "before–after" comparisons applied on the same subjects, cross-classification of responses in matched pairs designs, problems of rater agreement, social mobility tables , or models of preference in opinion surveys. In this framework, interest lies on the off-diagonal cells and the models of symmetry and marginal homogeneity consist the starting or reference point. If symmetry is not a meaningful choice, which is usually the case, there is need to consider special models of asymmetry that measure departures from symmetry toward certain direction.

9.2.1 Complete Symmetry

The standard hypothesis of complete symmetry (S) is defined as

$$\mathrm{H}_0 \ : \ \pi_{ij} = \pi_{ji} \ , \quad i > j \ , \quad i, j = 1, \ldots, I \ , \tag{9.8}$$

or equivalently in terms of the expected cell frequencies as

$$\mathrm{H}_0 \ : \ m_{ij} = m_{ji} \ , \quad i > j \ , \quad i, j = 1, \ldots, I \ .$$

The S model does not model the diagonal cells of the table. They are kept fixed or, in other words, modeled exactly. Actually, this is true not only for the S model but for all models for square tables considered in this section.

It can easily be verified that the ML estimates of the expected under (9.8) cell frequencies are

$$\hat{m}_{ij} = \frac{n_{ij} + n_{ji}}{2}, i, j = 1, ..., I. \tag{9.9}$$

Hypothesis (9.8) is tested through the classical test of Bowker (1948), based on the test statistic

$$X^2 = \sum_{i>j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}} , \tag{9.10}$$

for which the asymptotic distribution under (9.8) is the chi-squared with associated degrees of freedom $df(\mathrm{S}) = I(I-1)/2$. Hypothesis (9.8) is rejected for high values of (9.10). This test of Bowker is a generalization of McNemar's test for $I \times I$ tables with $I > 2$. For $I = 2$, (9.8) is equivalent to (9.1) and (9.10) reduces to (9.4). The corresponding LR statistic G^2, given by expression (2.37) for $\hat{m}_{ij}$ as in (9.9), is asymptotically equivalent.

The hypothesis of complete symmetry can equivalently be expressed, in terms of a log-linear model, as the model defined by (4.5), i.e.,

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \quad i, j = 1, \ldots, I ,$$

with the additional constraints

$$\lambda_i^X = \lambda_j^Y , \tag{9.11}$$

$$\lambda_{ij}^{XY} = \lambda_{ji}^{XY} , \quad i, j = 1, \ldots, I . \tag{9.12}$$

9.2.2 Marginal Homogeneity

The hypothesis of marginal homogeneity (MH)

$$\mathrm{H}_0 : \pi_{i+} = \pi_{+i}, \quad i = 1, \ldots, I , \tag{9.13}$$

states that the marginal distributions of a square contingency table are equal. One of the Eq. in (9.13) is redundant, due to $\sum_{i,j} \pi_{ij} = 1$. For an $I \times I$ table, complete symmetry (9.8) implies MH (9.13), while for the special case of 2×2 tables, models S and MH are equivalent and tested by the McNemar test.

The tests proposed for marginal homogeneity are asymptotically chi-squared distributed with $df(\mathrm{MH}) = I - 1$. For a short reference to the early MH tests, see Sect. 9.7. Alternatively, MH can be tested conditionally, as we shall see in Sect. 9.2.3.

MH has not a straight representation in log-linear model form but is a characteristic case of marginal model and inference can easily be developed in the marginal models framework (see Sect. 5.6), since (9.13) is equivalent to

$$m_{i+} - m_{+i} = 0, \quad i = 1, \ldots, I-1.$$

If $\mathbf{m} = (m_{11}, m_{12}, \ldots, m_{1I}, m_{21}, \ldots, m_{II})'$ is the $I^2 \times 1$ vector of expected cell frequencies, expanded by row, then MH can be expressed as a special MPH model of the type (5.30) by

$$\mathbf{C}\log(\mathbf{M}\mathbf{m}) = \mathbf{0}\,, \tag{9.14}$$

where $\mathbf{M}$ is the $2I \times I^2$ matrix

$$\mathbf{M} = \begin{pmatrix} & & \mathbf{I}_b & \\ \mathbf{I} & \mathbf{I} & \ldots & \mathbf{I} \end{pmatrix}\,,$$

with $\mathbf{I}$ the $I \times I$ identity matrix and $\mathbf{I}_b$ the $I \times I^2$ block-identity matrix

$$\mathbf{I}_b = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \ldots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{1} \end{pmatrix}\,,$$

with $\mathbf{1}$ and $\mathbf{0}$ the $1 \times I$ vectors of 1's and 0'w, respectively. $\mathbf{C}$ is the $(I-1) \times 2I$ matrix

$$\mathbf{C} = \begin{pmatrix} \mathbf{I}_{I-1} & -\mathbf{I}_{I-1} \end{pmatrix}\,,$$

where $\mathbf{I}_{I-1}$ is the $(I-1) \times I$ subtable of $\mathbf{I}$, formed by deleting the last row, and $\mathbf{0}$ the $(I-1) \times 1$ vector of 0's.

The fit of the MH model by expression (9.14) in Liang's `mph` function is discussed in Sect. 9.2.6 and illustrated in Sect. 9.2.7 below.

9.2.3 *Quasi Symmetry*

The model of complete symmetry S is a parsimonious model with sound interpretation. However, tables that satisfy model S are not common in practice. A less strict model for square tables is the quasi-symmetry (QS) model, introduced by Caussinus (1965). It is the model under which the local odds ratios of the table $\theta_{ij}^L = \frac{\pi_{ij}\pi_{i+1.j+1}}{\pi_{i+1.j}\pi_{i.j+1}}$ are symmetric instead of the cell probabilities π_{ij}, i.e., QS is defined by

$$\theta_{ij}^L = \theta_{ji}^L, \quad i, j = 1, \ldots, I-1, \quad i > j\,.$$

In terms of a log-linear model representation, it is defined by (4.5) when (9.12) holds, i.e., it is produced by the S model by relaxing the constraint (9.11). Thus, complete symmetry implies quasi symmetry. The associated degrees of freedom are $df(\text{QS}) = (I-1)(I-2)/2$.

The ML estimates of the expected cell frequencies under QS are not derived in closed-form expressions. The corresponding likelihood equations that have to be solved iteratively are

$$\hat{\pi}_{i+} = p_{i+} , \qquad i = 1, \ldots, I-1 , \tag{9.15}$$

$$\hat{\pi}_{+j} = p_{+j} , \qquad j = 1, \ldots, I-1 , \tag{9.16}$$

$$\hat{\pi}_{ij} + \hat{\pi}_{ji} = p_{ij} + p_{ji} , \qquad i, j = 1, \ldots, I, \quad i > j . \tag{9.17}$$

Actually, only one set of (9.15) and (9.16) is needed, since the other is redundant given (9.17). It is thus clear that expression (4.5) is overparameterized for QS.

Although its physical interpretation is not straightforward, it is a powerful model due to its links to other models, as, for example, the association models (Sect. 9.4.1) and the Bradley–Terry model (Sect. 9.6). The QS model is invariant under any permutation of the categories of the classification variables, provided that the same permutation is applied to both rows and columns.

Its nature can be better understood by its connection to the models S and MH, as stated by the property

$$\mathrm{S} = \mathrm{MH} \wedge \mathrm{QS} . \tag{9.18}$$

This means that when MH and QS both hold, then symmetry does also hold. Property (9.18) is used to build a conditional test for MH, as the test of symmetry, given that the quasi-symmetry model holds

$$G^2_{\mathrm{MH|QS}} = G^2_{\mathrm{S}} - G^2_{\mathrm{QS}} ,$$

which is asymptotically chi-squared distributed with $df(\mathrm{MH|QS}) = df(\mathrm{S}) - df(\mathrm{QS}) = I - 1$. This conditional test of MH is usually applied in the context of log-linear models, since the fit of S and QS is direct while of MH is not.

In the log-linear models setup, QS can alternatively be defined by

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^X + a_i + \lambda_{ij}^{XY} , \quad i, j = 1, \ldots, I , \tag{9.19}$$

subject to (9.12), which is not overparameterized and expresses QS as departure from the complete symmetry (model S). The parameters a_i, $i = 1, \ldots, I$, capture this departure and in view of property (9.18) they are indicative of sources of marginal inhomogeneity in the table (Becker 1990b). For identifiability purposes we set $a_1 = 0$ or

$$\sum_{i=1}^{I} a_i = 0 . \tag{9.20}$$

We adopt constraint (9.20), since it treats the categories symmetrically and thus the a_i's express the contribution of each category to marginal inhomogeneity in

reference to the overall mean. For $a_1 = a_2 = \ldots = a_I = 0$, (9.19) implies the S model while the larger (in absolute value) a_i is, the larger the contribution of category i to marginal inhomogeneity is.

Equivalent to (9.19) is the multiplicative expression in terms of cell probabilities

$$\pi_{ij} = \frac{2\alpha_i}{\alpha_i + \alpha_j}\pi_{ij}^S, \quad i,j = 1,\ldots I, \tag{9.21}$$

where π_{ij}^S is the expected cell probability under the S model (Kateri and Papaioannou 1997). Under (9.20), it holds $a_i = \log \alpha_i$ and

$$\alpha_i = \exp\left(\frac{1}{I}\sum_j \log\left(\frac{\pi_{ij}}{\pi_{ji}}\right)\right) . \tag{9.22}$$

9.2.4 Conditional (or Triangular) Symmetry

The model of conditional symmetry (Bishop et al. 1975; McCullagh 1978), known also as triangular (T) symmetry model (Goodman 1979c), is defined as

$$\pi_{ij} = \tau^* \pi_{ji}, \quad i > j,$$

and states that the cell probabilities of the lower triangle of the table are proportional to the corresponding upper triangle symmetric cells and the proportionality constant is the same for all cells.

The equivalent definition for T, as a departure from S model, is

$$\pi_{ij} = \tau\pi_{ij}^S \mathrm{I}(i > j) + (2-\tau)\pi_{ij}^S \mathrm{I}(i < j), \quad i,j = 1,\ldots,I,\ i \neq j\ , \tag{9.23}$$

with $\mathrm{I}(\cdot)$ the indicator function. Note that $\tau^* = \frac{\tau}{2-\tau}$. Model T has just one parameter more than S, namely the τ, and thus $df(\mathrm{T}) = I(I-1)/2 - 1$. The ML estimate of τ is derived in closed-form

$$\hat{\tau} = \frac{2\sum_{i>j} p_{ij}}{\sum_{i>j} p_{ij} + \sum_{i<j} p_{ij}} .$$

The log-linear expression for (9.23) is

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^X + \lambda_{ij}^{XY} + t\mathrm{I}(i > j), \quad i,j = 1,\ldots,I,\ i \neq j\ , \tag{9.24}$$

with the additional constraint (9.12).

Obviously, the T model implies S when $\tau^* = 1$ (or $\tau = 1$). Furthermore, it holds

$$\begin{aligned} \mathrm{S} &= \mathrm{T} \wedge \mathrm{Q}S\ , \\ \mathrm{S} &= \mathrm{T} \wedge \mathrm{M}H\ . \end{aligned} \tag{9.25}$$

9.2.5 *Diagonal Symmetry*

The model of diagonal (D) symmetry (Goodman 1979c) is adequate for ordinal contingency tables and models departures from complete symmetry in terms of the distance between the classification categories of rows and columns, i.e., in terms of the secondary diagonals of the table. Thus, it is defined as

$$\pi_{ij} = \delta^*_{i-j}\pi_{ji} \ , \quad i > j \ , \tag{9.26}$$

or equivalently

$$\pi_{ij} = \delta_{i-j}\pi^S_{ij}\mathrm{I}(i > j) + (2 - \delta_{i-j})\pi^S_{ij}\mathrm{I}(i < j), \quad i, j = 1, \ldots, I \ , \ i \neq j \ , \tag{9.27}$$

with $\mathrm{I}(\cdot)$ the indicator function and $\delta^*_{i-j} = \frac{\delta_{i-j}}{2-\delta_{i-j}}$. It has $I-1$ additional parameters than S, namely the parameters δ_k, $k = 1, \ldots, I-1$, and $df(\mathrm{D}) = (I-1)(I-2)/2$. It is especially useful in modeling rater agreement (see Sect. 9.5.2).

The ML estimates of δ_k are

$$\hat{\delta}_k = \frac{2\sum_{i-j=k} p_{ij}}{\sum_{|i-j|=k} p_{ij}} \ , \quad k = 1, \ldots I-1 \ .$$

As a log-linear model, (9.27) is given by

$$\log m_{ij} = \lambda + \lambda^X_i + \lambda^X_j + \lambda^{XY}_{ij} + d_k\mathrm{I}(i - j = k), \quad i, j = 1, \ldots, I, \ i \neq j \ , \tag{9.28}$$

with the interaction parameters $\{\lambda^{XY}_{ij}\}$ satisfying (9.12).

More parsimonious models can be derived by setting homogeneity constraints among some of the δ_k-parameters. For each homogeneity constraint, one degree of freedom is released. In particular, if all δ_ks are equal, i.e., $\delta_1 = \delta_2 = \ldots = \delta_{I-1} = \delta_0$, model D is reduced to the triangular symmetry model T with $\delta_0 = \tau$. Due to the ordinal classification variables, it may make sense to assume the diagonal parameters to be ordered, i.e.,

$$\delta_1 \leq \delta_2 \leq \ldots \leq \delta_{I-1} \ , \tag{9.29}$$

implying that the proportionality constant of row and column probabilities in symmetric cells increases as they come further apart. In this case, the parameters δ_k, $k = 1, \ldots, I-1$, are estimated under the order restriction (9.29) by isotonic regression. The order-restricted D model will be illustrated in Example 9.2 (Sect. 9.2.7).

9.2.6 Software for Symmetry Models

The models of symmetry are not directly fitted in statistical packages. In SPSS, models S, QS, T, and D can be fitted in MATRIX by the macro provided in the web appendix (see Sect. A.4). This requires the data to be inserted in table format. In R, they can be fitted by function `glm`, as log-linear models, as shown below. The model of MH can be tested conditionally on QS. An unconditional test, providing also cell estimates under MH, can be performed through marginal models and Liang's `mph` function in R.

To fit the symmetry models in R, let `NI` be the number of rows I of the $I \times I$ table and `freq` the vector of cell frequencies, given by rows. The associated row and column factors are defined by the commands

```
> row <- gl(NI,NI,length=NI^2) ;  col <- gl(NI,1,length=NI^2)
```

and can be bound along with the frequencies' vector in the data frame `example`

```
> example <- data.frame(freq,row,col)
```

The model of complete symmetry (S) is then fitted by

```
> S <- gnm(freq~Symm(row,col),data=example,family=poisson)
```

while that of QS by

```
> QS <- gnm(freq~Symm(row,col)+row,data=example,family=poisson)
```

This fits the QS model in its log-linear form (9.19) subject to (9.12) but under the constraint $a_1 = 0$. These a_i's have to be rescaled linearly to satisfy constraint (9.20). This is achieved by

```
> r1 <- NI*(NI+1)/2+1 ;  r2 <- r1+ (NI-2)
> a <- c(0, coef(QS)[r1:r2]) ;  a <- a-sum(a)/NI
```

Vector `a` is the vector of the ML estimates of the model's (9.19) parameters a_i ($i = 1,\ldots,I$). The ML estimates of the α_i parameters of model expression (9.21) are then

```
> alpha <- exp(a)
```

Alternatively, the $\hat{\alpha}_i$'s can be obtained by substituting the cell probabilities π_{ij} in (9.22) by their ML estimates under QS $\hat{\pi}_{ij}$, avoiding thus the rescaling of vector `a` above. This is implemented by

```
A <- exp(rowSums(log(matrix(QS$fitted.values, nrow=NI, byrow=T)/
+          t(matrix(QS$fitted.values, nrow=NI, byrow=T))))/NI)
```

For the T and D models, we need to define two further vectors `t` and `d`. For this, we define the vectors (not factors) indicating the row and column for each entry of `freq`, `X` and `Y`, respectively, as

```
> v <- c(1:NI) ;  X <-rep(v,each=NI) ;  Y <-rep(v, NI)
```

Then, the factors `t` and `d` are

```
> t<-as.numeric(X>Y)
```

and

```
> d <- X-Y
> for(i in 1:NI^2) {if (d[i]<0)
+                   {d[i]<-NI+abs(d[i])} else {d[i]<-d[i]+1}}
> d<-factor(d)
```

Table 9.2 The factors needed to fit the symmetry models in GLM

row				col				sm				sqs				t				d			
1	1	1	1	1	2	3	4	2	3	4	5	1	2	3	4	0	0	0	0	1	5	6	7
2	2	2	2	1	2	3	4	3	4	5	6	2	1	5	6	1	0	0	0	2	1	5	6
3	3	3	3	1	2	3	4	4	5	6	7	3	5	1	7	1	1	0	0	3	2	1	5
4	4	4	4	1	2	3	4	5	6	7	8	4	6	7	1	1	1	1	0	4	3	2	1

respectively. Finally, the T and D models are defined by

```
> TS <- gnm(freq Symm(row,col)+t,data=example,family=poisson)
> DS <- gnm(freq~Symm(row,col)+d,data=example,family=poisson)
```

The derived ML estimates for the parameters associated to `t` and `d` are the $\hat{\tau}^*$ and $\hat{\delta}_k^*$, $k = 1, \ldots, I-1$.

The symmetry models discussed above can alternatively be fitted in `glm`. For this, the factors `row`, `col`, `sm`, `sqs`, `t`, and `d` are required. To understand the structure of these factors, they are provided for $I = 4$ in Table 9.2.

For any value of `NI`, these factors, expanded in a vector form by rows, are created by function `SYMV(NI)` of the web appendix mentioned in Sect. A.3.6. Thus, for the data frame defined above, we set

```
> row <-SYMV(NI)$row ;  col <- SYMV(NI)$col
> t <- SYMV(NI)$t ;     d <- SYMV(NI)$d
> sm <- SYMV(NI)$sm ;   sqs <- SYMV(NI)$sqs
```

and the symmetry models are fitted by

```
> S.model <- glm(freq~sm+sqs,data=example, family=poisson)
> QS.model <- glm(freq~row+col+sqs,data=example, family=poisson)
> TS.model <- glm(freq~sm+sqs+t,data=example, family=poisson)
> DS.model <- glm(freq~sm+sqs+d,data=example, family=poisson)
```

To fit the MH model in `mph`, the function

```
> h.fct <- function(p){
  p.row <- M.fct(row)%*%p
  p.col <- M.fct(col)%*%p
  as.matrix(c(p.row[-NI] - p.col[-NI]))}
```

is required, that returns the $I-1$ differences between the row and column marginals, excluding the last category marginal that is redundant. The MH model is then applied by

```
> MH <- mph.fit(y=freq, h.fct=h.fct)
```

and summary results, the cell estimates and residuals included, are printed by

```
> mph.summary(MH,T)
```

Alternatively, MH can be tested conditionally, provided the QS model holds as

```
> anova(S.model, QS.model, test="Chisq")
```

Table 9.3 Cross-classification of GSS male respondents by degree of pride with regard to America's economic (rows) vs. scientific and tech (columns) achievements

	Science and tech				
Economic	1	2	3	4	Total
1: Very proud	369	59	6	1	435
2: Somewhat proud	226	238	10	3	477
3: Not very proud	60	67	14	2	143
4: Not proud at all	7	16	3	2	28
Total	662	380	33	8	1083

Table 9.4 Goodness of fit for symmetry models applied on Table 9.3

Model	G^2	df	p-value
I	210.3535	9	0.0000
S	217.9977	6	0.0000
MH	197.8918	3	0.0000
QS	3.4181	3	0.3315
T	8.4775	5	0.1318
D	4.3021	3	0.2306
D_{23}	4.3288	4	0.3633

9.2.7 Example 9.2

From the GSS data set, males' degree of pride with regard to America's economic vs. scientific and tech achievements is cross-tabulated as shown in Table 9.3.

The symmetry models discussed above are applied on this data set and the corresponding goodness-of-fit statistics are provided in Table 9.4. For comparative purposes the statistic for testing independence is also provided.

Models S, QS, T, D, and MH are fitted in `R`, as described in Sect. 9.2.6, for

```
> NI <- 4
> freq <- c(369,59,6,1,226,238,10,3,60,67,14,2,7,16,3,2)
> vision.w <- data.frame(freq,row,col)
```

and by changing the data frame name in `gnm` model specification commands from `data=example` to `data=vision.w`.

The test statistics for S and MH are highly significant while models QS, T, and D are of acceptable fit. The test for model MH in Table 9.4 is the unconditional, based on marginal models. MH can also be tested, conditional on QS, engaging property (9.18). Indeed, $G^2_{\text{MH|QS}} = G^2_{\text{s}} - G^2_{\text{QS}} = 217.9977 - 3.4182 = 214.5795$ with associated $df(\text{MH|QS}) = df(\text{S}) - df(\text{QS}) = 3$.

Regarding the QS model, the ML estimates of the a_i parameters are obtained by the procedure described in Sect. 9.2.6 and for this example are equal to $\hat{a}_1 = -1.766$, $\hat{a}_2 = -0.487$, $\hat{a}_3 = 1.112$, and $\hat{a}_4 = 1.140$, leading further to the ML estimates of the $\alpha_i = \log a_i$ parameters, $\hat{\alpha}_1 = 0.171$, $\hat{\alpha}_2 = 0.615$, $\hat{\alpha}_3 = 3.040$, and $\hat{\alpha}_4 = 3.128$.

The parameter estimated by model T in `gnm` is $\hat{\tau}^* = 1.5431$, leading to $\hat{\tau} = \frac{2\exp(\hat{\tau}^*)}{\exp(\hat{\tau}^*)+1} = 1.6478$. Similarly, for the D model, the `gnm` output provides $\hat{\delta}_1^* = 1.4277$, $\hat{\delta}_2^* = 2.1335$, $\hat{\delta}_3^* = 1.9459$ and thus $\hat{\delta}_1 = 1.6131$, $\hat{\delta}_2 = 1.7882$, $\hat{\delta}_3 = 1.7500$.

We observe that $\hat{\delta}_2$ and $\hat{\delta}_3$ violate the order $\delta_1 \leq \delta_2 \leq \delta_3$. The order-restricted ML estimates are $\hat{\delta}_1^* = 1.4277$, $\hat{\delta}_2^* = \hat{\delta}_3^* = 2.1163$ and the corresponding model is denoted by D_{23}, the subscript indicating the δ-parameters that are equated. To obtain D_{23} we need to modify accordingly the factor `d` of the D model in the corresponding `gnm` model formula. The factor `d` for $I = 4$ is provided in Table 9.2 and to fit the D_{23} model, it has to be modified to `d2`, which is

1	5	6	6
2	1	5	6
3	2	1	5
3	3	2	1

This can easily be derived by

```
> d23<-d
> for(i in 1:NI^2) {if (d23[i]==7) {d23[i]<-6}
+          else if (d23[i]==4) {d23[i]<-3}}
```

The model is finally obtained by

```
> d23 <- factor(d23)
> D23 <- gnm(freq~Symm(row,col)+d23,data=vision.w,family=poisson)
```

Among the symmetry models tried, based on their goodness of fit and parsimony, our final choice is the D_{23}. However, different adopted models highlight different features of the data set. Thus, in this GSS example, the T, D, and D_{23} models all state that responders are more proud for scientific and tech achievements of the USA than for economic. Their difference lies on the refining of the comparison. Under T the ratio of the expected probabilities in symmetric cells is constant for any level comparison of the pride scale, while under D it is constant within level comparisons that are equally far apart. Under D_{23} this ratio differs for comparing levels that are one unit apart and levels that are two or three units apart. In particular, model T states that comparing the respondents' feelings toward scientific and economic achievements, when opposing any two levels of the pride scale, the probability that the "more proud" level is assigned to the scientific achievements is 1.54 times higher than that of being assigned to the economic ones. On the other hand, under the proposed D_{23} model, the probability of being more proud for scientific than for economic achievements is 1.4 times higher than that of being more proud for economic achievements, when comparing "very proud" to "somewhat proud," and 2.1 times higher for any comparison between the levels "somewhat proud," "not very proud," and "not proud at all."

Furthermore, we observed that the MH model is of very bad fit. The estimates of the QS model a_i's indicate that marginal inhomogeneity is mainly due to the first category, followed by the last. More detailed, the relative contribution of each category to inhomogeneity is dictated by the ordering of the absolute values of $\hat{\alpha}_i$'s from highest to lowest.

9.3 Quasi-Independence Models for Square Tables

Quasi-independence models have been introduced in Sect. 5.5. In the special case of square tables with commensurable classification variables, a basic model of special interest is the QI model that excludes the main-diagonal entries. In a mobility table or a panel study this corresponds in excluding from the analysis cases they were stable and focusing thus on items changing status. The QI model fitted on the non-diagonal entries of a $I \times I$ table is defined in terms of expected cell frequencies as

$$m_{ij} = \alpha_i \beta_j \,, \qquad i \neq j \,, i, j = 1, \ldots, I \,. \tag{9.30}$$

According to the problem under study, the diagonal entries are either structural zeros or their values are fitted exactly. In practice, model (9.30) can be fitted by function `QI.model()`, discussed in Sect. 5.5, defining appropriately the index vector `zer`. Alternatively, it can be fitted in the `gnm` package, activating the `Diag()` structure. Such a case is exhibited in the Example given in Sect. 9.3.1 below. For $I = 3$ the QI model (9.30) is equivalent to the QS model while for $I > 3$, it implies QS.

Interesting applications of model (9.30) are met in the context of analyzing rater agreement (see Sect. 9.5.2) and mobility tables of any type (like social or occupational) and of measuring change (e.g., in opinion or voting), like Example 9.3 below.

Another QI model for square tables occurs in cases under which all cells above (or below) the diagonal of the table are structural zeros. This triangular type QI model makes sense when the commensurable classification variables are of ordinal scale and is known that in the second occasion, the situation measured cannot be worse (or better) than in the first, producing thus structural zeros. For example,

$$m_{ij} = \alpha_i \beta_j \,, \qquad i \geq j \,, i, j = 1, \ldots, I \,, \tag{9.31}$$

fits the independence model only on the lower triangle of the table. Model (9.31) can be fitted by function `QI.model()`, for structural zeros index vector given by

```
zer <- rep(1, NI^2)-t
```

for `NI`=I and `t` the corresponding vector introduced in Sect. 9.2.6. Model (9.31) is illustrated in Example 9.4.

9.3.1 Example 9.3

The example analyzed here is an election example, exhibiting the voting shifts of a sample of voters between the 1970 and 1974 (February) elections in Great Britain. The data, provided in Table 9.5, are initially from Fuller ("New Society," 12 September 1974) and analyzed by Morgan and Titterington (1977). A voting shift table, formulated by cross-classifying the vote of a sample of voters at two

Table 9.5 Voting transitions between 1970 and 1974 Morgan and Titterington (1977)

	1974 election				
1970 election	Con.	Lab.	Lib.	Abst.	Total
Con.	619 (619.00)	62 (62.15)	115 (115.24)	79 (78.61)	875
Lab.	41 (40.85)	604 (604.00)	99 (96.31)	74 (76.84)	818
Lib.	11 (10.76)	11 (13.69)	82 (82.00)	9 (6.55)	113
Abst.	73 (73.39)	112 (109.16)	63 (65.45)	98 (98.00)	346
Total	744	789	359	260	2152

In parentheses are given the ML estimates of the expected cell frequencies under the QS model

elections, is a classical example of a table with augmented diagonal entries (stable party preference). It is unrealistic to assume (or test) independence for such tables. It is however of interest to test whether the shifts in voting behavior are independent for voters that changed preference, i.e., to test model (9.30).

The QI model (9.30), excluding the main-diagonal cells, is fitted on Table 9.5 in `gnm` of `R`, as is illustrated next:

```
> NI <- 4
> row <- gl(NI,NI,length=NI^2)
> col <- gl(NI,1,length=NI^2)
> freq <- c(619,62,115,79,41,604,99,74,11,11,82,9,73,112,63,98)
> election <- data.frame(freq,row,col)
> QI<-gnm(freq~row+col+Diag(row,col),data=election,family=poisson)
```

The degrees of freedom for model (9.30) applied on Table 9.5 are $df(\text{QI}) = 4$ and the associated LR statistic value is $G^2 = 39.874$, highly significant. Thus, the QI model is rejected for this data set. However, the improvement of goodness of fit achieved by the QI model, compared to the classical independence model ($G^2 = 1417.275$ with $df(\text{I}) = 9$), is impressive. The QS model is of very good fit for this data set with $G^2 = 1.745$ ($df(\text{QS}) = 2$, p-value$=0.8832$). The estimates of its α parameters are $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4) = (2.044, 1.343, 0.191, 1.908)$. The $\hat{\alpha}_i$ values, compared relative to 1, indicate that the largest shift occurred toward the Liberal party that strengthened its power in 1974 while all other parties lost power, the greatest loss observed for the Conservatives.

9.3.2 Example 9.4

The data in Table 9.6 are from Bishop and Fienberg (1969) and refer to 121 stroke patients at the Massachusetts General Hospital. The patients' physical disability following a stroke was graded on their admission on a five-point scale A–E of increasing severity. The patients were re-rated on the same scale in discharge. One patient's score on the second examination could be the same or better than the first one, since none of the patients had a second stroke. This data set was reanalyzed by Altham (1975) and Goodman (1979a).

Table 9.6 Initial and final rating on physical disability of 121 stroke patients

	Final state					
Initial state	A	B	C	D	E	Total
A	5 (5.00)	–	–	–	–	5
B	4 (3.75)	5 (5.25)	–	–	–	9
C	6 (4.43)	4 (6.20)	4 (3.37)	–	–	14
D	9 (6.16)	10 (8.63)	4 (4.69)	1 (4.52)	–	24
E	11 (15.66)	23 (21.92)	12 (11.93)	15 (11.48)	8 (8.00)	69
Total	35	42	20	16	8	121

In parentheses are given the ML estimates of the expected cell frequencies under the QI model

In this setup, QI is defined by model (9.31), which is fitted on data of Table 9.6 in R by function `QI.model()` with the associated `zer` vector defined through the `X` and `Y` vectors of Sect. 9.2.6. In particular,

```
> v <- c(1:5) ;  X <-rep(v,each=5) ;  Y <-rep(v, 5)
zer <- rep(1,25)-as.numeric(X>=Y)
```

and finally

```
> freq<-c(5,0,0,0,0,4,5,0,0,0,6,4,4,0,0,9,10,4,1,0,11,23,12,15,8)
> QI.T <- QI.model(freq,zer,5,5)
```

lead to $G^2 = 9.5958$ with an asymptotic p-value equal to 0.1427 on $df = 6$. Thus, QI is plausible, indicating that for these stroke patients, the state of their physical ability in discharge can be regarded as independent of their initial state.

9.4 Symmetry Models with Scores

9.4.1 Homogeneous Association Models

For an $I \times I$ contingency table, the homogeneous RC(M) association model, $1 \leq M \leq I-1$, is denoted by $\text{RC}_h(M)$ and is model (6.18) with homogeneous row and column scores, i.e., with

$$\propto_{im} = \nu_{im} \ , \ i = 1, \ldots, I, \ m = 1, \ldots, M \ .$$

Hence, it is defined as

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \sum_{m=1}^{M} \varphi_m \propto_{im} \propto_{jm}, i, j = 1, \ldots, I \ , \tag{9.32}$$

with $df(\text{RC}_h(M)) = (I-M-1)^2 + \sum_{\rho=0}^{\rho^*}(I-2-\rho)$, where $\rho^* = \min\{M-1, I-2\}$.

For $M = I-1$, $df(\text{RC}_h(I-1)) = (I-1)(I-2)/2$ and the $\text{RC}_h(I-1)$ model is equivalent to the QS model. For $1 \leq M < I-1$, more parsimonious models of

symmetric interactions are derived. The most well-known model of this type is the homogeneous RC model, RC_h, derived for $M = 1$. In this case, the degrees of freedom given by the formula above reduce to $df(\mathrm{RC}_h) = (I-1)(I-2)$. However the $\mathrm{RC}_h(M)$ model with $M < I-1$ does not fit the diagonal entries exactly. To consider special types of quasi-symmetric model, in the sense that the interactions are symmetric and the diagonal cells are excluded from the model, I extra parameters have to be added in the model. Thus, for $1 \leq M < I-1$, the model

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + d_i \mathrm{I}(i=j) + \sum_{m=1}^{M} \varphi_m \propto_{im} \propto_{jm}, \quad i,j = 1,\ldots,I\,, \tag{9.33}$$

is a quasi-symmetric model and will be denoted by $\mathrm{RC}_h^d(M)$. Of special interest is the simplest model of this class RC_h^d, achieved for $M = 1$,

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + d_i \mathrm{I}(i=j) + \varphi \propto_i \propto_j, i,j = 1,\ldots,I\,. \tag{9.34}$$

For $I = 3$, model (9.34) is saturated while for $I = 4$ it is equivalent to the QS model and $df(\mathrm{RC}_h^d) = 3$. For $M > 4$, RC_h^d is more parsimonious than QS and $df(\mathrm{RC}_h^d) = I^2 - 4I + 2$ (Goodman 1985).

The simplest quasi-symmetric model is the homogeneous uniform association model with the diagonal fitted exactly, U_h^d, which is defined by (9.34) with the scores $\propto_i$, $i = 1,\ldots,I$, being not parametric but known and equidistant for successive categories. The associated degrees of freedom are $df(\mathrm{U}_h^d) = (I-1)^2 - I - 1$. Recall that under the standard uniform (U) association model, all local odds ratios are constant, as stated by (6.1). For the U_h^d model, due to the exclusion of the diagonal cells, (6.1) is extended to

$$\begin{aligned} \log \theta_{ij}^L = \; & \varphi(\propto_i - \propto_{i+1})(\propto_j - \propto_{j+1}) + (d_i + d_{i+1})\mathrm{I}(i=j) \\ & - d_i \mathrm{I}(i=j+1) - d_{i+1}\mathrm{I}(i=j-1)\,, \quad i,j = 1,\ldots,I-1\,, \;\; i<j\,, \end{aligned} \tag{9.35}$$

while $\theta_{ij}^L = \theta_{ji}^L$, due to the quasi-symmetric nature of the U_h^d model.

Note that for the homogeneous association models RC_h, RC_h^d, and U_h^d, the scores $\propto_i$, $i = 1,\ldots,I$, are under constraints (6.5), i.e., for uniform weights in the general form of constraints (6.17). Marginal weights are not possible, since they have to be applied simultaneously for rows and columns. This is obviously true also for the $\mathrm{RC}_h(M)$ and $\mathrm{RC}_h^d(M)$ models with $M > 1$.

9.4.2 Ordinal Quasi Symmetry

For ordinal classification variables a special QS model has been proposed by Agresti (1983b), which is based on the assignment of known score values to their categories and has just one parameter more than the model of complete symmetry,

independently of the size of the table, the model of ordinal quasi symmetry (OQS). It is a simpler version of Goodman's D model, for the case the log-odds parameters δ^*_{i-j} in (9.26) follow a pattern linear in the distance between the row and column categories $i-j$, i.e., for

$$\delta^*_{i-j} = \delta^{i-j}\,, i > j\,. \tag{9.36}$$

This model is called the OQS model, since it is simultaneously a special type of quasi-symmetry model as well, as can easily be verified by setting $\alpha_i = \delta^i$ in (9.21). Whereas likelihood equations for QS equate observed and fitted margins, this model equates observed and fitted means.

In log-linear formulation, a generalized OQS model that allows to scale the differences $i-j$ in (9.36), for different i's and j's, $i > j$, is

$$\log m_{ij} = \lambda + \lambda_i^X + \lambda_j^X + \beta u_i + \lambda_{ij}^{XY}, \quad i,j = 1,\ldots,I, \tag{9.37}$$

with $\lambda_{ij}^{XY} = \lambda_{ji}^{XY}$, $i,j = 1,\ldots,I$, and $u_1, u_2, \ldots, u_I$ known scores. For $u_i = i$, $i = 1,\ldots,I$, this model is equivalent to (9.36) with $\beta = -\log\delta$. For $\beta = 0$, (9.37) reduces to the S model.

For ordinal square tables, OQS is a very powerful model, since it can be interpreted as a D and a QS model. It is very parsimonious having just one parameter more than the model of complete symmetry S. Furthermore, it fits well when there is an underlying bivariate normal distribution (Agresti 1983b).

The OQS model is invariant under linear transformations of the scores; thus we can set, without loss of generality, the scores to satisfy the constraints

$$\sum_i u_i = 0 \;\text{ and }\; \sum_i u_i^2 = 1. \tag{9.38}$$

These constraints are useful for the interpretation of β and in agreement with standard identifiability restrictions set on scores.

9.4.3 *Example 9.2 (Continued)*

Reconsidering the example introduced in Sect. 9.2.7, since the classification scale is ordinal from very proud (1) to not proud at all (4), the quasi-symmetric models based on scores, discussed in this section, can be applied.

Recall that for this 4×4 example `X` and `Y` are the vectors of the row and column classification variables, respectively, `row` and `col` the corresponding factors, and `vision.w` the associated data frame.

Then, the OQS model can easily be fitted in `gnm` by specifying the model as

```
> OQS <- gnm(freq~Symm(row,col)+X, data=vision.w,family=poisson)
```

This way, the scores used in (9.37) are the simple row scores appearing in `X`, i.e.,

$u_i = i$ ($i = 1, \ldots, 4$). To fit model (9.37) under the constraints (9.38), the scores in `x` need to be rescaled before the application of the model. This rescaling is achieved by function `rescale.square()`, to be found in the web appendix mentioned in Sect. A.3.6. Thus, if `Xu` is the vector of the linearly rescaled scores, model OQS under constraints (9.38) is fitted by

```
> NI<- 4 ;  Xu<-rep(rescale.square(NI), each=NI)
> OQS <- gnm(freq~Symm(row,col)+Xu, data=vision.w,family=poisson)
```

The RC_h^d is obtained in `gnm` by

```
> gnm(freq~row+col+Diag(row,col)+MultHomog(row,col),
+           family=poisson,data=vision.w)
```

In this case, model RC_h^d is equivalent to the classical QS model, since $I = 4$. For the simpler model of homogeneous uniform association U_h^d with scores constrained by (6.5), we have

```
> Yu<-rep(rescale.square(NI), NI) ;  u <- Xu*Yu
> U.hd<-gnm(freq~row+col+Diag(row,col)+u,family=poisson,
+           data=vision.w)
```

More generally, the homogeneous RC model RC_h, with no special consideration for the diagonal entries, is fitted as above by omitting the `Diag(row,col)` term. For the $RC_h^d(K)$ model, with $K > 1$, the term `MultHomog(row,col)` is replaced by `instances(MultHomog(row,col),K)`.

For data in Table 9.3, both models, OQS and U_h^d, provide an adequate fit. In particular, $G^2(\text{OQS}) = 8.8867$ with $df = 5$ (p-value=0.1137) and $G^2(U_h^d) = 3.9002$ with $df = 4$ (p-value=0.4197).

For scores satisfying (9.38), the ML estimate of parameter β in (9.37) is $\hat{\beta} = 2.876$. The positive sign of $\hat{\beta}$ indicates that the lower triangle of Table 9.3 is more probable than the upper, i.e., responders are less proud for economic than for science and tech achievements. Furthermore, for the probabilities in symmetric cells under OQS, it holds by (9.37) that

$$\frac{\pi_{ij}}{\pi_{ji}} = \exp[\beta(u_i - u_j)] \, , i > j \, .$$

That is, according to the OQS model's structure, the odds of an observation falling a certain distance under the main diagonal of the table (instead of the same distance above it) are estimated as

$$\frac{\hat{\pi}_{ij}}{\hat{\pi}_{ji}} = \exp[2.876(u_i - u_j)] \, , i > j \, .$$

Hence, the probability of being not very proud for economic achievements ($i = 3$) and very proud for science and tech achievements ($j = 1$) is

$$\frac{\hat{\pi}_{31}}{\hat{\pi}_{13}} = \exp\{2.876[0.2236 - (-0.6708)]\} = \exp(2.572) = 13.097$$

times the probability of being not very proud for science and tech and very proud for economic achievements. The same proportionality constant applies also when comparing any i, j levels two units apart, i.e., for $I = 4$, and also when comparing $i = 4$ to $j = 2$.

Fitting the U_h^d model subject to (6.5) leads to $\hat{\varphi} = 2.097$ for the intrinsic association parameter φ. Under this model and in view of (9.35), interpretation results based on specific odds ratios can be extracted.

9.5 Rater Agreement

Consider an $I \times I$ contingency table $(n_{ij})_{I\times I}$, cross-classifying items according to their ranking by two separate raters (medical doctors, teachers, book critics, wine tasters, etc.) on the same scale. Interest lies in analyzing their agreement. The natural choice for measuring agreement of the raters is the probability of their actual agreement $\pi_a = \sum_{i=1}^{I} \pi_{ii}$.

The most popular measure of two raters' agreement is Cohen's kappa (Cohen 1960), defined as

$$\kappa = \frac{\pi_a - \pi_I}{1 - \pi_I} , \tag{9.39}$$

where $\pi_I = \sum_{i=1}^{I} \pi_{i+}\pi_{+i}$ is the probability of the expected agreement if the two raters were rating independently. Thus, it adjusts the probability of actual agreement for "random" agreement by chance, which is captured by π_I. It is estimated by

$$k = \kappa(\hat{\pi}_a = p_a, \hat{\pi}_I = p_I) = \frac{p_a - p_I}{1 - p_I} ,$$

with $p_a = \sum_{i=1}^{I} p_{ii} = \left(\sum_{i=1}^{I} n_{ii}\right)/n$ the observed proportions of actual agreement and $p_I = \sum_{i=1}^{I} p_{i+}p_{+i} = \sum_{i=1}^{I}(n_{i+}n_{+i})/n$ the ML estimate of π_I.

Coefficient k expresses the observed agreement as a proportion of the maximum possible agreement, depending on the marginal distributions. Theoretically, it ranges in the interval $\left[\frac{-p_I}{1-p_I},\ 1\right]$ but in practice,

$$0 \le k \le 1 ,$$

since agreement rarely is worse than expected under independence. Agreement is characterized as "perfect," "good," or "poor," if $k > 0.75$, $0.4 \le k \le 0.75$, or $k < 0.4$, respectively, although the cut points are not clear, depending upon the application area.

For large sample size n, $\hat{\kappa}$ is asymptotically normal distributed $\hat{\kappa} \sim \mathcal{N}\left(\kappa,\ \sigma_k^2\right)$. Based on its estimated asymptotic variance,

Table 9.7 Cross-classification of 120 patients (artificial data) according to their depression severity rating of two independent psychiatrists from negative (D0) to most severe (D3)

Psych.	Psychiatrist B				
A	D0	D1	D2	D3	Total
D0	34	5	3	0	42
D1	1	21	9	2	33
D2	0	2	18	7	27
D3	0	1	4	13	18
Total	35	29	34	22	120

$$\hat{\sigma}_k^2 = \frac{(1-p_a)}{n(1-p_I)^2}\left\{p_I + \frac{2[2p_a p_I - \sum_i p_{ii}(p_{i+}+p_{+i})]}{(1-p_I)} + \frac{(1-p_a)\left[\sum_{i,j} p_{ij}(p_{j+}+p_{+i})^2 - 4p_I^2\right]}{(1-p_I)^2}\right\}$$

(Fleiss et al. (1969)), the asymptotic $100(1-\alpha)\%$ confidence interval for κ can be derived or the level of its strength can be tested (H_0: $\kappa = \kappa_0$ vs. one- or two-sided alternatives, for given κ_0).

9.5.1 Example 9.5

The same patients are classified by two independent psychiatrists regarding the presence and severity of depression, as shown in Table 9.7.

For this data set, $p_a = 0.717$, $p_I = 0.260$, leading to $k = 0.617$ and $\hat{\sigma}_k = 0.056$. The difference between the observed agreement and that expected under independence is about 62% of the maximum possible difference; the asymptotic 95% CI is $(0.508, 0.726)$ indicating moderate to strong agreement between the two psychiatrists.

9.5.2 Agreement on Ordinal Rating Scales

Cohen's κ considers the classification variables as nominal, assuming equally serious the effect on disagreement of all off-diagonal cells, independently from their distance from the diagonal. In most cases, classification is based on an ordinal scale and the disagreement is stronger for larger distances between the ranking categories. The weighted kappa (Spitzer et al. 1967; Cohen 1968) is a measure that weights the disagreements, according to their severity. The system of weights used $(w_{ij})_{I\times I}$ satisfies $0 \le w_{ij} \le 1$, with $w_{ij} = w_{ji}$ and $w_{ii} = 1$, for all $i, j = 1, \ldots, I$. Given the weights, the weighted agreements, actual and under independence, are $\pi_a^w = \sum_{i=1,j} w_{ij}\pi_{ij}$ and $\pi_I^w = \sum_{i,j} w_{ij}\pi_{i+}\pi_{+j}$, respectively. The weighted kappa is then

$$\kappa_w = \frac{\pi_a^w - \pi_I^w}{1 - \pi_I^w}\,. \tag{9.40}$$

A common choice is the uniform spaced weights

$$w_{ij} = 1 - |i-j|/(I-1)\,, \quad i,j = 1,\ldots,I, \tag{9.41}$$

which consider stronger disagreement for cells farther apart from the main diagonal. Another option is the weights by Fleiss and Cohen (1973)

$$w_{ij} = 1 - (i-j)^2/(I-1)^2\,, \quad i,j = 1,\ldots,I. \tag{9.42}$$

Inequalities between Cohen's unweighted kappa and the weighted kappa with weights (9.41) and (9.42) are studied in Warrens (2013).

9.5.3 *Cohen's Kappa in* `R`

Cohen's kappa, weighted and unweighted, can be computed in `R` in many packages. Most of them, like `psych` and `irr`, do not apply directly on the cross-classification table but on the analytical by subject ratings in $n \times 2$ matrix.

A handy function is `Kappa` of library `vcd` that applies on the $I \times I$ table and provides the unweighted and weighted kappa estimates along with their asymptotic standard errors.

Thus, for Table 9.7, k and k_w are computed as

```
> library(vcd);  freq <- c(34,5,3,0,1,21,9,2,0,2,18,7,0,1,4,13)
> depression <- matrix(freq, byrow=T, ncol=4)
> k.est <- Kappa(depression)
```

Under `k.est` is saved

	value	ASE
Unweighted	0.6172249	0.05557287
Weighted	0.7239476	0.12554597

The 95% CI is obtained by function `confint()` of `vcd` as

```
> confint(k.est)
```

Kappa	lwr	upr
Unweighted	0.5083041	0.7261457
Weighted	0.4778820	0.9700131

Argument `level=` of `confint` controls the significance level of the CI.

For k_w, the default weights used in (9.40) are the uniform spaced weights (9.41). Alternatively, the Fleiss–Cohen weights (9.42) can be applied by

```
> Kappa(depression, weights = "Fleiss-Cohen")
```

giving $k_w = 0.816$ and $\hat{\sigma}_{k.w} = 0.108$.

Although measures of agreement are convenient and easy to use and interpret, a model-based analysis of the agreement–disagreement tables is much more informative. Beyond measuring the agreement, in most applications, is of special

interest to detect special patterns of disagreement or draw conclusions with regard to its direction. This is possible through the models QI, QS, T, and D, discussed previously in this chapter. The discussion on interpreting symmetry models applied on the examples in Sects. 9.2.7, 9.3.1, and 9.4.3 is toward this direction.

For Table 9.7, the T model is of very good fit with $G^2(\mathrm{T}) = 3.023$ (p-value=0.696, $df = 5$) and $\hat{\tau}^* = -1.179$, leading to $\hat{\tau} = 0.47$ (for model fit in `R` and derivation of $\hat{\tau}$, see the example in Sect. 9.2.7). This means that psychiatrist A is "looser" in his diagnosis than B. In particular, whenever a disagreement occurs between them, it is 2.13 (=1/0.47) times more probable that psychiatrist B did the more severe diagnosis. The QS model describes also very well these data ($G^2(\mathrm{T}) = 1.012$, p-value=0.798, $df = 3$), highlighting different aspects of the comparison and evaluating marginal inhomogeneity.

9.6 The Bradley–Terry Model

The Bradley–Terry (BT) model is a well-known model used whenever items of a group are repeatedly compared to each other in pairs. Characteristic applications occur in sports, opposing teams' win–losses, and in customers' preference for competitive projects. Let us assume that I items (teams, products, etc.) are cross-classified in pairs and within the pair (i, j) let Π_{ij} be the probability that i "won" (or is preferred to) j with $\Pi_{ij} + \Pi_{ji} = 1$. The model is then defined by

$$\log\left(\frac{\Pi_{ij}}{\Pi_{ji}}\right) = \gamma_i - \gamma_j \ , \quad i, j = 1, \ldots, I \ ,$$

(Bradley and Terry 1952). One of the γ parameters is redundant, say $\gamma_1 = 0$. If $\gamma_i = \gamma_j$, then $\Pi_{ij} = \Pi_{ji}$, while $\Pi_{ij} > \Pi_{ji}$, for $\gamma_i > \gamma_j$. The equivalent model's expression in terms of the preference probability of i over j is

$$\Pi_{ij} = \frac{e^{\gamma_i}}{e^{\gamma_i} + e^{\gamma_j}} \ , \quad i, j = 1, \ldots, I \ .$$

It is connected to the classical QS model. In particular, Fienberg and Larntz (1976) showed that BT is the logit version of the QS model. To see this, consider a probability table (π_{ij}) and let $\pi^*_{ij} = \pi_{ij}/(\pi_{ij} + \pi_{ji})$ be the conditional probability of cell $(i,\ j)$, conditional on the symmetric pairs $(i,\ j)$ and $(j,\ i)$. Then under the QS model, as defined by (9.21), it holds

$$\frac{\pi^*_{ij}}{\pi^*_{ji}} = \frac{\pi_{ij}}{\pi_{ij}} = \frac{a_i}{a_j} \ .$$

Hence, $\Pi_{ij} = \pi^*_{ij}$ for $a_i \propto \exp(\gamma_i)$.

The BT model is strongly related to the choice axiom of Luce (1959); therefore it is also known as the Bradley–Terry–Luce model. Imrey et al. (1976) expressed the BT model as a quasi-independence model. Extensions of the BT model in order to allow for ties have been considered by Glenn and David (1960), Rao and Kupper (1967), and Davidson (1970). Davidson and Beaver (1977) incorporated in the model the within-pair order effect (or home team advantage). Tutz (1986) and Bockenholt and Dillon (1997) proposed extensions for ordered responses. The BT and the Davidson models have been extended for more than one response variables through logistic models by Bockenholt (1988). Su and Zhou (2006) connected the BT to the proportional hazards model (Cox 1972). The literature on the Bradley–Terry model and its generalizations, known as *paired comparison models*, is vast (see David 1988; Train 2009).

In `R`, BT models can be fitted by the `BradleyTerry2` package of Turner and Firth (2012b).

9.7 Overview and Further Reading

Stuart (1955) introduced a test for marginal homogeneity of $I \times I$ tables with the asymptotic null distribution of his test statistic being $\mathscr{X}^2$ with df=$I-1$. The test by Bhapkar (1966, 1979a) is of Wald type and asymptotic equivalent to Stuart's test. Similar is the $\mathscr{X}^2_{I-1}$ test proposed by Madansky (1963), who maximized the log-likelihood kernel subject to the marginal homogeneity constraints and handled the maximization problem by a gradient method (see Bishop et al. 1975, pp. 294–295). Ireland et al. (1969) discussed another $\mathscr{X}^2_{I-1}$ test, yielding an iterative estimation of the expected cell frequencies under marginal homogeneity by a modified version of the minimum discrimination information estimation. All these standard tests do not take into consideration the possible ordering of the classification variables' categories. Agresti (1983c) proposed alternative approaches that make use of category order and showed by power comparisons for certain common alternatives that they tend to yield more powerful tests. McCullagh (1977) developed statistics to measure the lack of marginal homogeneity in square tables for paired data based on global odds ratios and a logistic model.

The model of quasi symmetry (QS) was introduced by Caussinus (1965) and is presented in detail in the classical reference of Bishop et al. (1975). McCullagh (1978) connected the QS model to Markov chains; the transition probabilities matrix of a Markov chain is quasi symmetric. An application-orientated presentation of the QS model is provided in McCullagh (1982). Fienberg and Larntz (1976) proved that the Bradley–Terry model can be defined as the logit version of the QS model. Agresti (1993) showed that generalized Rasch models for ordinal response items are related to quasi-symmetric log-linear models with diagonal parameters. Bavaud (2002) provided an interesting approach of quasi symmetry and connected it to gravity model. For an overview and connections to other family of models, see Goodman (2002a).

The QS model exhibits a similar property to that of association models. In particular, QS can be viewed as a departure from complete symmetry model, and it can be proved that under certain conditions, this is the closest to complete symmetry in terms of the Kullback–Leibler divergence (Kateri and Papaioannou 1997). This leads to the generation of a class of quasi-symmetry models, based on ϕ-divergence, which includes the standard QS model as a special case. The ordinal QS model can be generalized toward this direction as well (Kateri and Agresti 2007).

9.7.1 Mobility Tables and the Mover–Stayer Model

The analysis of mobility tables relies mainly on log-linear and association models with or without special treatment for the diagonal entries (see Goodman 1979d). Characteristic examples of social and class mobility tables are Tables 4 and 6, respectively, in Breen (2008). Models for mobility tables have been extended to $I \times I \times K$ tables, modeling mobility in different layers (Xie 1992). Mobility tables have also been analyzed by modeling simultaneously their joint and marginal distributions (Becker 1994; Lang and Agresti 1994; Lang and Eliason 1997; Sobel et al. 1998).

Well known for mobility tables is the Mover–Stayer (*MS*) model. The MS model assumes that there are two types of individuals: the "stayer," who remains in the same category during the entire period of study, and the "mover," who changes categories over time and whose moves are described by a Markov chain of constant transition probabilities matrix. MS was initially introduced in the context of industrial mobility tables by Blumen et al. (1955) and further developed, among others, by Goodman (1961), Spilerman (1972), Frydman (1984), and Fuchs and Greenhouse (1988) and for panel data by Cook et al. (2002).

Extensions of the association models to handle square tables with diagonal values of strong effect, like mobility tables, have been presented in Sect. 9.4.1. The adjustment of correspondence analysis (CA) to apply to such tables has been considered by Greenacre (2000) and is based on the orthogonal decomposition of the data table into two tables, a symmetric and a skew symmetric, before its analysis by CA. This decomposition is due to Gower (1977) and Constantine and Gower (1978).

9.7.2 Measuring Agreement

Though Cohen's kappa is the predominant measure of rater agreement, alternative measures have been proposed. Hutchinson (1993), for example, argued that the tetrachoric correlation coefficient is preferable to Cohen's kappa, because the latter is sensitive in different placing of the boundaries between categories. For a review on measures of agreement, see Banerjee et al. (1999). See also Schuster (2004) and Fay (2005).

References on measures of agreement among multiple raters include Fleiss (1971), Landis and Koch (1977), James (1983), Kraemer (1980), Janson and Olsson (2004), and Schuster (2004).

Beyond criticism and contradictions between measures, an assessment of agreement based only on measures can lead to severe loss of information. Furthermore, a model-based approach is extended naturally to problems of multi-rater agreement.

Special log-linear, association, or symmetry models for modeling rater agreement were proposed by Goodman (1979c), Tanner and Young (1985), Darroch and McCloud (1986), Agresti (1988), Becker and Agresti (1992), Agresti and Lang (1993), Perkins and Becker (2002), and Valet et al. (2007). Schuster (2002) proposed a mixture model for rater agreement that includes the models of Tanner and Young (1985) and Agresti (1988), among other, as special cases. Agresti and Lang (1993) and Yang and Becker (1997) modeled agreement by latent class models. A logistic regression model, adjusting for covariates, has been proposed by Barlow (1996).

9.7.3 Symmetry Models for Multi-way Tables

The models of symmetry, quasi symmetry, and marginal homogeneity have been generalized to higher-order contingency tables I^k, $k > 2$, mainly by Bhapkar (1979a,b) and Bhapkar and Darroch (1990) and Lovison (2000). With respect to the conditional symmetry model T, Read (1978) and Sobel (1988) considered the conditional within levels T symmetry model for K stratified $I \times I$ contingency tables by applying the two-way T model on each of the K partial square tables. A generalization of the concept of conditional symmetry to three-way tables and definition of conditional symmetry models of first and second order for three-way tables is provided by Kateri and Dellaportas (2012). These models retain all desirable properties with respect to connections between the three-way conditional symmetry models as well as to the symmetry and marginal homogeneity models for three-way tables, i.e., extensions of property (9.25).

9.7.4 Clustered Categorical Data

We have seen that square tables with commensurable ordinal classification variables occur often in panel studies of two occasions. Such correlated ordinal data can be modeled through global odds ratios based on a time- and subject-dependent covariate vector by the maximum likelihood method (Molenberghs and Lesaffre 1994) or by the GEE approach (Williamson et al. 1995), both methods being also applicable to studies where the response variable is measured more than twice. Williamson and Kim (1996) propose a regression model for bivariate correlated ordinal data that associates the outcome data with grouping variables and other covariates.

Clustered categorical data of two or more occasions occur often in longitudinal studies. They are treated either through marginal models (see also Sect. 5.7.2) or random effect models. Marginal models treat the pairwise among cluster-level dependence structure as nuisance while random effect models include also cluster-specific effects that are random. They are applied through the *generalized linear mixed models* (GLMM), i.e., GLMs are extended to include also random effects. Characteristic related references are Breslow and Clayton (1993) and Hedeker and Gibbons (1994). Furthermore, in repeated categorical measures analysis, the transition (or Markov) models are also an option, when previous responses are considered as predictors (see Kalbfleisch and Lawless (1985) and references cited therein).

The estimation method used in marginal models is maximum likelihood (see Fitzmaurice and Laird 1993; Bergsma and Rudas 2002a). The weighted least squares (WLS) of procedure of Grizzle et al. (1969) is an alternative. For repeated categorical measures it has been applied by Koch et al. (1977) and Landis and Koch (1979). Though WLS is computationally simpler than the ML, the WLS estimators do not share the nice statistical properties of the MLEs.

Liang and Zeger (1986) introduced the generalized estimating equations (GEE) method, a quasi-likelihood approach, that overcomes the computational difficulties of the MLEs while leads to estimators of good properties. The GEE method was further developed by Lipsitz et al. (1994) and Diggle et al. (2002). Touloumis et al. (2013) developed a GEE approach for correlated nominal or ordinal multinomial responses using a local odds ratio parameterization. This approach is implemented in the `R` package `multgee` (Touloumis 2012). Other fundamental contributions to marginal modeling deal with response variables (Glonek and McCullagh 1995), ordinal variables (Colombi and Forcina 2001; Forcina and Dardanoni 2008), or panel data (Croon et al. 2000; Vermunt et al. 2001). Marginal models have been considered for medical applications, among others, by Balagtas et al. (1995) and Molenberghs and Lesaffre (1999).

A comprehensive reference book on modeling clustered categorical data is Molenberghs and Verbeke (2005), while they are extensively treated also in Agresti (2010, 2013). For a detailed insight to marginal models for clustered categorical data and the underlying theory we refer to Bergsma et al. (2009). An `R` package for fitting marginal models is `hmmm` of Colombi et al. (2013).

Chapter 10
Further Topics

Abstract This epilogue chapter refers briefly to alternative methods and approaches in the analysis of contingency tables (latent class models, graphical models, and smoothing), not covered in the book. Furthermore, a bibliography on small sample inference, Bayesian inference, and the analysis of high-dimensional sparse contingency tables is discussed.

Keywords Association measures • Latent class models • Graphical models • Small sample inference • Bayesian inference

10.1 Overview

The focus of this book is on model-based approaches, so we did not refer to association measures other than the odds ratio which was in the kernel of the models developed. Measures of association however played an important role in the early development of categorical data analysis and continue to be of special interest in areas of social sciences and psychology. For this, an overview of the related literature is provided in Sect. 10.2 below.

The predominant modeling approach for contingency tables is log-linear models based. Alternative approaches of contingency table analysis with links to log-linear models are mentioned in Sect. 10.3.

Our approach was the classical frequentist approach, assuming large samples and non-sparse situations so that standard asymptotic theory applies. It is often the case that small samples occur. Sect. 10.4 refers to methods for analysis of small samples. Furthermore the Bayesian analysis of contingency tables is an attractive alternative in situations the asymptotic assumptions are not met. Beyond small sample setups, the Bayesian method is interesting for giving the option of incorporating prior information upon availability. Section 10.5 is devoted to Bayesian methods and applications for contingency tables.

M. Kateri, *Contingency Table Analysis: Methods and Implementation Using R*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4811-4_10,

Finally, clustering of categorical data has gained the last years renewed interest due to huge data sets and the need to organize their presentation and detect association structures. Huge data sets are normally extremely high dimensional. Bibliography on analyzing extreme high-dimensional categorical data sets and clustering techniques is to be found in Sect. 10.6.

10.2 On Measures of Association

Upon rejection of independence, information on the significant underlying relationship is traditionally summarized through *measures of association*. These measures differentiate for nominal and ordinal data. Measures for nominal variables refer only to the strength of the association while for ordinal variables, they incorporate information about the direction (positive or negative) of the association as well. Their values range preferably in the [0, 1] and [−1, 1] intervals, for the nominal and ordinal measures, respectively, with their absolute value being increasing in the strength of the association. Special interest has been attracted by measures for binary cross-classifications.

The interest in defining and measuring association in $I \times J$ contingency tables dates back to the 1840s and the related bibliography is enormously rich, motivated from diverse scientific fields, like social sciences, education, and meteorology. The history of their development, interesting and often generating controversies, is reviewed exhaustively in the book by Goodman and Kruskal (1979), which is the most classical reference on association measures. This book republishes their four *JASA* papers of Goodman and Kruskal in 1954, 1959, 1963, and 1972. In the first two papers, Goodman and Kruskal organized existing measures, presenting them unified. Furthermore, they focused on the general $I \times J$ table and developed new measures, for nominal and ordinal variables. They suggested measures taking into account the existence of a response variable in the table, introducing symmetric and asymmetric versions of their measures. In their last two papers they proved the asymptotic normality of their measures, making thus asymptotic inference feasible.

Association measures are organized in classes accordingly to their basis of formulation. They can be based on the X^2 statistic for independence, on the assumption of an underlying joint normal distribution, or they can be based on a probabilistic model (like measures based on pairs of observations or on scores assigned to the categories of the classification variables). The measures based on the X^2 statistic, as the contingency coefficient ϕ, the association coefficient C, and Cramer's V, are indicative only about the strength of the underlying association.

Focusing on 2×2 contingency tables, the odds ratio θ, $\theta \in [0, +\infty)$, introduced by Yule (1900, 1903, 1912), is undoubtedly the "gold standard" of measures. Yule proposed transformations of θ, the Q and Y, that range in [−1, 1]. A competitor of Yule's odds ratio for measuring association in 2×2 tables is the *tetrachoric correlation coefficient* of Pearson (1900b, 1904, 1913), a product–moment correlation between two unobserved quantitative variables that have been dichotomized.

The tetrachoric correlation opposes the odds ratio by this underlying continuum assumption. Pearson preferred to view contingency tables as discretizations of underlying multivariate normal distributions (Pearson and Heron 1913) while for Yule contingency tables were formed by discrete variables of fixed categories. The dispute between Yule and Pearson was long and strong. Over the years both approaches had their supporters and opposers. Edwards (1963) argued that the odds ratio should be the basis of an association measure for 2×2 tables. On the other hand, when the classification variables are of continuous nature and dichotomized at a certain cut point, as it is often the case in psychometrics, then the tetrachoric correlation is preferable. The tetrachoric correlation coefficient and related inference problems are discussed analytically by Bonett and Price (2005). Transformations of the odds ratio that approximate the tetrachoric correlation have been proposed by Digby (1983) and Becker and Clogg (1988). Bonett and Price (2007) proposed a generalized Yule coefficient that is similar in value to the contingency coefficient ϕ and has, among others, Yule's Q and the coefficient of Digby (1983) as special cases. A review of association coefficients for 2×2 tables with focus on their general properties is provided by Warrens (2008). Overall, the odds ratio predominated, also due to its connection to log-linear models. The odds ratio is the precursor of log-linear models and Yule can be considered as their "father." For an overview of the odds ratios and their role in contingency table analysis, see Rudas (1998).

For $I \times J$ contingency tables, the tetrachoric correlation was extended to the *polychoric correlation* by Lancaster and Hamdan (1964) while measures, generalizations of the odds ratio, were proposed by Altham (1970) for nominal association and by Agresti (1980) and Edwardes and Baltzan (2000) for ordinal. Cumulative odds ratios for ordinal variables have been considered earlier by Clayton (1974), who developed statistics based on them to summarize the difference in location between two distributions of an ordered categorical variable and for describing association between two such variables. The approach was generalized also in case some observations were subject to censorship (Clayton 1976).

For ordinal association, famous are Goodman and Kruskal's λ, λ_a, λ_b and γ. Kendall (1938, 1948) introduced a measure of rank correlation, known as Kendall's τ, and its asymmetric version τ_b. Stuart (1953) proposed a measure of association for contingency tables, the τ_c, based on Kendall's τ, which is compared to Goodman and Kruskal's γ in Hamdan (1977). Another popular measure of ordinal association is Somers' d (Somers 1962). The measures of *raters' agreement* consist a special category of association measures, presented in Sect. 9.5.2.

Different measures refer to different types of association; thus they should not be used blindly and routinely. The choice of measure should take into account the nature of the contingency table under consideration and be compatible with its distribution. Efficacies of the measures of association for ordinal contingency tables are discussed in Simon (1978). Contingency tables are better treated through model-based approaches, which are flexible regarding the structure imposed on the association and meanwhile are more informative, providing cell estimates under the assumed model. Since our approach is model focused, we will not discuss any further measures of association. For ordinal measures of association we refer to Agresti (2010, Chap. 7).

10.3 Alternative Approaches in Contingency Table Analysis

10.3.1 Latent Class Models

The classical latent class model is defined as a finite mixture of unobserved (latent) multinomial distributions, each of which exhibits statistical independence. Latent class models play an important role in multivariate data analysis and receive special attention in psychology and social sciences. They were first treated by Lazarsfeld (1950). For their connection to log-linear models, see Goodman (1974), Haberman (1979), Heinen (1996), and Hagenaars (1998). Formann (1992) connected latent class models to polytomous logistic models and Gilula (1979, 1984) to correspondence analysis. Goodman (1987) provides a nice overview of the connection between the approaches of CA, latent class analysis, and log-linear and association models. The connection of association models to latent class models is also discussed in Anderson and Vermunt (2000). Their link to models for rater agreement is due to Uebersax (1993), Uebersax and Grove (1990, 1993), and Yang and Becker (1997). Becker and Yang (1998) discuss latent class models for modeling marginal associations in contingency tables and provide an extended literature review on analysis of contingency tables by latent class models.

The volume *Applied Latent Class Analysis*, edited by Hagenaars and McCutcheon (Cambridge University Press, 2002), provides an interesting collection of papers on traditional latent class analysis as well as in connection to special topics as clustering, logistic regression, longitudinal data, missing data, and nonresponse (see, e.g., Goodman 2002b). For an updated source on latent variable models and their applications, we refer to Bartholomew et al. (2011). An overview on latent variable models for categorical responses is provided by Agresti and Kateri (2014).

10.3.2 Graphical Models

The graphical log-linear models and their role in the analysis of high-dimensional contingency tables are discussed in Sects. 4.7.2, 4.9.3, and 4.9.4. Beyond log-linear models, conditional independence graphs of a multi-way contingency tables are connected to the RC model and to correspondence analysis by de Falguerolles et al. (1995). Bounds on the cell counts of contingency tables for decomposable log-linear models and their related graphs are proposed by Dobra and Fienberg (2000, 2003). Bidirected graphical models of marginal independencies for categorical data are discussed in Lupparelli et al. (2009). Quasi-symmetry models are presented as graphical models by Gottard et al. (2011).

10.3.3 Smoothing Categorical Data

Alternative to modeling, nonparametric approaches can be adopted for analyzing contingency tables. The oldest and most well-known is the partitioning of the X^2 statistic for testing independence (see Sect. 2.5.4). Furthermore, smoothing methods bridge the gap between the parametric and nonparametric approaches, from strict assumptions to no assumptions at all. Smoothing methods for ordinal data were studied and compared by Titterington and Bowman (1985), who organized them into three major groups, the kernel-based methods (Aitchison and Aitken 1976), Bayesian-based methods (Leonard 1975), and the penalized minimum distance methods, which relate to maximizing the penalized likelihood (Scott et al. 1980). Smoothing methods are appropriate for the analysis of large sparse contingency tables (Simonoff 1983). The application of smoothing to categorical data analysis is reviewed in Simonoff (1995, 1998). For a literature review on the smoothing methods for categorical data, see the references cited in Titterington and Bowman (1985) and Simonoff (1995).

Coull and Agresti (2003) introduced a generalized log-linear model with random effects, useful for smoothing large sparse contingency tables by maximizing a penalized likelihood. The structures considered for the random effects mimic Goodman's association models. Geenens and Simar (2010) propose two nonparametric tests for testing independence of two categorical cross-classified variables, conditional on a set of explanatory variables, based on kernel estimation of the conditional probabilities.

10.4 Small Sample Inference for Contingency Tables

Starting with Fisher's exact test (see Sect. 2.1.7) for testing independence for 2×2 contingency tables with small cell entries, the development of methods and algorithms for exact inference for a variety of models, applied on two- and multi-way contingency tables, has a long history. In the early years, Yule's correction was quite popular, while it was criticized later. Despite this, adding a constant to the cells of a contingency table to avoid problems from small or zero count cells is still common in practice. Greenland (2010) argued against this practice, showing in the Bayesian framework that it can lead to a form of Simpson's paradox, and proposed more sophisticated methods of smoothing. In case of high-dimensional contingency tables, sparseness problems occur even for moderate to large sample sizes, leading to inferential implications (see also Sect. 10.6).

Fundamental is the network algorithm of Mehta and Patel (1983) for sampling from an $I \times J$ contingency table with given marginals, which served in extending the Fisher's exact test to tables of higher size. Sequent results by them with coauthors established the exact analysis of contingency tables. For example, Agresti et al. (1990) extended this algorithm for the exact analysis of two-way contingency

tables with ordinal classification variables. Basic summarizing reference is Mehta and Patel (1995). Exact analysis of models for binary response is provided in Cox (1970a) and Cox and Snell (1989). Exact conditional tests for testing quasi-independence in incomplete tables are considered by McDonald and Smith (1995). A survey on exact inference for categorical data is provided in the discussion paper of Agresti (1992) and in Agresti (2001). A detailed treatment is provided in the book by Hirji (2006).

Small sample inference can be developed based on Markov chain, as in Forster et al. (1996), or bootstrap algorithms. For applications of bootstrap methods on categorical data, we refer to Jhun and Jeong (2000) and Amiri and von Rosen (2011). Model-based bootstrap tests for independence in two-way tables are considered in Pettersson (2002) and Jeong et al. (2005). Bootstrapping for log-linear models in large, sparse contingency tables has been considered by Sauermann (1989). Alternatively to log-linear models, Streitberg (1999) developed a bootstrap approach for analyzing interactions in high-dimensional tables, based on the additive approach (see Darroch and Speed (1983) and references therein).

The problem of studying the exact distribution in a contingency table under a model assumption can also be faced through algebraic statistics, based on the pioneer work by Diaconis and Sturmfels (1998). They proposed an algorithm for sampling from a set of tables with given marginals, based on Markov bases computation, which is achieved by finding a Gröbner basis. Aoki and Takemura (2005) and Rapallo (2003, 2006) derive Gröbner bases for some classical log-linear models taking structural zeros into account. Dobra (2003) applied graphical models to identify special settings which lead to reduction of the required computations for the identification of a Markov basis. Dobra et al. (2009) dealt with the maximum likelihood estimation for log-linear models and a related disclosure limitation problem, focusing on the disclosure of small cell counts to protect the confidentiality of individual responses. Hara et al. (2012) proposed a new class of models for the analysis of multi-way contingency tables, more parsimonious than the usual hierarchical log-linear models, by modeling the interaction terms in each maximal compact component of a hierarchical model. They proceed to exact tests via Markov bases while their approach considers also the presence of structural zeros.

Exact inference for the model of symmetry for square contingency tables based on Diaconis and Sturmfels's algorithm is provided by Rapallo (2003) and Krampe and Kuhnt (2007). In the context of rater agreement, Rapallo (2005) provides algebraic testing procedures for Cohen's kappa, the quasi symmetry, and quasi-independence model. Krampe et al. (2011) develop algebraic tests for the models of conditional, diagonal, and ordinal quasi symmetry.

Alternatively, inferential problems due to sparseness, sampling zeros, or small frequencies can be treated in the Bayesian analysis framework.

10.5 Bayesian Analysis of Contingency Tables

Bayesian analysis can be the solution in situations of small samples or sparse tables, where standard asymptotic inference does not apply. Furthermore, the incorporation of prior inference can be essential in some applications' areas. The model selection procedure is benefited in the Bayesian framework. MCMC methods enable an efficient search of the model space even if it is large. For the models visited, the associated algorithm provides the posterior model probabilities, a powerful tool for models' evaluation and estimation of their uncertainty. These issues are of great importance in high multidimensional contingency tables. Model uncertainty might be high in small samples or in existence of more than one models of similar performance. High model uncertainty can be incorporated in the Bayesian statistical inference.

Early attempts on Bayesian analysis of categorical data go back to the 1950s and were based on conjugate prior analysis. Good (1956) proposed smoothing proportions in contingency tables while his approach for hierarchical Bayesian inference (Good 1965) is related to the early work by Johnson in the 1920s on the Dirichlet priors for the multinomial distribution (see Fienberg (2006) for a detailed discussion on the early and key Bayesian developments). Lindley (1964) focused on the Bayesian analysis of contingency tables (two- and three-way) and developed the Bayesian inference for the odds ratio. Altham (1969, 1971) dealt with the Bayesian analysis of 2×2 tables for small samples based on conjugate priors. Since then, the development of the Bayesian approach was rapid, mainly due to the progress of computer-intensive numerical methods for the evaluation of posterior distributions, which made the Bayesian analysis and Bayesian model selection of multidimensional problems and complex models feasible. For an overview, see Congdon (2005) and the review paper by Agresti and Hitchcock (2005).

The Bayesian analysis of log-linear models with non-conjugate priors originates from Leonard (1975) and Laird (1978). They proposed univariate normal priors for the parameters of the saturated model. Knuiman and Speed (1988) and King and Brooks (2001) considered multivariate normal prior for the parameter vector and extended the approach to multi-way contingency tables. In contingency tables' framework, the model fit evaluation through the Bayes factor has been considered by Spiegelhalter and Smith (1982), Raftery (1986), and Albert (1997). Issues for the Bayesian analysis of the 2×2 table are discussed in Howard (1998). For Bayesian log-linear model selection we refer to Dellaportas and Forster (1999) and Ntzoufras et al. (2000). The Bayesian analysis of log-linear models is reviewed in Forster (2010). Consonni and Pistone (2007) considered the Bayesian analysis of contingency tables with structural zeros based on algebraic statistics. The basic reference on Bayesian logit models is Albert and Chib (1993).

The Bayesian analysis of the simple U association model has been considered by Agresti and Chuang (1989). They imposed a Dirichlet prior distribution on the probability table $\pi = (\pi_{ij})_{I \times J}$. The prior mean was assigned from the U model. Alternatively to the conjugate prior-type analysis they proposed the Bayesian

log-linear analysis by considering independent uniform priors for the main effect parameters λ_i^X and λ_j^Y and normal priors for the interaction parameters $\lambda_{ij}^{XY} \sim N(\varphi \propto_i \nu_j, \sigma^2)$.

The first attempt for fitting the RC association model in the Bayesian framework was due to Chuang (1982). He set independent uniform priors on the main effect parameters λ_i^X and λ_j^Y and normal priors on the parametric row and column scores $\propto_i \sim \mathcal{N}(0, \sigma_1^2)$, $\nu_j \sim \mathcal{N}(0, \sigma_2^2)$ and proceeded with empirical variance estimation. Evans et al. (1993) adopted a different approach for the Bayesian analysis of the RC model. They based their analysis on the Bayesian estimation of the saturated log-linear model with normal priors on all its parameters and then concluded to the posterior for the RC by Euclidean projection from the posterior of the saturated log-linear model. Further, they studied the posterior distribution of the Euclidean distance between the interaction matrices of the saturated and the RC models, (λ_{ij}^{XY}) and $(\varphi \propto_i \nu_j)$, respectively. Finally, Bayesian inference for the more general RC(M) association model has been developed by Kateri et al. (2005). This procedure can be also applied for fitting the RC model (for $M = 1$). Albert (1997) provided an interesting Bayesian approach for testing the fit of simple models such as independence, quasi-independence, and uniform association models, as well as for modeling outliers via mixture models.

With respect to merging categories and the associated role of the association models (as discussed in Sect. 7.5), in the Bayesian framework, Tarantola et al. (2008) used methodology adopted from product partition models to make inferences about the clustering of scores in the row effect model. For the two-group comparison of an ordinal scale, Kateri and Agresti (2013) discussed stochastic orderings, based on generalized odds ratios for ordinal responses for $2 \times J$ contingency tables, from the Bayesian point of view.

We referred in Sect. 6.8.2 to the order-restricted inference for association models, through large sample asymptotic methods. The Bayesian inference for association models with order-constrained parametric scores has been developed by Iliopoulos et al. (2007). Their approach for identifying possible score equalities was based on calculating the posterior probabilities of possible order violations for successive categories in the unrestricted model. These probabilities were used in an isotonic regression-type logic, indicating which scores should be merged. Furthermore, the deviance information criterion (DIC, Spiegelhalter et al. 2002) was applied to identify the most appropriate model in terms of goodness of fit. However, this approach forms not a formal Bayesian evaluation in favor or against merging specific scores, since it is not based on the posterior model odds and probabilities (for details, see Kass and Raftery 1995). Toward this direction, Iliopoulos et al. (2009) proposed an alternative approach for this problem, focusing on the estimation of posterior model probabilities of the RC order-constrained model, in a full Bayesian way, by allowing for ties in the prior distribution level. They constructed a *trans*-dimensional MCMC algorithm (reversible jump MCMC, Green 1995) for assessing the equality of successive row and column scores.

For Bayesian graphical models we refer to Madigan and Raftery (1994) and Madigan and York (1995). Massam et al. (2009) developed a family of conjugate prior for a class of discrete hierarchical log-linear models for multi-way tables, with graphical models being in this class. Webb and Forster (2008) dealt with Bayesian graphical model selection for multivariate ordinal data. Ng et al. (2008) provided a conjugate Bayesian analysis of incomplete contingency tables based on a new family of distributions, the grouped Dirichlet distributions, which includes the classical Dirichlet distribution as special case.

10.6 Extreme High-Dimensional Categorical Data

High-dimensional contingency tables often lead to sparseness and related inferential discrepancies. Approaches for high-dimensional data discussed in Hastie et al. (2009) and Bühlmann and van de Geer (2011) apply also on contingency tables. In particular, Dahinden et al. (2007) extended the lasso algorithm, to the group lasso, in order to fit log-linear models for high-dimensional and sparse data arising in computational biology. An alternative approach based on graphical models is given by Dahinden et al. (2010). The lasso penalty for high-dimensional GLMs is considered by van de Geer (2008).

In high-dimensional problems the clustering of the subjects or items under study becomes often an important issue. This way customers, patients, or genes, for example, can be assigned to groups of similar profile (with respect to some characteristics). Clustering methods are based on measuring the dissimilarity between items with respect to their characteristics, captured in variables. Most of the clustering algorithms refer to continuous variables (see, e.g., Everitt et al. 2011). Bock (1986) developed clustering methods for categorical data, based on a logistic or log-linear models probability distribution. For an overview on clustering methods that apply also on categorical data, see also in van Mechelen et al. (2004). On clustering of categorical data, refer to Agresti (2013, Sect. 15.3).

Appendix A
Appendix: Contingency Table Analysis in Practice

A.1 Software for Categorical Data Analysis

The free software R, for statistical computing and graphics, is of increasing popularity and usage (R web site: http://www.r-project.org/). Many researchers support their published papers with the related R code. This way, R software is continuously updated and one can find a variety of functions for basic or advanced analysis of categorical data and special types of them. R language and environment is similar to S and code written for S-Plus runs usually under R as well. Furthermore, standard statistical packages, such as SAS, SPSS, and Stata, are well supplied to treat categorical data. Especially in their updated versions, their features concerning categorical data analysis are enriched. They incorporate procedures for applying the recently developed methods and models in categorical data analysis following the new computing strategies. Briefly, one could say that their major new features concern mainly options for exact analysis and analysis of repeated categorical data. Thus, NLMIXED of SAS fits generalized linear mixed models while GEE analysis for marginal models can be performed in GENMOD. SPSS offers the "generalized estimating equations" sub-option under the "GLMs" option. The related R function is `gee()`.

For categorical data analysis with SAS, we refer to Stokes et al. (2012) while a variety of SAS codes are presented and discussed in the Appendix of Agresti (2007, 2013). Advanced models are fitted in R using special functions, developed individually, and included in different libraries. Orientated toward categorical data analysis and models for ordinal data as well are the libraries MASS (Venables and Replay) and VGLM, VGAM developed by Yee (2008). For example, generalized linear mixed models can be fitted through the `glmmPQL()` function of the MASS library.

Other software, as BMDP, Minitab, and SYSTAT, have also components for categorical data inference.

M. Kateri, *Contingency Table Analysis: Methods and Implementation Using R*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4811-4,

Bayesian analysis of categorical data can be carried out through WINBUGS (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml), which is a free software. Another option is to perform categorical data analysis through MATLAB, as Johnson and Albert (2000). The MATLAB functions they used are described in their Appendix.

For categorical data analysis, there have been developed also some special packages. Thus, exact analysis of categorical data is performed by StatXact while exact conditional logistic regression can be fitted by LogXact. SUDAAN is specialized for analysis of mixed data from stratified multistage cluster designs. It has also the feature of analyzing marginal models for nominal and ordinal responses by GEE. Software tool for estimating marginal regression models is also MAREG.

Finally, some algorithms may be found in Fortran. For example, Haberman (1995) provided a Fortran program for fitting the association model RC(K) by the Newton–Raphson method while Ait-Sidi-Allal et al. (2004) implemented their algorithms for estimating parameters in association and correlation models also in Fortran.

A.2 Contingency Table Analysis with R

All procedures and models discussed in this book are worked out in R, in a fashion aiming that even readers not familiar with R will be able to apply in practice all the models discussed here, even the nontrivial ones, fast and directly. A web companion of the book serves this goal. This section of the Appendix is basically the content description of the web companion of the book, to be found under

http://cta.isw.rwth-aachen.de

A.2.1 R Packages for Contingency Table Analysis

An extensive list of special R packages, useful in the analysis of contingency tables, is provided in the web appendix.

A.2.2 Data Input in R

Alternative forms of defining contingency tables data in R are presented (`matrix()`, `array()`, and `data.frame()`) and transformations from one type to another are illustrated. Ways of entering or reading data are discussed.

A.3 R Functions Used

The R functions constructed for the descriptive and inferential needs of this book are given in the corresponding section of the web appendix, organized by chapter of their first use.

A.3.1 R Functions of Chap. 1

- Binomial–Normal Distribution Graph: `bin_norm( )`

A.3.2 R Functions of Chap. 2

- Likelihood Ratio Statistic for Testing Independence in Two-way Contingency Tables: `G2( )`
- Odds Ratio for a 2×2 Table: `odds.ratio( )`
- Local Odds Ratios for an $I \times J$ Table: `local.odds.DM( )`
- Global Odds Ratios for an $I \times J$ Table: `global.odds.DM( )`
- Cumulative Odds Ratios for an $I \times J$ Table: `cum.odds.DM( )`
- Continuation Odds Ratios for an $I \times J$ Table: `cont.odds.DM( )`
- Linear Trend Test: `linear.trend( )`
- Midrank Scores Computation: `midrank( )`
- Fourfold Plots for the Local Odds Ratios of an $I \times J$ Table: `ffold.local( )`

A.3.3 R Functions of Chap. 3

- Breslow–Day–Tarone Test of Homogeneous Association: `BDT( )`
- Woolf's Test of Homogeneous Association: `woolf( )`

A.3.4 R Functions of Chap. 5

- Independence (I) Model for Two-way Contingency Tables: `fit.I( )`
- Quasi-Independence (QI) Model for Two-way Contingency Tables: `fit.QI( )`

A.3.5 R Functions of Chap. 6

- Scores' Rescaling to Obey the Weighted Constraints (6.17): `rescale( )`
- Uniform (U) Association Model: `fit.U( )`
- Row Effect (R) Association Model: `fit.R( )`
- Column Effect (C) Association Model: `fit.C( )`
- Row–Column (RC) Association Model: `fit.RC( )`
- RC(M) Association Model: `fit.RCm( )`
- Plotting the Row and Column Scores in Two Dimensions: `plot_2dim( )`

A.3.6 R Functions of Chap. 9

- $(1-a)100\%$ Asymptotic Confidence Interval for the Difference of Correlated Proportions: `McNemar.CI( )`
- Factors Needed to Fit Symmetry Models on an $I \times I$ Table in `glm`: `SYMV( )`
- Scores' Rescaling to Satisfy Constraints (9.38): `rescale.square( )`

A.4 Contingency Table Analysis with SPSS

The association and symmetry models cannot be fitted directly in SPSS through the options of the windows commands. Association models that are GLM can be fitted through the GLM option by defining the appropriate vectors, as explained in Sect. 6.6. For all two-way association models (RC(M) included, which is nonlinear and thus cannot be fitted in GLM) and the symmetry models, we provide appropriate syntax codes to be fitted in SPSS MATRIX.

In particular, we provide MATRIX codes for:

- Independence for two-way tables using SPSS MATRIX
- Association models for two-way tables
  (uniform (U), row effect (R), column effect (C), and RC (M) association models)
- Symmetry models

References

Agresti, A.: Generalized odds ratios for ordinal data. Biometrics **36**, 59–67 (1980)

Agresti, A.: A survey of strategies for modeling cross–classifications having ordinal variables. J. Am. Stat. Assoc. **78**, 184–198 (1983a)

Agresti, A.: A simple diagonals–parameters symmetry and quasi-symmetry model. Stat. Probab. Lett. **1**, 313–316 (1983b)

Agresti, A.: Testing marginal homogeneity for ordinal categorical variables. Biometrics **39**, 505–510 (1983c)

Agresti, A.: Applying R^2-type measures to ordered categorical data. Technometrics **28**, 133–138 (1986)

Agresti, A.: A model for agreement between ratings on an ordinal scale. Biometrics **44**, 539–548 (1988)

Agresti, A.: A survey of exact inference for contingency tables (with discussion). Stat. Sci. **7**, 131–177 (1992)

Agresti, A.: Computing conditional maximum likelihood estimates for generalized Rasch models using simple loglinear models with diagonal parameters. Scand. J. Stat. **20**, 63–71 (1993)

Agresti, A.: On logit confidence intervals for the odds ratio with small samples. Biometrics **55**, 597–602 (1999)

Agresti, A.: Exact inference for categorical data: recent advances and continuing controversies. Stat. Med. **20**, 2709–2722 (2001)

Agresti, A.: Dealing with discreteness: making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact. Stat. Meth. Med. Res. **12**, 3–21 (2003)

Agresti, A:. An Introduction to Categorical Data Analysis, 2nd edn. Wiley, New York (2007)

Agresti, A.: Analysis of Ordinal Categorical Data, 2nd edn. Wiley, New York (2010)

Agresti, A.: Categorical Data Analysis, 3rd edn. Wiley, Hoboken (2013)

Agresti, A., Chuang, C.: Model-based Bayesian methods for estimating cell proportions in cross-classification tables having ordered categories. Comput. Stat. Data Anal. **7**, 245–258 (1989)

Agresti, A., Coull, B.A.: Approximate is better than 'exact' for interval estimation of binomial proportions. Am. Stat. **52**, 119–126 (1998)

Agresti, A., Coull, B.A.: The analysis of contingency tables under inequality constraints. J. Stat. Plann. Infer. **107**, 45–73 (2002)

Agresti, A., Hartzel, J.: Strategies for comparing treatments on a binary response with multi-centre data. Stat. Med. **19**, 1115–1139 (2000)

Agresti, A., Hitchcock, D.B.: Bayesian inference for categorical data analysis. Stat. Methods Appl. **14**, 297–330 (2005)

Agresti, A., Kateri, M.: Some remarks on latent variable models in categorical data analysis. Commun. Stat. Theory Meth. **43**, 1–14 (2014)

M. Kateri, *Contingency Table Analysis: Methods and Implementation Using R*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4811-4,

Agresti, A., Kezouh, A.: Association models for multidimensional cross-classifications of ordinal variables. Commun. Stat. Ser. A **12**, 1261–1276 (1983)
Agresti, A., Lang, J.B.: Quasi-symmetric latent class models, with application to rater agreement. Biometrics **49**, 131–139 (1993)
Agresti, A., Min, Y.: On small-sample confidence intervals for parameters in discrete distributions. Biometrics **57**, 963–971 (2001)
Agresti, A., Min, Y.: Unconditional small-sample confidence intervals for the odds ratio. Biostatistics **3**, 379–386 (2002)
Agresti, A., Yang, M.: An empirical investigation of some effects of sparseness in contingency tables. Comput. Stat. Data Anal. **5**, 9–21 (1987)
Agresti, A., Chuang, C., Kezouh, A.: Order-restricted score parameters in association models for contingency tables. J. Am. Stat. Assoc. **82**, 619–623 (1987)
Agresti, A., Mehta, C.R., Patel, N.R.: Exact inference for contingency tables with ordered categories. J. Am. Stat. Assoc. **85**, 453–458 (1990)
Aitchison, J., Aitken, C.G.: Multivariate binary discrimination by the kernel method. Biometrika **63**, 413–420 (1976)
Ait-Sidi-Allal, L., Baccini, A., Mondot, A.M.: A new algorithm for estimating the parameters and their asymptotic covariance in correlation and association models. Comput. Stat. Data Anal. **45**, 389–421 (2004)
Akaike, H.: A new look at the statistical model identification. IEEE Trans. Automat. Control **19**, 716–723 (1974)
Albert, J.H.: Bayesian testing and estimation of association in a two-way contingency table. J. Am. Stat. Assoc. **92**, 685–693 (1997)
Albert, J.H., Chib, S.: Bayesian analysis of binary and polychotomous response data. J. Am. Stat. Assoc. **88**, 669–679 (1993)
Aldrich, J.: R. A. Fisher and the making of maximum likelihood 1912–1922. Stat. Sci. **12**, 162–176 (1997)
Altham, P.M.E.: Exact Bayesian analysis of a 2×2 contingency table, and Fisher's 'exact' significance test. J. Roy. Stat. Soc. Ser. B **31**, 261–269 (1969)
Altham, P.M.E.: The measurement of association of rows and columns for an $r \times c$ contingency table. J. Roy. Stat. Soc. Ser. B **32**, 63–73 (1970)
Altham, P.M.E.: The analysis of matched proportions. Biometrika **58**, 561–576 (1971)
Altham, P.M.E.: Quasi-independent triangular contingency tables. Biometrics **31**, 233–238 (1975)
Amiri, S., von Rosen, D.: On the efficiency of bootstrap method into the analysis contingency table. Comput. Meth. Programs Biomed. **104**, 182–187 (2011)
Andersen, E.B.: Conditional inference for multiple-choice questionnaires. Br. J. Math. Stat. Psychol. **26**, 31–44 (1973)
Andersen, A.H.: Multidimensional contingency tables. Scand. J. Stat. **1**, 115–127 (1974)
Andersen, E.B.: Discrete Statistical Models with Social Science Applications. North Holland, Amsterdam (1980)
Anderson, J.A.: Regression and ordered categorical variables (with discussion). J. Roy. Stat. Soc. B **46**, 1–30 (1984)
Andersen, E.B.: Diagnostics in categorical data analysis. J. Roy. Stat. Soc. Ser. B **54**, 781–791 (1992)
Anderson, C.J.: The Analysis of three-way contingency tables by three-mode association models. Psychometrika **61**, 465–483 (1996)
Andersen, E.B.: Introduction to the Statistical Analysis of Categorical Data. Springer, New York (2001)
Anderson, J.A., Philips, P.R.: Regression, discrimination, and measurement models for ordered categorical variables. Appl. Stat. **30**, 22–31 (1981)
Anderson, C.J., Vermunt, J.K.: Log-multiplicative association models as latent variable models for nominal and/or ordinal data. Socio. Meth. **30**, 81–121 (2000)
Anderson, C.J., Li, Z., Vermunt, J.K.: Estimation of models in a Rasch family for polytomous items and multiple latent variables. J. Stat. Software **20**, 1–36 (2007)

Andrich, D.: Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. Psychometrika **75**, 292–308 (2010)
Anscombe, F.J., Tukey, J.W.: The examination and analysis of residuals. Technometrics **5**, 141–160 (1963)
Aoki, S., Takemura, A.: Markov chain Monte Carlo exact tests for incomplete two-way contingency tables. J. Stat. Comput. Simul. **75**, 787–812 (2005)
Armitage, P.: Tests for linear trends in proportions and frequencies. Biometrics **11**, 375–386 (1955)
Asmussen, S., Edwards, D.: Collapsibility and reponse variables in contingency tables. Biometrika **70**, 567–578; **11**, 375–386 (1983)
Baccini, A., Khoudraji, A.: A least squares procedure for estimating the parameters in the RC-association model for contingency tables. Comput. Stat. **7**, 287–300 (1992)
Baccini, A., Fekri, M., Fine, J.: Generalized least squares estimation in contingency tables analysis: asymptotic properties and applications. Statistics **34**, 267–300 (2000)
Balagtas, C.C., Becker, M.P., Lang, J.B.: Marginal modelling of categorical data from crossover experiments. Appl. Stat. **44**, 63–77 (1995)
Banerjee, C., Capozzoli, M., McSweeney, L., Sinha, D.: Beyond kappa: a review of interrater agreement measures. Can. J. Stat. **27**, 3–23 (1999)
Baptista, J., Pike, M.C.: Exact two-sided confidence limits for the odds ratio in a 2×2 table. Appl. Stat. **26**, 214–220 (1977)
Barlow, W.: Measurement of interrater agreement with adjustment for covariates. Biometrics **52**, 695–702 (1996)
Barnard, G.A.: A new test for 2×2 tables. Nature **156**, 177 (1945)
Barnard, G.A.: Significance tests for 2×2 tables. Biometrika **34**, 123–138 (1947)
Bartholomew, D., Knott, M., Moustaki, I.: Latent Variable Models and Factor Analysis: A Unified Approach, 3rd edn. Wiley, Hoboken (2011)
Bartlett, M.S.: Contingency table interactions. J. Roy. Stat. Soc. Suppl. **2**, 248–252 (1935)
Bartlett, M.S.: Some examples of statistical methods of research in agriculture and applied biology. J. Roy. Stat. Soc. Suppl. **4**, 137–183 (1937)
Bartolucci, F., Forcina, A.: Extended RC association models allowing for order restrictions and marginal modeling. J. Am. Stat. Assoc. **97**, 1192–1199 (2002)
Bartolucci, F., Forcina, A., Dardanoni, V.: Positive quadrant dependence and marginal modelling in two-way tables with ordered margins. J. Am. Stat. Assoc. **96**, 1497–1505 (2001)
Bartolucci, F., Colombi, R., Forcina, A.: An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. Statistica Sinica **17**, 691–711 (2007)
Basu, A., Harris, I.R., Hjort, N.L., Jones, M.C.: Robust and efficient estimation by minimising a density power divergence. Biometrika **85**, 549–559 (1998)
Bavaud, F.: The quasi-symmetric side of gravity modelling. Environ. Plann. A **34**, 61–79 (2002)
Becker, M.: On the bivariate normal distribution and association models for ordinal categorical data. Stat. Probab. Lett. **8**, 435–440 (1989a)
Becker, M.: Models for the analysis of association in multivariate contingency tables. J. Am. Stat. Assoc. **84**, 1014–1019 (1989b)
Becker, M.P.: Maximum likelihood estimation of the RC(M) association model. Appl. Stat. **39**, 152–167 (1990a)
Becker, M.P.: Quasisymmetric models for the analysis of square contingency tables. J. Roy. Stat. Soc. **52**, 369–378 (1990b)
Becker, M.P.: Explanatory analysis of association models using loglinear models and singular value decomposition. Comput. Stat. Data Anal. **13**, 253–267 (1992)
Becker, M.P.: Analysis of cross-classifications of counts using models for marginal distributions: an application to trends in attitudes on legalizzed abortion. Socio. Meth. **24**, 229–265 (1994)
Becker, M., Agresti, A.: Log-linear modelling of pairwise interobserver agreement on a categorical scale. Stat. Med. **11**, 101–114 (1992)
Becker, M.P., Clogg, C.C.: A note on approximating correlations from odds ratios. Socio. Meth. Res. **16**, 407–424 (1988)

Becker, M.P., Clogg, C.C.: Analysis of sets of two-way contingency tables using association models. J. Am. Stat. Assoc. **84**, 142–151 (1989)
Becker, M.P., Yang, I.: Latent class marginal models for cross-classifications of counts. Socio. Meth. **28**, 293–325 (1998)
Becker, M.P., Minick, S., Yang, I.: Specifications of models for cross-classified counts: comparisons of the log-linear models and marginal models perspectives. Socio. Meth. Res. **26**, 511–529 (1998)
Beh, E.J.: Simple correspondence analysis of ordinal cross-classifications using orthogonal polynomials. Biomed. J. **39**, 589–613 (1997)
Beh, E.J.: A comparative study of scores for correspondence analysis with ordered categories. Biomed. J. **40**, 413–429 (1998)
Beh, E.J.: Simple correspondence analysis: a bibliographic review. Int. Stat. Rev. **72**, 257–284 (2004)
Beh, E.J., Farver, T.B.: An evaluation of non-iterative methods for estimating the linear-by-linear parameter of ordinal log-linear models. Aust. N. Z. J. Stat. **51**, 335–352 (2009)
Benzécri, J.P.: L' analyse des Données (L' Analyse des Correspondance), vol. 2. Dunod, Paris (1973)
Berger, V.W.: Admissibility of exact conditional tests of stochastic order. J. Stat. Plann. Infer. **66**, 39–50 (1998)
Berger, V.W., Ivanova, A.: The bias of linear rank tests when testing for stochastic order in ordered categorical data. J. Stat. Plann. Infer. **107**, 237–247 (2002)
Berger, V.W., Permutt, T., Ivanova, A.: Convex hull test for ordered categorical data. Biometrics **54**, 1541–1550 (1998)
Bergsma, W.P., Rudas, T.: Marginal models for categorical data. Ann. Stat. **30**, 140–159 (2002a)
Bergsma, W.P., Rudas, T.: Modeling conditional and marginal association in contingency tables. Ann. Fac. Sci. Toulouse Math. **11**, 443–454 (2002b)
Bergsma, W.P., Croon, M., Hagenaars, J.A.: Marginal Models: For Dependent, Clustered, and Longitudinal Categorical Data. Springer, New York (2009)
Berkson, J.: Some difficulties of interpretation encountered in the application of the chi-square test. J. Am. Stat. Assoc. **39**, 357–365 (1938)
Berkson, J.: Application of the logistic function to bio-assay. J. Am. Stat. Assoc. **39**, 357–365 (1944)
Berkson, J.: Why I prefer logits to probits. Biometrics **7**, 327–339 (1951)
Berkson, J.: A statistically precise and relatively simple method of estimating the bioassay with quantal response, based on the logistic function. J. Am. Stat. Assoc. **48**, 565–599 (1953)
Berkson, J.: Maximum likelihood and minimum logit X^2 estimation of the logistic function. J. Am. Stat. Assoc. **50**, 130–162 (1955)
Berry, G., Armitage, P.: Mid-*P* confidence intervals: a brief review. J. Stat. **44**, 417–423 (1995)
Best, D.J., Rayner, J.C.W.: Nonparametric analysis for doubly ordered two-way contingency tables. Biometrics **52**, 1153–1156 (1996)
Bhapkar, V.P.: A note on the equivalence of two criteria for hypotheses in categorical data. J. Am. Stat. Assoc. **61**, 228–235 (1966)
Bhapkar, V.P.: On tests of marginal symmetry and quasi-symmetry, in two- and three-dimensional contingency tables. Biometrics **35**, 417–426 (1979a)
Bhapkar, V.P.: On tests of symmetry when higher order interactions are absent. J. Ind. Stat. Assoc. **17**, 17–26 (1979b)
Bhapkar, V.P., Darroch, J.N.: Marginal symmetry and quasi symmetry of general order. J. Multivariate Anal. **34**, 173–184 (1990)
Birch, M.W: Maximum likelihood in three-way contingency tables. J. Roy. Stat. Soc. Ser. B **25**, 220–233 (1963)
Birch, M.W: The detection of partial association, I: the 2×2 case. J. Roy. Stat. Soc. Ser. B **26**, 313–324 (1964)
Birch, M.W.: The detection of partial association II: the general case. J. Roy. Stat. Soc. Ser. B **27**, 111–124 (1965)

Bishop, Y.M.M.: Full contingency tables, logits, and split contingency tables. Biometrics **25**, 383–400 (1969)
Bishop, Y.M.M., Fienberg, S.E.: Incomplete two-dimensional contingency tables. Biometrics **25**, 119–128 (1969)
Bishop, Y.M.M., Fienberg, S.E., Holland, P.W.: Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge (1975)
Blaker, H.: Confidence curves and improved exact confidence intervals for discrete distributions. Can. J. Stat. **28**, 783–798 (2000)
Bliss, C.I.: The method of probits. Science **79**, 38–39 (1934)
Bliss, C.I.: The calculation of the dosage - mortality curve. Ann. Appl. Biol. **22**, 134–167 (1935)
Blumen, I., Kogan, M., McCarthy, P.J.: The Industrial Mobility of Labor as a Probability Process. Cornell Studies of Industrial and Labor Relations, vol. 6. Cornell University Press, Ithaca (1955)
Blyth, C.R.: On Simpson's paradox and the sure-thing principle. J. Am. Stat. Assoc. **67**, 364–366 (1972)
Bock, H.H.: Loglinear models and entropy clustering methods for qualitative data. In: Gaul, W., Schader, M. (eds.) Classification as a Tool of Research, pp. 19–26. Elsevier Science Publishers B.V. (North-Holland), Amsterdam (1986)
Bockenholt, U.: A logistic representation of multivariate paired-comparison models. J. Math. Psychol. **32**, 44–63 (1988)
Bockenholt, U., Dillon, W.: Modelling within-subject dependencies in ordinal paired comparison data. Psychometrika **62**, 411–434 (1997)
Bonett, D.G., Price, R.M.: Inferential methods for the tetrachoric correlation coefficient. J. Educ. Behav. Stat. **30**, 213–225 (2005)
Bonett, D.G., Price, R.M.: Statistical inference for generalized Yule coefficients in 2×2 contingency tables. Socio. Meth. Res. **35**, 429–446 (2007)
Boulesteix, A.L., Strobl, C.: Maximally selected Chi-squared statistics and non-monotonic associations: an exact approach based on two cutpoints. Comput. Stat. Data Anal. **51**, 6295–6306 (2007)
Bowker, A.H.: A test for symmetry in contingency tables. J. Am. Stat. Assoc. **43**, 572–574 (1948)
Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs I. The method of paired comparisons. Biometrika **39**, 324–345 (1952)
Breen, R.: Statistical models of association for comparing cross-classifications. Socio. Meth. Res. **36**, 442–461 (2008)
Breslow, N., Clayton, D.G.: Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. **88**, 9–25 (1993)
Breslow, N., Day, N.E.: Statistical Methods in Cancer Research, vol. I. IARC, Lyon (1980)
Bross, I.: Misclassification in 2×2 tables. Biometrics **10**, 478–486 (1954)
Brown, M.B.: Identification of the sources of significance in two-way contingency tables. J. Roy. Stat. Soc. (C) **23**, 405–413 (1974)
Brown, L.D., Li, X.: Confidence intervals for two sample binomial distribution. J. Stat. Plann. Infer. **130**, 359–375 (2005)
Brown, L.D., Cai, T.T., DasGupta, A.: Interval estimation for a binomial proportion (with discussion). Stat. Sci. **16**, 101–133 (2001)
Brunswick, A.F.: Adolescent health, sex, and fertility. Am. J. Publ. Health **61**, 711–720 (1971)
Bühlmann, P., van de Geer, S.: Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, New York (2011)
Buonaccorsi, J.P.: Measurement Error: Models, Methods, and Applications. Chapman & Hall/CRC, Boca Raton (2010)
Burman, P.: On some testing problems for sparse contingency tables. J. Multivariate Anal. **88**, 1–18 (2004)
Burnham, K.P., Anderson, D.R.: Multimodel inference: understanding AIC and BIC in model selection. Socio. Meth. Res. **33**, 261–304 (2004)
Cai, T.T.: One-sided confidence intervals in discrete distributions. J. Stat. Plann. Infer. **131**, 63–88 (2005)

Carlier, A., Kroonenberg, P.M.: Decompositions and biplots in three-way correspondence analysis. Psychometrika **61**, 355–373 (1996)
Carroll, J.D., Green, P.E., Schaffer, C.M.: Interpoint distance comparisons in correspondence analysis. J. Market. Res. **23**, 271–280 (1986)
Carroll, J.D., Green, P.E., Schaffer, C.M.: Comparing interpoint distances in correspondence analysis. J. Market. Res. **24**, 445–450 (1987)
Carroll, J.D., Green, P., Schaffer, C.M.: Reply to Greenacre's commentary on the Carroll-Green-Schaffer scaling of two-way correspondence analysis solutions. J. Market. Res. **26**, 366–368 (1989)
Caussinus, H.: Contribution à l' analyse statistique des tableaux de corrélation. Ann. Facul. Sci. Univ. Toulouse **29**, 77–182 (1965)
Chen, T.T.: Log-linear models for categorical data with misclassification and double sampling. J. Am. Stat. Assoc. **74**, 481–488 (1979)
Cheng, P.E., Liou, M., Aston, J.A.D., Tsai, A.C.: Information identities and testing hypotheses: power analysis for contingency tables. Statistica Sinica **18**, 535–558 (2008)
Cheng, P.E., Liou, M., Aston, J.A.D.: Likelihood ratio tests with three-way tables. J. Am. Stat. Assoc. **105**, 740–749 (2010)
Choulakian, V.: Exploratory analysis of contingency tables by log-linear formulation and generalizations of correspondence analysis. Psychometrika **53**, 235–250 (1988)
Chuang, C.: Empirical Bayes methods for a two-way multiplicative-interaction model. Commun. Stat. Theory Meth. **11**, 2977–2989 (1982)
Clayton, D.G.: Some odds ratio statistics for the analysis of ordered categorical data. Biometrika **61**, 525–531 (1974)
Clayton, D.G.: An odds ratio comparison for ordered categorical data with censored observations. Biometrika **63**, 405–408 (1976)
Clayton, D.G.: A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika **65**, 141–151 (1978)
Clogg, C.C.: Using association models in sociological research: Some examples. Am. J. Socio. **88**, 114–134 (1982a)
Clogg, C.C.: Some models for the analysis of association in multiway crossclassifications having ordered categories. J. Am. Stat. Assoc. **77**, 803–815 (1982b)
Clogg, C.C., Shihadeh, E.S.: Statistical Models for Ordinal Variables. Sage, Thousand Oaks (1994)
Cochran, W.G.: Some methods for strengthening the common X^2 tests. Biometrics **10**, 417–451 (1954)
Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Meas. **20**, 37–46 (1960)
Cohen, J.: Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol. Bull. **70**, 213–220(1968)
Cohen, A., Sackrowitz, H.B.: Directional tests for one-sided alternatives in multivariate models. Ann. Stat. **26**, 2321–2338 (1998)
Cohen, A., Sackrowitz, H.B., Sackrowitz, M.: Testing whether treatment is 'better' than control with ordered categorical data: an evaluation of new methodology. Stat. Med. **19**, 2699–2712 (2000)
Collet, D.: Modelling Binary Data, 2nd edn. CRC Press, London (2003)
Colombi, R., Forcina, A.: Marginal regression models for the analysis of positive association of ordinal response variables. Biometrika **88**, 1007–1019 (2001)
Colombi, R., Giordano, S., Cazzaro, M., Lang, J.B.: Package `hmmm`. `R` package version 1.0-1 (2013)
Colombo, A.G., Ihm, P.: A quasi-independence model to estimate failure rates. Reliab. Eng. Syst. Saf. **21**, 309–318 (1988)
Congdon, P.: Bayesian Models for Categorical Data. Wiley, New York (2005)
Conover, W.J.: Some reasons for not using the Yates continuity correction on 2×2 contingency tables. J. Am. Stat. Assoc. **69**, 374–376 (1974)
Consonni, G., Pistone, G.: Algebraic Bayesian analysis of contingency tables with possibly zero-probability cells. Statistica Sinica **17**, 1355–1370 (2007)

Constantine, A.G., Gower, J.C.: Graphical representation of asymmetric matrices. J. Roy. Stat. Soc. Ser. C (Appl. Stat.) **27**, 297–304 (1978)
Cook, R.J., Kalbfleisch, J.D., Yi, G.Y.: A generalized mover-stayer model for panel data. Biostatistics **3**, 407–420 (2002)
Cornfield, J.: A statistical problem arising from retrospective studies. In: Neyman, J. (ed.) Proceedings of 3rd Berkeley Symposium on Mathematical Statistics and Probability, vol. 4, pp. 135–148. Statistical Laboratory of the University of California, Berkeley (1956)
Coull, B.A.: Continuity correction. In: Encyclopedia of Biostatistics. Wiley, New York (2005)
Coull, B.A., Agresti, A.: Generalized log-linear models with random effects, with application to smoothing contingency tables. Stat. Model. **0**, 1–21 (2003)
Cox, C., Chuang, C.: A comparison of chi-square partitioning and two logit analyses of ordinal pain data from a pharmaceutical study. Stat. Med. **3**, 273–285 (1984)
Cox, D.R.: The regression analysis of binary sequences. J. Roy. Stat. Soc. Ser. B **20**, 215–242 (1958a)
Cox, D.R.: Two further applications of a model for binary regression. Biometrika **45**, 562–565 (1958b)
Cox, D.R.: Analysis of Binary Data. Chapman & Hall, New York (1970a)
Cox, D.R.: The continuity correction. Biometrika **57**, 217–219 (1970b)
Cox, D.R.: Regression models and life-tables (with discussion). J. Roy. Stat. Soc. Ser. B **34**, 187–220 (1972)
Cox, D.R.: Some remarks on overdispersion. Biometrika **70**, 269–274 (1983)
Cox, D.R., Hinkley, D.V. Theoretical Statistics. Chapman & Hall, Boca Raton (1974)
Cox, D.R., Snell, E.J.: A general definition of residuals. J. Roy. Stat. Soc. B **30**, 248–265 (1968)
Cox, D.R., Snell, E.J.: Analysis of Binary Data, 2nd edn. Chapman & Hall, New York (1989)
Cramer, J.S.: The origins of logistic regression. Tinbergen Institute Discussion Paper, TI 2002–119/4, Tinbergen Institute (2002)
Cressie, N., Read, T.R.C.: Multinomial goodness-of-fit tests. J. Roy. Stat. Soc. B **46**, 440–464 (1984)
Croon, M.A., Bergsma, W., Hagenaars, J.A.: Log-linear models analyzing change in categorical variables by generalized log-linear models. Socio. Meth. Res. **29**, 195–229 (2000)
Dahinden, C., Parmigiani, G., Emerick, M.C., Bühlmann, P.: Penalized likelihood for sparse contingency tables with an application to full-length cDNA libraries. BMC Bioinformatics **8**, 476 (2007) (http://www.biomedcentral.com/1471-2105/8/476)
Dahinden, C., Kalisch, M., and Bühlmann, P.: Decomposition and model selection for large contingency tables. Biometrical J. **52**, 233–252 (2010)
Dale, J.R.: Local versus global association for bivariate ordered responses. Biometrika **71**, 507–514 (1984)
Dale, J.R.: Global cross-ratio models for bivariate, discrete, ordered responses. Biometrics **42**, 909–917 (1986)
Dardanoni, V., Forcina, A.: A unified approach to likelihood inference on stochastic orderings in a nonparametric context. J. Am. Stat. Assoc. **93**, 1112–1123 (1998)
Darroch, J.N., McCloud, P.I.: Category distinguishability and observer agreement. Aust. J. Stat. **28**, 371–388 (1986)
Darroch, J.N., Ratcliff, D.: Generalized iterative scaling for log-linear models. Ann. Math. Stat. **43**, 1470–1480 (1972)
Darroch, J.N., Speed, T.P.: Additive and multiplicative models and interactions. Ann. Stat. **11**, 724–738 (1983)
Darroch, J.N., Lauritzen, S.L., Speed, T.P.: Markov fields and log-linear interaction models for contingency tables. Ann. Stat. **8**, 522–539 (1980)
David, H.A.: The Method of Paired Comparisons. Oxford University Press, New York (1988)
Davidson, R.R.: On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. J. Math. Psychol. **65**, 317–328 (1970)
Davidson, R.R., Beaver, R.J.: On extending the Bradley-Terry model to incorporate within-pair order effects. Biometrics **33**, 693–702 (1977)

Davis, L.J.: Modification of the empirical logit to reduce bias in simple linear logistic regression. Biometrika **72**, 199–202 (1985)

Davison, A.C.: Treatment effect heterogeneity in paired data. Biometrika **79**, 463–474 (1992)

Day, N.E., Byar, D.P.: Testing hypotheses in case-control studies: equivalence of Mantel-Haenszel statistics and logit score tests. Biometrics **35**, 623–630 (1979)

de Falguerolles, A., Jmel, S., Whittaker, J.: Correspondence analysis and association models constrained by a conditional independence graph. Psychometrika **60**, 161–180 (1995)

de Leeuw, J.: Models and methods for the analysis of correlation coefficients. J. Econometrics **22**, 113–137 (1983a)

de Leeuw, J.: On the prehistory of correspondence analysis. Statistica Neerlandica **37**, 161–164 (1983b)

de Leeuw, J., Mair, P.: Simple and canonical correspondence analysis using the `R` package `anacor`. J. Stat. Software **31**, 1–18 (2009)

de Leeuw, J., van der Heijden, P.: Reduced rank models for contingency tables. Biometrika **78**, 229–232 (1991)

De Rooij, M., Heiser, W.J.: Graphical representations and odds ratios in a distance-association model for the analysis of cross-classied data. Psychometrika **70**, 99–123 (2005)

Dellaportas, P., Forster, J.: Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. Biometrika **86**, 615–633 (1999)

Deming, W.E., Stephan, F.F.: On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Ann. Math. Stat. **11**, 427–444 (1940)

Diaconis, P., Efron, B.: Testing for independence in a two-way table: new interpretations of the chi-square statistic (with discussion). Ann. Stat. **13**, 845–913 (1985)

Diaconis, P., Sturmfels, B.: Algebraic algorithms for sampling from conditional distributions. Ann. Stat. **26**, 363–397 (1998)

Digby, P.G.N.: Approximating the tetrachoric correlation coefficient. Biometrics **39**, 753–757 (1983)

Diggle, P.J., Heagerty, P., Liang, K.Y., Zeger, S.L.: Analysis of Longitudinal Data, 2nd edn. Clarendon Press, Oxford (2002)

Dobra, A.: Markov bases for decomposable graphical models. Bernoulli **9**, 1093–1108 (2003)

Dobra, A., Fienberg, S.E.: Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. Proc. Natl. Acad. Sci. **97**, 11885–11892 (2000)

Dobra, A., Fienberg, S.E.: Bounding entries in multi-way contingency tables given a set of marginal totals. In: Haitovsky, Y., Lerche, H.R., Ritov, Y. (eds.) Foundations of Statistical Inference. Proceedings of the Shoresh Conference 2000, pp. 3–16. Springer, Berlin (2003)

Dobra, A., Fienberg, S.E., Rinaldo, A., Slavkovic, A., Zhou, Y.: Algebraic statistics and contingency table problems: log-linear models, likelihood estimation, and disclosure limitation. In: Putinar, M., Sullivant, S. (eds.) Emerging Applications of Algebraic Geometry, pp. 63–88. Springer, New York (2009)

Dobson, A.J., Barnett, A.: An Introduction to Generalized Linear Models, 3rd edn. Chapman and Hall/CRC, Boca Raton (2008)

Douglas, R., Fienberg, S.E.: An overview of dependency models for cross-classified categorical data involving ordinal variables. In: Topics in Statistical Dependence. Insitute of Mathematical Statistics Lecture Notes, pp. 167–188. Institute of Mathematical Statistics, Hayward (1990)

Douglas, R., Fienberg, S.E., Lee, M.T., Sampson, A.R., Whitaker, L.R.: Positive dependence concepts for ordinal contingency tables. In: Topics in Statistical Dependence. Insitute of Mathematical Statistics Lecture Notes, pp. 189–202. Institute of Mathematical Statistics, Hayward (1990)

Ducharme, G.R., Lepage, Y.: Testing collapsibility in contingency tables. J. Roy. Stat. Soc. Ser. B **48**, 197–205 (1986)

Dykstra, R., Kochar, S., Robertson, T.: Inference for likelihood ratio ordering in the two-sample problem. J. Am. Stat. Assoc. **90**, 1034–1040 (1995)

Edwards, A.W.F.: The measure of association in a 2×2 Table. J. Roy. Stat. Soc. Ser. A **126**, 109–114 (1963)

Edwards, D., Kreiner, S.: The analysis of contingency tables by graphical models. Biometrika **70**, 553–565 (1983)
Edwardes, M.D.deB., Baltzan, M.: The generalization of the odds ratio, risk ratio and risk difference to $r \times k$ tables. Stat. Med. **19**, 1901–1914 (2000)
Efron, B., Hinkley, D.V.: Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information (with discussion). Biometrika **65**, 457–487 (1978)
Emerson, J.D., Moses, L.E.: A note on the Wilcoxon-Mann-Whitney test for $2 \times K$ ordered tables. Biometrics **41**, 303–309 (1985)
Enke, H.: On the analysis of incomplete two-dimensional contingency tables. Biometrical J. **19**, 561–573 (1977)
Escoufier, B.: Analyse factorielle en référence à un modèle: application à l' analyse de tableaux d' échange. Rev. Stat. Appl. **32**, 25–36 (1984)
Espeland, M.A., Odoroff, C.L.: Log-linear models for doubly sampled categorical data fitted by the EM algorithm. J. Am. Stat. Assoc. **80**, 663–670 (1985)
Evans, M., Gilula, Z., Guttman, I.: Computational issues in the Bayesian analysis of categorical data: log-linear and Goodman's RC model. Statistica Sinica **3**, 391–406 (1993)
Everitt, B.S., Landau, S., Leese, M., Stahl, D.: Cluster Analysis, 5th edn. Wiley, New York (2011)
Fahrmeir, F., Tutz, G.: Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd edn. Springer, New York (2001)
Farrington, C.P.: On assessing goodness of fit of generalized linear models to sparse data. J. Roy. Stat. Soc. Ser. B **58**, 349–360 (1996)
Fay, M.P.: Random marginal agreement coefficients: rethinking the adjustment for chance when measuring agreement. Biostatistics **6**, 171–180 (2005)
Fay, M.P.: Confidence intervals that match Fisher's exact or Blaker's exact tests. Biostatistics **11**, 373–374 (2010a)
Fay, M.P.: Two-sided exact tests and matching confidence intervals for discrete data. R J. **2**, 53–58 (2010b)
Fienberg, S.E.: Preliminary graphical analysis and quasi-independence for two-way contingency tables. J. Roy. Stat. Soc. Ser. C **18**, 153–168 (1969)
Fienberg, S.E.: Quasi-independence and maximum likelihood estimation in incomplete contingency tables. J. Am. Stat. Assoc. **65**, 1610–1616 (1970a)
Fienberg, S.E.: An iterative procedure for estimation in contingency tables. Ann. Math. Stat. **41**, 907–917 (1970b)
Fienberg, S.E.: Perspective canada as a social report. Soc. Indicat. Res. **2**, 153–174 (1975)
Fienberg, S.E.: When did Bayesian inference become 'Bayesian'? Bayesian Anal. **1**, 1–40 (2006)
Fienberg, S.E.: The Analysis of Cross-Classified Categorical Data. Springer, New York (reprint of the 1980 edition by MIT Press) (2007)
Fienberg, S.E., Larntz, K.: Loglinear representation for paired and multiple comparisons models. Biometrika **63**, 245–254 (1976)
Fienberg, S.E., Mason, W.M.: Identification and estimation of age-period-cohort effects in the analysis of discrete archival data. In: Schuessler, K.F. (ed.) Sociological Methodology, pp. 1–67. Jossey-Bass, San Francisco (1978)
Fienberg, S.E., Rinaldo, A.: Three centuries of categorical data analysis: log-linear models and maximum likelihood estimation. J. Stat. Plann. Infer. **137**, 3430–3445 (2007)
Fienberg, S.E., Rinaldo, A.: Maximum likelihood estimation in log-linear models. Ann. Stat. **40**, 996–1023 (2012)
Fischer, G.H., Molenaar, I.W.: Rasch Models: Foundations, Recent Developments and Applications. Springer, New York (1995)
Fisher, R.A.: On the interpretation of chi-square from contingency tables, and the calculation of P. J. Roy. Stat. Soc. **85**, 87–94 (1922)
Fisher, R.A.: Statistical Methods for Research Workers, 5th edn. Oliver and Boyd, Edinburgh (1934)
Fisher, R.A.: The precision of discriminant functions. Ann. Eugen. **10**, 422–429 (1940)

Fitzmaurice, G.M., Laird, N.M.: A likelihood-based method for analysing longitudinal binary responses. Biometrika **80**, 141–151 (1993)
Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**, 378–382 (1971)
Fleiss, J.L., Cohen, J.: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ. Psychol. Meas. **33**, 613–619 (1973)
Fleiss, J.L., Cohen, J., Everitt, B.S.: Large sample standard errors of kappa and weighted kappa. Psychol. Bull. **72**, 323–327 (1969)
Forcina, A., Dardanoni, V.: Regression models for multivariate ordered responses via the Plackett distribution. J. Multivariate Anal. **99**, 2472–2478 (2008)
Formann, A.K.: Linear logistic latent class analysis for polytomous data. J. Am. Stat. Assoc. **87**, 476–486 (1992)
Forster, J.J.: Bayesian inference for Poisson and multinomial log-linear models. Stat. Meth. **7**, 210–224 (2010)
Forster, J.J., McDonald, J.W., Smith, P.W.F.: Monte Carlo exact conditional tests for log-linear and logistic models. J. Roy. Stat. Soc. Ser. B **58**, 445–453 (1996)
Freeman, M.F., Tukey, J.W.: Transformations related to the angular and the square root. Ann. Math. Stat. **21**, 607–611 (1950)
Friendly, M.: Mosaic displays for multiway contingency tables. J. Am. Stat. Assoc. **89**, 190–200 (1994)
Friendly, M.: Conceptual and visual models for categorical data. Am. Stat. **49**, 153–160 (1995)
Friendly, M.: Extending mosaic displays: marginal, conditional, and partial views of categorical data. J. Comput. Stat. Graph. **8**, 373–395 (1999)
Friendly, M.: Visualizing Categorical Data. SAS Institute, Cary (2000)
Friendly, M.: A brief history of the mosaic display. J. Comput. Graph. Stat. **11**, 89–107 (2002)
Friendly, M.: Working with categorical data with R and the `vcd` and `vcdExtra` packages. http://cran.r-project.org/web/packages/vcdExtra/vignettes/vcd-tutorial.pdf (2013)
Frydman, H.: Maximum likelihood estimation in the mover-stayer model. J. Am. Stat. Assoc. **79**, 632–638 (1984)
Fuchs, C., Greenhouse, J.B.: The EM algorithm for maximum likelihood estimation in the mover-stayer model. Biometrics **44**, 605–613 (1988)
Fuchs, C., Kennet, R.: A test for detecting outlying cells in the multinormal distribution and two-way contingency tables. J. Am. Stat. Assoc. **75**, 395–398 (1980)
Gabriel, K.R.: Goodness of fit of biplots and correspondence analysis. Biometrika **89**, 423–436 (2002)
Galindo-Garre, F., Vermunt, J.K.: The order-restricted association model: two estimation algorithms and issues in testing. Psychometrika **69**, 641–654 (2004)
García-Pérez, M.A., Núñez-Antón, V.: Cellwise residual analysis in two-way contingency tables. Educ. Psychol. Meas. **63**, 825–839 (2003)
Gart, J.J., Nam, J.: Approximate interval estimation of the ratio of binomial parameters: a review and correction for skewness. Biometrics **44**, 323–338 (1988)
Gart, J.J., Thomas, D.G.: The performance of three approximate confidence limit methods for the odds ratio. Am. J. Epidemiol. **115**, 453–470 (1982)
Gart, J.J., Zweifel, J.R.: On the bias of various estimators of the logit and its variance with application to quantal bioassay. Biometrika **54**, 181–187 (1967)
Gart, J.J., Pettigrew, H.M., Thomas, D.G.: The effect of bias, variance estimation, skewness and kurtosis of the empirical logit on weighted least squares analyses. Biometrika **72**, 179–190 (1985)
Gautam, S.: Test for linear trend in 2xK ordered tables with open-ended categories. Biometrics **53**, 1163–1169 (1997)
Gautam, S., Sampson, A.R., Singh, H.: Iso-chi-squared testing of $2 \times k$ ordered tables. Can. J. Stat. **29**, 609–619 (2001)
Geenens, G., Simar, L.: Nonparametric tests for conditional independence in two-way contingency tables. J. Multivariate Anal. **101**, 765–788 (2010)

Genter, F.C., Farewell, V.T.: Goodness-of-link testing in ordinal regression models. Can. J. Stat. **13**, 37–44 (1985)
Ghosh, D.: Semiparametric global cross-ratio models for bivariate censored data. Scand. J. Stat. **33**, 609–619 (2006)
Gifi, D.: Nonlinear Multivariate Analysis. Wiley, Chichester (1990)
Gilula, Z.: Singular value decomposition of probability matrices: Probabilistic aspects of latent dichotomous variables. Biometrika **66**, 339–344 (1979)
Gilula, Z.: On some similarities between canonical correlation models and latent class models for two-way contingency tables. Biometrika **71**, 523–529 (1984)
Gilula, Z.: Grouping and association in contingency tables: an exploratory canonical correlation approach. J. Am. Stat. Assoc. **81**, 773–779 (1986)
Gilula, Z., Haberman, S.J.: Canonical analysis of contingency tables by maximum likelihood. J. Am. Stat. Assoc. **81**, 780–788 (1986)
Gilula, Z., Krieger, A.M.: Collapsed two-way contingency tables and the chi-square reduction principle. J. Roy. Stat. Soc. B **51**, 425–433 (1989)
Gilula, Z., Krieger, A.M., Ritov, Y.: Ordinal association in contingency tables: some interpretive aspects. J. Am. Stat. Assoc. **83**, 540–545 (1988)
Glassman, A.H., Helzer, J.E., Covey, L.S., Cottler, L.B., Stetner, F., Tipp, J.E., Johnson, J.: Smoking, smoking cessation, and major depression. J. Am. Med. Assoc. **264**, 1546–1549 (1990)
Glenn, W., David, H.: Ties in paired comparison experiments using a modified Thurstone-Mosteller model. Biometrics **16**, 86–109 (1960)
Glonek, G.F.V., McCullagh, P.: Multivariate logistic models. J. Roy. Stat. Soc. Ser. B **57**, 533–546 (1995)
Glonek, G.F.V., Darroch, J.N., Speed, T.P.: On the existence of maximum likelihood estimators for hierarchical loglinear models. Scand. J. Stat. **15**, 187–193 (1988)
Gokhale, D.V., Johnson, N.S.: A class of alternatives to independence in contingency tables. J. Am. Stat. Assoc. **73**, 800–804 (1978)
Gokhale, D.V., Kullback, S.: The Information in Contingency Tables. Marcel Dekker Inc, New York (1978a)
Gokhale, D.V., Kullback, S.: The minimum discrimination information approach in analyzing categorical data. Commun. Stat. Theory Meth. **7**, 987–1005 (1978b)
Gollob, H.F.: A statistical model which combines features of factor analytic and analysis of variance techniques. Psychometrika **33**, 73–116 (1968)
Good, I.J.: On the estimation of small frequencies in contingency tables. J. Roy. Stat. Soc. Ser. B **18**, 113–124 (1956)
Good, I.J.: The Estimation of Probabilities: An Essay on Modern Bayesian Methods. MIT Press, Cambridge (1965)
Goodman, L.A.: Statistical methods for the mover-stayer model. J. Am. Stat. Assoc. **56**, 841–868 (1961)
Goodman, L.A.: Statistical methods for the preliminary analysis of transaction flows. Econometrica **31**, 197–208 (1963a)
Goodman, L.A.: On Plackett's test for contingency table interactions. J. Roy. Stat. Soc. Ser. B **25**, 179–188 (1963b)
Goodman, L.A.: Interactions in multi-dimensional contingency tables. Ann. Math. Stat. **35**, 632–646 (1964)
Goodman, L.A.: On the statistical analysis of mobility tables. Am. J. Socio. **70**, 564–585 (1965)
Goodman, L.A.: The analysis of cross-classified data: independence, quasi independence, and interactions in contingency tables with or without missing entries. J. Am. Stat. Assoc. **63**, 1091–1131 (1968)
Goodman, L.A.: On partitioning chi-square and detecting partial association in three-way contingency tables. J. Roy. Stat. Soc. Ser. B **31**, 486–498 (1969)
Goodman, L.A.: The multivariate analysis of qualitative data: interactions among multiple classifications. J. Am. Stat. Assoc. **65**, 226–256 (1970)

Goodman, L.A.: A simple simultaneous test procedure for quasi-independence in contingency tables. Appl. Stat. **20**, 165–177 (1971a)
Goodman, L.A.: The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. Technometrics **13**, 33–61 (1971b)
Goodman, L.A.: The partitioning of chi-square, the analysis of marginal contingency tables, and the estimation of expected frequencies in multidimensional contingency tables. J. Am. Stat. Assoc. **66**, 339–344 (1971c)
Goodman, L.A.: Some multiplicative models for the analysis of cross-classified data. In: Sixth Berkely Symposium, vol. 1, pp. 649–696. University of California, California (1972)
Goodman, L.A.: Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika **61**, 215–231 (1974)
Goodman, L.A.: On quasi-independence in triangular contingency tables. Biometrics **35**, 651–655 (1979a)
Goodman, L.A.: Simple models for the analysis of association in cross–classifications having ordered categories. J. Am. Stat. Assoc. **74**, 537–552 (1979b)
Goodman, L.A.: Multiplicative models for square contingency tables with ordered categories. Biometrika **66**, 413–418 (1979c)
Goodman, L.A.: Multiplicative models for the analysis of occupational mobility tables and other kinds of cross-classification tables. Am. J. Socio. **84**, 804–819 (1979d)
Goodman, L.A.: Association models and canonical correlation in the analysis of cross–classifications having ordered categories. J. Am. Stat. Assoc. **76**, 320–334 (1981a)
Goodman, L.A.: Association models and the bivariate normal for contingency tables with ordered categories. Biometrika **68**, 347–355 (1981b)
Goodman, L.A.: Criteria for determining whether certain categories in a cross–classification table should be combined, with special reference to occupational categories in an occupational mobility table. Am. J. Socio. **87**(3), 612–650 (1981c)
Goodman, L.A.: Three elementary views of log-linear models for the analysis of cross-classifications having ordered categories. Socio. Meth. **12**, 193–239 (1981d)
Goodman, L.A.: The analysis of contingency tables having ordered categories using log-linear and log-nonlinear models. Metron **XL**, 37–52 (1982)
Goodman, L.A.: 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 (1983)
Goodman, L.A.: 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. Stat. **13**, 10–69 (1985)
Goodman, L.A.: Some useful extensions of the usual correspondence analysis and the usual log-linear models approach in the analysis of contingency tables with or without missing entries (with discussion). Int. Stat. Rev. **54**, 243–309 (1986)
Goodman, L.A.: New methods for analyzing the intrinsic character of qualitative variables using cross- classified data. Am. J. Socio. **93**, 529–583 (1987)
Goodman, L.A.: Total-score models and Rasch-type models for the analysis of a multidimensional contingency table, or a set of multidimensional contingency tables, with specified and/or unspecified order for response categories. Socio. Meth. **20**, 249–294 (1990)
Goodman, L.A.: Measures, models, and graphical displays in the analysis of cross-classified data (with discussion). J. Am. Stat. Assoc. **86**, 1085–1138 (1991)
Goodman, L.A.: On quasi-independence and quasi-dependence in contingency tables, with special inference to ordinal triangular contingency tables. J. Am. Stat. Assoc. **89**, 1059–1063 (1994)
Goodman, L.A.: 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. Am. Stat. Assoc. **91**, 408–427 (1996)

Goodman, L.A.: Contributions to the statistical analysis of contingency tables: Notes on quasi-symmetry, quasi-independence, log-linear models, log-bilinear models, and correspondence analysis models. Ann. Facul. Sci. Toulouse **XI**, 525–540 (2002a)
Goodman, L.A.: Latent class analysis: the empirical study of latent types, latent variables and latent structures. In: Hagenaars, J.A., McCutcheon, A.L. (eds.) Applied Latent Class Analysis, pp. 3–55. Cambridge University Press, Cambridge (2002b)
Goodman, L.A., Haberman, S.J.: The analysis of nonadditivity in two-way analysis of variance. J. Am. Stat. Assoc. **85**, 139–145 (1990)
Goodman, L.A., Hout, M.: Statistical methods and graphical displays for analyzing how the association between two qualitative variables differs among countries, among groups, or over time: a modified regression-type approach. Socio. Meth. **28**, 175–230 (1998)
Goodman, L.A., Kruskal, W.H.: Measures of Association for Cross Classifications. Springer, New York (1979)
Gottard, A., Marchetti, G.M., Agresti, A.: Quasi-symmetric graphical log-linear models. Scand. J. Stat. **38**, 447–465 (2011)
Gower, J.C.: The analysis of asymmetry and orthogonality. In: Barra, J., et al. (ed.) Recent Developments in Statistics, pp. 109–123. North Holland, Amsterdam (1977)
Gower, J.C.: Fisher's optimal scores and multiple correspondence analysis. Biometrics **46**, 947–961 (1990)
Graubard, B.I., Korn, E.L.: Choice of column scores for testing independence in ordered $2 \times K$ contingency tables. Biometrics **43**, 471–476 (1987)
Green, P.: Reversible Jump markov chain monte carlo computation and Bayesian model determination. Biometrika **82**, 711–732 (1995)
Greenacre, M.J.: Theory and Applications of Correspondence Analysis. Academic, New York (1984)
Greenacre, M.J.: Clustering the rows and columns of a contingency table. J. Classification **5**, 39–51 (1988a)
Greenacre, M.J.: Correspondence analysis of multivariate categorical data by weighted least squares. Biometrika **75**, 457–467 (1988b)
Greenacre, M.J.: The Carroll-Green-Schaffer scaling in correspondence analysis: a theoretical and empirical appraisal. J. Market. Res. **26**, 358–365 (1989)
Greenacre, M.J.: Correspondence analysis of square asymmetric matrices. J. Roy. Stat. Soc. C **49**, 297–310 (2000)
Greenacre, M.: Correspondence Analysis in Practice, 2nd edn. Chapman & Hall, Boca-Raton (2007)
Greenacre, M.J., Blasius, J.: Multiple Correspondence Analysis and Related Methods. Chapman & Hall, Boca-Raton (2006)
Greenacre, M.J., Hastie, T.: The geometric interpretation of correspondence analysis. J. Am. Stat. Assoc. **82**, 437–447 (1987)
Greenland, S.: Simpson's paradox from adding constants in contingency tables as an example of Bayesian noncollapsibility. Am. Stat. **64**, 340–344 (2010)
Grizzle, J.E.: Continuity correction in the X^2 test for 2×2 tables. Am. Stat. **21**, 28–32 (1967)
Grizzle, J.E., Williams, O.D.: Loglinear models and tests of independence for contingency tables. Biometrics **28**, 137–156 (1972)
Grizzle, J.E., Starmer, C.F., Koch, G.G.: Analysis of categorical data by linear models. Biometrics **25**, 489–504 (1969)
Gross, S.T.: On asymptotic power and efficiency of tests of independence in contingency tables with ordered classifications. J. Am. Stat. Assoc. **76**, 935–941 (1981)
Grove, D.M.: A test of independence against a class of ordered alternatives in a $2 \times C$ contingency table. J. Am. Stat. Assoc. **75**, 454–459 (1980)
Grove, D.M.: Positive association in a two-way contingency table: likelihood ratio tests. Commun. Stat. Theory Meth. **13**, 931–945 (1984)
Haber, M.: A comparison of some continuity corrections for the chi-squared test on 2×2 tables. J. Am. Stat. Assoc. **75**, 510–515 (1980)

Haber, M.: The continuity correction and statistical testing. Int. Stat. Rev. **50**, 135–144 (1982)
Haber, M.: Maximum likelihood methods for linear and log-linear models in categorical data. Comput. Stat. Data Anal. **3**, 1–10 (1985)
Haber, M., Brown, M.: Maximum likelihood methods for log-linear models when expected frequencies are subject to linear constraints. J. Am. Stat. Assoc. **81**, 477–482 (1986)
Haberman, S.J.: Log-linear models for frequency data: sufficient statistics and likelihood equations. Ann. Stat. **1**, 617–632 (1973a)
Haberman, S.J.: The analysis of residuals in cross-classified tables. Biometrics **29**, 205–220 (1973b)
Haberman, S.J.: The Analysis of Frequency Data. University of Chicago Press, Chicago (1974a)
Haberman, S.J.: Log-linear models for frequency tables with ordered classifications. Biometrics **30**, 589–600 (1974b)
Haberman, S.J.: Log-linear models and frequency tables with small expected cell counts. Ann. Stat. **5**, 1148–1169 (1977)
Haberman, S.J.: Analysis of Qualitative Data, vols. 1 and 2. Academic, New York (1979)
Haberman, S.J.: Tests for independence in two-way contingency tables based on canonical correlation and on linear-by-linear interaction. Ann. Stat. **9**, 1178–1186 (1981)
Haberman, S.J.: A warning on the use of Chi-squared statistics with frequency tables with small expected cell counts. J. Am. Stat. Assoc. **83**, 555–560 (1988)
Haberman, S.J.: Computation of maximum likelihood estimates in association models. J. Am. Stat. Assoc. **90**, 1438–1446 (1995)
Hagenaars, J.A.: Categorical causal modeling: latent class analysis and directed log-linear models with latent variables. Socio. Meth. Res. **26**, 436–486 (1998)
Haldane, J.B.S.: The estimation and significance of the logarithm of a ratio of frequencies. Ann. Hum. Genet. **20**, 309–311 (1956)
Hamdan, M.A.: Comparison of two measures of association in two-way contingency tables. Can. J. Stat. **5**, 235–240 (1977)
Hara, H., Sei, T., Takemura, A.: Hierarchical subspace models for contingency tables. J. Multivariate Anal. **103**, 19–34 (2012)
Harris, J.A., Treloar A.E.: On a limitation in the applicability of the contingency coefficient. Am. Stat. **22**, 460–472 (1927)
Harris, J.A., Tu, C.: A second category of limitations in the applicability of the contingency coefficient. Am. Stat. **242**, 367–375 (1929)
Hartigan, J.A., Kleiner, B.: Mosaics for contingency tables. In: Eddy, W.F. (ed.) Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface, pp. 268–273. Springer, New York (1981)
Hartigan, J.A., Kleiner, B.: A mosaic of television ratings. Am. Stat. **38**, 32–35 (1984)
Hastie, T.J., Pregibon, D.: Generalized linear models, chapter 6. In: Chambers, J.M., Hastie, T.J. (eds.) Statistical Models in S. Wadsworth & Brooks/Cole, Pacific Grove (1992)
Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York (2009)
Heagerty, P.J., Zeger, S.L.: Marginal regression models for clustered ordinal measurements. J. Am. Stat. Assoc. **91**, 1024–1036 (1996)
Hauck, W.W.: A comparative study of conditional maximum likelihood estimation of a common odds ratio. Biometrics **40**, 1117–1123 (1984)
Hedeker, D., Gibbons, R.D.: A Random-effects ordinal regression model for multilevel analysis. Biometrics **50**, 7–11 (1994)
Heinen, T.: Latent Class and Discrete Latent Trait Models. Sage Publications, Thousand Oaks (1996)
Heiser, W.J.: Correspondence analysis with least absolute residuals. Comput. Stat. Data Anal. **5**, 337–356 (1987)

Heiser, W.J., Meulmann, J.J.: Homogeneity analysis: exploring the distribution of variables and their nonlinear relationships. In: Greenacre, M., Blasius, J. (eds.) Correspondence Analysis in the Social Sciences: Recent Developments and Applications, pp. 179–209. Academic, London (1994)

Hill, M.O.: Correspondence analysis: a neglected multivariate method. Appl. Stat. **23**, 340–354 (1974)

Hilton, J.F., Mehta, C.R., Patel, N.R.: An algorithm for conducting exact Smirnov tests. Comput. Stat. Data Anal. **17**, 351–361 (1994)

Hirji, K.F.: Exact Analysis of Discrete Data. Chapman & Hall/CRC, Boca Raton (2006)

Hirji, K.F., Tan, S.J., Elashoff, R.M.: A quasi-exact test for comparing two binomial proportions. Stat. Med. **10**, 1137–1153 (1991)

Hirotsu, C.: Defining the pattern of association in two-way contingency tables. Biometrika **70**, 579–589 (1983)

Hirschfeld, H.O.: A connection between correlation and contingency. Cambridge Phil. Soc. Proc. (Math. Proc.) **31**, 520–524 (1935)

Højsgaard, S., Edwards, D., Lauritzen, S.: Graphical Models with R. Springer, New York (2012)

Holst, L.: Asymptotic normality and efficiency for certain goodness of fit tests. Biometrika **59**, 137–145 (1972)

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression, 3rd edn. Wiley, New York (2013)

Howard, J.V.: The 2×2 table: a discussion from a Bayesian viewpoint. Stat. Sci. **13**, 351–367 (1998)

Hutchinson, T.P.: Focus on psychometrics. Kappa muddles together two sources of disagreement: tetrachoric correlation is preferable. Res. Nurs. Health **16**, 313–315 (1993)

Hwang, J.T.G., Yang, M.-C.: An optimality theory for mid p-values in 2×2 contingency tables. Statistica Sinica **11**, 807–826 (2001)

Hwang, H., Tomiuk, M.A., Takane, Y.: Correspondence analysis, multiple correspondence analysis, and recent developments, chapter 11. In: Millsap, R.E., Maydeu-Olivares, A. (eds.) The Sage Handbook of Quantitative Methods in Psychology. Sage Publications, Thousand Oaks (2009)

Iliopoulos, G., Kateri, M., Ntzoufras, I.: Bayesian estimation of unrestricted and order-restricted association models for a two-way contingency table. Comput. Stat. Data Anal. **51**, 4643–4655 (2007)

Iliopoulos, G., Kateri, M., Ntzoufras, I.: Bayesian model comparison for the order restricted RC association model. Psychometrika **74**, 561–587 (2009)

Imrey, P.B., Johnson, W.D., Koch, G.G.: An incomplete contingency table approach to paired-comparison experiments. J. Am. Stat. Assoc. **71**, 614–623 (1976)

Ireland, C.T., Kullback, S.: Contingency tables with given marginals. Biometrika **55**, 179–188 (1968a)

Ireland, C.T., Kullback, S.: Minimum discrimination information estimation. Biometrics **24**, 707–713 (1968b)

Ireland, C.T., Ku, H.H., Kullback, S.: Symmetry and marginal homogeneity of an $r \times r$ contingency table. J. Am. Stat. Assoc. **64**, 1323–1341 (1969)

Irwin, J.O.: A note on the subdivision of chi-square into components. Biometrika **36**, 130–134 (1949)

James, I.R.: Analysis of nonagreements among multiple raters. Biometrics **39**, 651–657 (1983)

Janson, H., Olsson, U.: A measure of agreement for interval or nominal multivariate observations by different sets of judges. Educ. Psychol. Meas. **64**, 62–70 (2004)

Jeong, H.C., Jhun, M., Kim, D.: Bootstrap tests for independence in two-way ordinal contingency tables. Comput. Stat. Data Anal. **48**, 623–631 (2005)

Jewell, N.P.: Small-sample bias of point estimators of the odds ratio from matched sets. Biometrics **40**, 421–435 (1984)

Jewell, N.P.: On the bias of commonly used measures of association for 2×2 tables. Biometrics **42**, 351–358 (1986)

Jhun, M., Jeong, H.C.: Applications of bootstrap methods for categorical data analysis. Comput. Stat. Data Anal. **35**, 83–91 (2000)
Joe, H.: Relative entropy measures of multivariate dependence. J. Am. Stat. Assoc. **84**, 157–164 (1989)
Johnson, N.S.: C_a method for testing for significance in the $r \times c$ contingency table. J. Am. Stat. Assoc. **70**, 942–947 (1975)
Johnson, T.R.: Discrete choice models for ordinal response variables: a generalization of the stereotype model. Psychometrika **72**, 489–504 (2007)
Johnson, V.E., Albert J.H.: Ordinal Data Modeling. Springer, New York (2000)
Jones, M.C.: Constant local dependence. J. Multivariate Anal. **64**, 148–155 (1998)
Jones, M.P., O'Gorman, T.W., Lemke, J.H., Woolson, R.F.: A Monte Carlo investigation of homogeneity tests of the odds ratio under various sample size configurations. Biometrics **45**, 171–181 (1989)
Kalbfleisch, J.D., Lawless, J.F.: The analysis of panel data under a Markov assumption. J. Am. Stat. Assoc. **80**, 863–871 (1985)
Kass, R.E., Raftery, A.E.: Bayes factors. J. Am. Stat. Assoc. **90**, 773–795 (1995)
Kastenbaum, M.A.: A note on the additive partitioning of chi-square in contingency tables. Biometrics **16**, 416–422 (1960)
Kateri, M.: Categorical data. In: Encyclopedia of Statistical Sciences. Wiley, New York (2008)
Kateri, M.: On the comparison of two ordinal responses. Commun. Stat. Theory Meth. **40**, 3748–3763 (2011)
Kateri, M., Agresti, A.: A class of ordinal quasi symmetry models for square contingency tables. Stat. Probab. Lett. **77**, 598–603 (2007)
Kateri, M., Agresti, A.: A generalized regression model for a binary response. Stat. Probab. Lett. **80**, 89–95 (2010)
Kateri, M., Agresti, A.: Bayesian inference about odds ratio structure in ordinal contingency tables. Environmetrics **24**, 281–288 (2013)
Kateri, M., Balakrishnan, N.: Statistical evidence in contingency tables analysis. J. Stat. Plann. Infer. **138**, 873–887 (2008)
Kateri, M. and Dellaportas, A.: Conditional symmetry models for three-way contingency tables. J. Stat. Plann. Infer. **142**, 2430–2439 (2012)
Kateri, M., Iliopoulos, G.: On collapsing categories in two-way contingency tables. Statistics **37**, 443–455 (2003)
Kateri, M., Papaioannou, T.: f-divergence association models. Int. J. Math. Stat. Sci. **3**, 179–203 (1995)
Kateri, M., Papaioannou, T.: Asymmetry models for contingency tables. J. Am. Stat. Assoc. **92**, 1124–1131 (1997)
Kateri, M., Ahmad, R., Papaioannou, T.: New features in the class of association models. Appl. Stoch. Model. Data Anal. **14**, 125–136 (1998)
Kateri, M., Nicolaou, A., Ntzoufras, I.: Bayesian Inference for the RC(M) association model. J. Comput. Graph. Stat. **14**, 116–138 (2005)
Kelderman, H.: Loglinear Rasch model tests. Psychometrika **49**, 223–245 (1984)
Kendall, M.G.: A new measure of rank correlation. Biometrika **30**, 81–93 (1938)
Kendall, M.G.: Rank Correlation Methods. Charles Griffin, London (1948)
Khamis, H.J.: The Association Graph and the Multigraph for Loglinear Models. Sage, Thousand Oaks (2011)
Kim, H.: Measures of influence in correspondence analysis. J. Stat. Comput. Simul. **40** 201–217 (1992)
Kim, S.H., Choi, H., Lee, S.: Estimate-based goodness-of-fit test for large sparse multinomial distributions. Comput. Stat. Data Anal. **53** , 1122–1131 (2009)
Kimeldorf, G., Sampson, A.R., Whitaker, L.R.: Min and max scoring for two-sample ordinal data. J. Am. Stat. Assoc. **87**, 241–247 (1992)
King, R., Brooks, S.P.: Prior induction in log-linear models for general contingency table analysis. Ann. Stat. **29**, 715–747 (2001)

Klimova, A., Rudas, T., Dobra, A.: Relational models for contingency tables. J. Multivariate Anal. **104**, 159–173 (2012)
Knuiman, M.W., Speed, T.P.: Incorporating prior information into the analysis of contingency tables. Biometrics **44**, 1061–1071 (1988)
Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H., Lehnen, R.G.: A general methodology for the analysis of experiments with repeated measurements of categorical data. Biometrics **33**, 133–158 (1977)
Koch, G.G., Amara, I.A., Davis, G.W., Gillings, D.B.: A review of some statistical methods for covariance analysis of categorical data. Biometrics **38**, 563–595 (1982)
Koehler, K.: Goodness-of-fit tests for log-linear models in sparse contingency tables. J. Am. Stat. Assoc. **81**, 483–493 (1986)
Koshimizu, T., Tsujitani, M.: Association models with location and dispersion scores for the analysis of singly-ordered contingency tables. Behaviormetrika **25**, 151–164 (1998)
Kotze, T.J.vW., Hawkins, D.M.: The identification of outliers in two-way contingency tables using 2×2 subtables. J. Appl. Stat. **33**, 215–223 (1984)
Kraemer, H.C.: Extension of the kappa coefficient. Biometrics **36**, 207–216 (1980)
Kraemer, H.C.: Reconsidering the odds ratio as a measure of 2×2 association in a population. Stat. Med. **23**, 257–270 (2004)
Krampe, A., Kuhnt, S.: Bowker's test for symmetry and modifications within the algebraic framework. Comput. Stat. Data Anal. **51**, 4124–4142 (2007)
Krampe, A., Kateri, M., Kuhnt, S.: Asymmetry models for square contingency tables: exact tests via algebraic statistics. Stat. Comput. **21**, 55–67 (2011)
Kroonenberg, P.M.: Singular value decompositions of interactions in three-way contingency tables. In: Coppi, R., Bolasco, S. (eds.) Multiway Data Analysis. North-Holland, Amsterdam (1989)
Kroonenberg, P.M., Lombardo, R.: Nonsymmetric correspondence analysis: a tool for analysing contingency tables with a dependence structure. Multivariate Behav. Res. **34**, 367–396 (1999)
Ku, H.H., Kullback, S.: Loglinear models in contingency table analysis. Am. Stat. **28**, 115–122 (1974)
Kuha, J., Firth, D.: On the index of dissimilarity for lack of fit in loglinear and log-multiplicative models. Comput. Stat. Data Anal. **55**, 375–388 (2011)
Kuhnt, S.: Outlier identification procedures for contingency tables using maximum likelihood and L_1 estimates. Scand. J. Stat. **31**, 431–442 (2004)
Kulinskaya, E., Morgenthaler, S., Staudte, R.G.: Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence. Wiley, New York (2008)
Kullback, S.: Information Theory and Statistics. Wiley, New York (1959)
Kuriki, S.: Asymptotic distribution of inequality-restricted canonical correlation with application to tests for independence in ordered contingency tables. J. Multivariate Anal. **94**, 420–449 (2005)
Kuss, O.: On the estimation of the stereotype regression model. Comput. Stat. Data Anal. **50**, 1877–1890 (2006)
Laird, N.M.: Empirical Bayes methods for two-way contingency tables. Biometrika **65**, 581–590 (1978)
Lancaster, H.O.: The derivation and partition of chi-square in certain discrete distributions. Biometrika **36**, 117–129 (1949)
Lancaster, H.O.: The exact partition of chi-square and its application to the problem of pooling small expectations. Biometrika **37**, 267–270 (1950)
Lancaster, H.O.: Significance tests in discrete distributions. J. Am. Stat. Assoc. **56**, 223–234 (1961)
Lancaster, H.O.: Canonical correlations and partitions of X^2. Quart. J. Math **14**, 220–224 (1963)
Lancaster, H.O., Hamdan, M.A.: Estimation of the correlation coefficient in contingency tables with possible nonmetrical characters. Psychometrika **29**, 383–391 (1964)
Landis, J.R., Koch, G.G.: An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics **33**, 363–374 (1977)

Landis, J.R., Koch, G.G.: The analysis of categorical data in longitudinal studies of behavioral development. In: Nesselroade, J.R., Baltes, P.B. (eds.) Longitudinal Research in the Study of Behavior and Development, pp. 233–262. Academic, New York (1979)

Landis, J.R., Heyman, E.R., Koch, G.G.: Average partial association in three-way contingency tables: a review and discussion of alternative tests. Int. Stat. Rev. **46**, 237–254 (1978)

Lang, J.B.: Maximum likelihood methods for a generalized class of loglinear models. Ann. Stat. **24**, 726–752 (1996a)

Lang, J.B.: On the comparison of multinomial and Poisson log-linear models. J. Roy. Stat. Soc. Ser. B **58**, 253–266 (1996b)

Lang, J.B.: Multinomial-Poisson homogeneous models for contingency tables. Ann. Stat. **32**, 340–383 (2004)

Lang, J.B.: Homogeneous linear predictor models for contingency tables. J. Am. Stat. Assoc. **100**, 121–134 (2005)

Lang, J.B., Agresti, A.: Simultaneously modelling the joint and marginal distributions of multivariate categorical responses. J. Am. Stat. Assoc. **89**, 625–632 (1994)

Lang, J.B., Eliason, S.R.: The application of association-marginal models to the study of social mobility. Socio. Meth. Res. **26**, 183–213 (1997)

Lang, J.B., McDonald, J.W., Smith, P.W.F.: Association-marginal modeling of multivariate categorical responses: a maximum likelihood approach. J. Am. Stat. Assoc. **94**, 1161–1171 (1999)

Lapp, K., Molenberghs, G., Lesaffre, E.: Models for the association between ordinal variables. Comput. Stat. Data Anal. **28**, 387–411 (1998)

Lauritzen, S.L.: The EM algorithm for graphical association models with missing data. Comput. Stat. Data Anal. **19**, 191–203 (1995)

Lauritzen, S.L.: Graphical Models. Clarendon Press, Oxford (1996)

Lauritzen, S.L., Richardson, T.S.: Chain graph models and their causal interpretations. J. Roy. Stat. Soc. Ser. B **64**, 321–361 (2002)

Lauro, N.C., Balbi, S.: The analysis of structured qualitative data. Appl. Stoch. Model. Data Anal. **15**, 1–27 (1999)

Lauro, C., D' Ambra, L.: Non symmetrical correspondence analysis. In: Diday, E., et al. (eds.) Data Analysis and Informatics III, pp. 433–446. North Holland, Amsterdam (1984)

Lazarsfeld, P.F.: The logical and mathematical foundation of latent structure analysis. In: Suchman, E.A., Lazarsfeld, P.F., Starr, S.A., Clausen, J.A. (eds.) Studies in Social Psychology in World War II. Vol 4: Measurement and Prediction, pp. 362–412. Princeton University Press, Princeton (1950)

Lebart, L., Morineau, A., Warwick, K.: Multivariate Descriptive Statistical Analysis. Wiley, New York (1984)

Lee, A.H., Fung, W.K.: Confirmation of multiple outliers in generalized linear and nonlinear regressions. Comput. Stat. Data Anal. **25**, 55–65 (1997)

Lee, A.H., Yick, J.S.: A perturbation approach to outlier detection in two-way contingency tables. Aust. N. Z. J. Stat. **41**, 305–314 (1999)

Lehmann, E.: Some concepts of dependence. Ann. Math. Stat. **37**, 1137–1153 (1966)

Leonard, T.: Bayesian estimation methods for two-way contingency tables. J. Roy. Stat. Soc. B **37**, 23–37 (1975)

Liang, K.Y., Self, S.G.: Tests for homogeneity of odds ratio when the data are sparse. Biometrika **72**, 353–358 (1985)

Liang, K.Y., Zeger, S.L.: Longitudinal data analysis using generalized linear models. Biometrika **73**, 13–22 (1986)

Lindley, D.V.: The Bayesian analysis of contingency tables. Ann. Math. Stat. **35**, 1622–1643 (1964)

Lipsitz, S.R., Kim, K., Zhao, L.: Analysis of repeated categorical data using generalized estimating equations. Stat. Med. **13**, 1149–1163 (1994)

Liu, I., Agresti, A.: The analysis of ordered categorical data: an overview and a survey of recent developments (with discussion). Test **14**, 1–73 (2005)

Liu, B., Guo, J.: Collapsibility of conditional graphical models. Scand. J. Stat. **40**, 191–203 (2012)
Liu, Q., Pierce, D.A.: Heterogeneity in Mantel-Haenszel-type models. Biometrika **80**, 543–556 (1993)
Lombardo, R., Beh, E.J., D'Ambra, L.: Non-symmetric correspondence analysis with ordinal variables using orthogonal polynomials. Comput. Stat. Data Anal. **52**, 566–577 (2007)
Lovison, G.: Generalized symmetry models for hypercubic concordance tables. Int. Stat. Rev. **68**, 323–338 (2000)
Luce, R.D.: Individual Choice Behavior: A Theoretical Analysis. Wiley, New York (1959)
Lupparelli, M., Marchetti, G.M., Bergsma, W.P.: Parameterizations and fitting of bi-directed graph models to categorical data. Scand. J. Stat. **36**, 559–576 (2009)
MacDonald, P.L., Gardner, R.C.: Type I error rate comparisons of post hoc procedures for $I \times J$ chi-square tables. Educ. Psychol. Meas. **60**, 735–754 (2000)
Madansky, A.: Tests of homogeneity for correlated samples. J. Am. Stat. Assoc. **58**, 97–119 (1963)
Madigan, D., Raftery, A.E.: Model selection and accounting for model uncertainty in graphical models using Occam's window. J. Am. Stat. Assoc. **89**, 1535–1546 (1994)
Madigan, D., York, J.: Bayesian graphical models for discrete data. Int. Stat. Rev. **63**, 215–232 (1995)
Mair, P., Hatzinger, R.: CML based estimation of extended Rasch models with the eRm package in R. Psychol. Sci. **49**, 26–43 (2007)
Mantel, N.: Chi-square tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. J. Am. Stat. Assoc. **58**, 690–700 (1963)
Mantel, N.: Incomplete contingency tables. Biometrics **26**, 291–304 (1970)
Mantel, N., Byar, D.P.: Marginal homogeneity, symmetry and independence. Commun. Stat. Ser. A **7**, 953–976 (1978)
Mantel, N., Haenszel, W.: Statistical aspects of the analysis of data from retrospective studies of disease. J. Natl. Cancer Inst. **22**, 719–748 (1959)
Marchetti, G.M., Lupparelli, M.: Chain graph models of multivariate regression type for categorical data. Bernoulli **17**, 827–844 (2011)
Mardia, K.V.: Families of Bivariate Distributions. Charles Griffin and Co, London (1970)
Marselos, M, Boutsouris, K., Liapi, H., Malamas, M., Kateri, M., Papaioannou, T.: Epidemiological aspects on the use of cannabis among university students in Greece. Eur. Addiction Res. **3**, 184–191 (1997)
Martìn, N., Pardo, N.: New families of estimators and test statistics in log-linear models. J. Multivariate Anal. **99**, 1590–1609 (2008)
Martín Andrés, A., Tapia García, J.M., Silva-Mato, A., Sánchez Quevedo, M.J.: On the validity condition of the chi-squared test in 2×2 tables. Test **14**, 99–128 (2005)
Massam, H., Liu, J., Dobra, A.: A conjugate prior for discrete hierarchical log-linear models. Ann. Stat. **37**, 3431–3467 (2009)
Maydeu-Olivares, A., Joe, H. Limited and full information estimation and goodness-of-fit testing in 2^n contingency tables: a unified framework. J. Am. Stat. Assoc. **100**, 1009–1020 (2005)
Maydeu-Olivares, A., Joe, H.: Limited information goodness-of-fit testing in multidimensional contingency tables. Psychometrika **71**, 713–732 (2006)
McCullagh, P.: A logistic model for paired comparisons with ordered categorical data. Biometrika **64**, 449–453 (1977)
McCullagh, P.: A class of parametric models for the analysis of square contingency tables with ordered categories. Biometrika **65**, 413–418 (1978)
McCullagh, P.: Regression models for ordinal data (with discussion). J. Roy. Stat. Soc. B **42**, 109–142 (1980)
McCullagh, P.: Some applications of quasi-symmetry. Biometrika **69**, 303–308 (1982)
McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. Chapman and Hall, London (1989)
McDonald, J.W., Smith, P.W.F.: Exact conditional tests of quasi-independence for triangular contingency tables: estimating attained significance levels. Appl. Stat. **44**, 143–151 (1995)

McDonald, L.L., Davis, B.M., Milliken, G.A.: A non-randomized unconditional test for comparing two proportions in a 2×2 contingency table. Technometrics **19**, 145–150 (1977)
McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**, 153–157 (1947)
Mehta, C.R., Patel, N.R.: A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. J. Am. Stat. Assoc. **78**, 427–434 (1983)
Mehta, C.R., Patel, N.R.: Exact logistic regression: theory and examples. Stat. Med. **14**, 2143–2160 (1995)
Mehta, C.R., Senchaudhuri, P.: Conditional versus unconditional exact tests for comparing two binomials. Technical Report, vol. 5. Cytel Software Corporation, Cambridge (2003)
Mehta, C.R., Walsh, S.J.: Comparison of exact, mid-p, and Mantel-Haenszel confidence intervals for the common odds ratio across several 2×2 contingency tables. Am. Stat. **46**, 146–150 (1992)
Meyer, D., Zeileis, A., Hornik, K.: The strucplot framework: Visualizing multi-way contingency tables with `vcd`. J. Stat. Software **17**(3), 1–48 (2006)
Michailidis, G., De Leeuw, J.: The Gifi system for descriptive multivariate analysis. Stat. Sci. **13**, 307–336 (1998)
Molenberghs, G., Lesaffre, E.: Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. J. Am. Stat. Assoc. **89**, 633–644 (1994)
Molenberghs, G., Lesaffre, E.: Marginal modelling of multivariate categorical data. Stat. Med. **18**, 2237–2255 (1999)
Molenberghs, G., Verbeke, G.: Models for Discrete Longitudinal Data. Springer, New York (2005)
Morgan, B.J.T., Titterington, D.M.: A comparison of iterative methods for obtaining maximum likelihood estimates in contingency tables with a missing diagonal. Biometrika **64**, 265–269 (1977)
Morris, C.: Central limit theorems for multinomial sums. Ann. Stat. **3**, 165–188 (1975)
Mote, V.L., Anderson, R.L.: The effect of misclassification on the properties of χ^2-tests. Biometrika **52**, 95–109 (1965)
Murrell, P.: R Graphics. Chapman and Hall/CRC, London (2006)
Nair, V.N.: Testing in industrial experiments with ordered categorical data (with discussion). Technometrics **28**, 283–311 (1986)
Nair, V.N.: Chi-squared-type tests for ordered alternatives in contingency tables. J. Am. Stat. Assoc. **82**, 283–291 (1987)
Nelder, J.A., Wedderburn, R.W.M.: Generalized linear models. J. Roy. Stat. Soc. A **135**, 370–384 (1972)
Nenadić, O., Greenacre, M.: Correspondence analysis in R, with two- and three- dimensional graphics: the `ca` package. J. Stat. Software **20**, 1–13 (2007)
Newcombe, R.G.: Interval estimation for the difference between independent proportions: Comparison of eleven methods. Stat. Med. **17**, 873–890 (1998)
Newcombe, R.G.: A deficiency of the odds ratio as a measure of effect size. Stat. Med. **25**, 4235–4240 (2006)
Neyman, J.: Contributions to the theory of the χ^2 test. In: Neyman, J. (ed.) Proceedings of First Berkeley Symposium on Mathematical Statistics and Probability, pp. 239–273. University of California Press, Berkeley (1949)
Ng, K.W., Tang, M.L., Tan, M., Tian, G.L.: Grouped Dirichlet distribution: a new tool for incomplete categorical data analysis. J. Multivariate Anal. **99**, 490–509 (2008)
Nikiforov, A.M.: Exact Smirnov two-sample tests for arbitrary distributions. Appl. Stat. **43**, 265–284 (1994)
Ntzoufras, I., Forster, J.J., Dellaportas, P.: Stochastic search variable selection for log-linear models. J. Stat. Comput. Simul. **68**, 23–37 (2000)
O' Neill, M.E.: Asymptotic distributions of the canonical correlations from contingency tables. Aust. J. Stat. **20**, 75–82 (1978a)
O' Neill, M.E.: Distributional expansions for canonical correlations from contingency tables. J. Roy. Stat. Soc. B **40**, 303–312 (1978b)

O' Neill, M.E.: A note on the canonical correlations from contingency tables. Aust. J. Stat. **20**, 58–66 (1981)

Odoroff, C.L.: A comparison of minimum logit chi-square estimation and maximum likelihood estimation in $2 \times 2 \times 2$ and $3 \times 2 \times 2$ contingency tables: tests for interaction. J. Am. Stat. Assoc. **65**, 1617–1631 (1970)

Palmgren, J.: The Fisher information matrix for log-linear models arguing conditionally on observed explanatory variables. Biometrika **68**, 563–566 (1981)

Pardo, L.: Statistical Inference Based on Divergence Measures. Chapman & Hall, New York (2006)

Park, M., Lee, J.W., Kim, C.: Correspondence analysis approach for finding allele associations in population genetic study. Comput. Stat. Data Anal. **51**, 3145–3155 (2007)

Parsa, A.R., Smith, W.B.: Scoring under ordered constraints in contingency tables. Commun. Stat. Theory Meth. **22**, 3537–3551 (1993)

Parzen, M., Lipsitz, S., Ibrahim, J., Klar, N.: An estimate of the odds ratio that always exists. J. Comput. Graph. Stat. **11**, 420–436 (2002)

Pearce, N.: What does the odds ratio estimate in a case-control study? Int. J. Epidemiol. **22**, 1189–1192 (1993)

Pearson, E.S.: The choice of statistical tests illustrated on the interpretation of data classed in a 2×2 table. Biometrika **34**, 139–167 (1947)

Pearson, K.: 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. 5th Ser. **50**, 157–175 (1900a) (Reprinted in Karl Pearson's Early Statistical Papers ed. by E.S. Pearson. Cambridge University Press, Cambridge, 1948)

Pearson, K.: Mathematical contribution to the theory of evolution VII: On the correlation of characters not quantitatively measurable. Phil. Trans. Roy. Soc. **195A**, 1–47 (1900b)

Pearson, K.: 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, vol. 1. Dulau und Co., Londaon (1904) (Reprinted in Karl Pearson's Early Statistical Papers ed. by E.S. Pearson. Cambridge University Press, Cambridge, 1948)

Pearson, K.: On the probable error of a coefficient of correlation found from a fourfold table. Biometrika **9**, 22–27 (1913)

Pearson, K., Heron, D.: On theories of association. Biometrika **14**, 186–191 (1913)

Pepe, M.S., Janes, H., Longton, G., Leisenring, W., Newcomb, P.: Limitations of the odds ratio in gauging the performance of diagnositic, prognostic, or screening marker. Am. J. Epidemiol. **159**, 882–890 (2004)

Perkins, S.M., Becker, M.P.: Assessing rater agreement using marginal association models. Stat. Med. **21**, 1743–1760 (2002)

Permutt, T., Berger, V.W.: A new look at rank tests in ordered $2 \times k$ contingency tables. Commun. Stat. Theory Meth. **29**, 989–1003 (2000)

Pettersson, T.: A comparative study of model-based tests of independence for ordinal data using the bootstrap. J. Stat. Comput. Simul. **72**, 187–203 (2002)

Pierce, D.A., Schafer, D.W.: Residuals in generalized linear models. J. Am. Stat. Assoc. **81**, 977–986 (1986)

Pirie, W.R., Hamdan, M.A.: Some revised continuity corrections for discrete distributions. Biometrics **28**, 693–701 (1972)

Plackett, R.L.: The continuity correction in 2×2 tables. Biometrika **51**, 327–337 (1964)

Plackett, R.L.: A class of bivariate distributions. J. Am. Stat. Assoc. **60**, 516–522 (1965)

Plackett, R.L.: The Analysis of Categorical Data. Charles Griffin, London (1974)

Plackett, R.L.: Karl Pearson and the chi-squared test. Int. Stat. Rev. **51**, 59–72 (1983)

Poleto, F.Z., Paulino, C.D., Molenberghs, G., Singer, J.M.: Inferential implications of over-parametrization: a case study in incomplete categorical data. Int. Stat. Rev. **79**, 92–113 (2011)

Prentice, R.: Use of the logistic model in retrospective studies. Biometrics **32**, 599–606 (1976)

Raftery, A.E.: A note on Bayes factors for log-linear contingency table models with vague prior information. J. Roy. Stat. Soc. B **48**, 249–250 (1986)

Rao, P.V., Kupper, L.L.: Ties in paired comparison experiments: a generalization of the Bradley-Terry model. J. Am. Stat. Assoc. **62**, 192–204 (1967)

Rapallo, F.: Algebraic Markov bases and MCMC for two-way contingency tables. Scand. J. Stat. **30**, 385–397 (2003)

Rapallo, F.: Algebraic exact inference for rater agreement models. Stat. Meth. Appl. **14**, 45–66 (2005)

Rapallo, F.: Markov bases and structural zeros. J. Symbolic Comput. **41**, 164–172 (2006)

Rapallo, F.: Outliers and patterns of outliers in contingency tables with algebraic statistics. Scand. J. Stat. **39**, 784–797 (2012)

Rasch, G.: Probabilistic Models for Some Intelligence and Attainment Tests (Danish Institute for Educational Research, Copenhagen), expanded edition. The University of Chicago Press, Chicago (Original work published in 1960) (1980)

Rayner, J.C.W., Best, D.J.: Analysis of singly ordered two-way contingency tables. J. Appl. Math. Decis. Sci. **4**, 83–98 (2000)

Read, C.B.: Tests of symmetry in three-way contingency tables. Psychometrika **43**, 409–420 (1978)

Read T.R.C., Cressie, N.A.C.: Goodness-of-Fit Statistics for Discrete Multivariate Data. Springer, New York (1988)

Riedwyl, H., Schüpbach, M.: Siebdiagramme: Graphische Darstellung von Kontingenztafeln. Technical Report No. 12. Institute for Mathematical Statistics, University of Bern, Bern, Switzerland (1983)

Riedwyl, H., Schüpbach, M.: Parquet diagram to plot contingency tables. In: Faulbaum, F. (ed.) Softstat'93: Advances in Statistical Software, pp. 293–299. Gustav Fischer, New York (1994)

Ritov, Y., Gilula, Z.: The order-restricted RC model for ordered contingency tables: estimation and testing of fit. Ann. Stat. **19** 2090–2101 (1991)

Ritov, Y., Gilula, Z.: Analysis of contingency tables by correspondence models subject to order constraints. J. Am. Stat. Assoc. **88**, 1380–1387 (1993)

Robertson, T., Wright, F.T.: Likelihood-ratio tests for and against a stochastic ordering between multinomial populations. Ann. Stat. **9**, 1248–1257 (1981)

Rom, D., Sarkar, S.K.: Approximating probability integrals of multivariate normal using association models. J. Stat. Comput. Simul. **35**, 109–119 (1990)

Rom, D., Sarkar, S.K.: A generalized model for the analysis of association in ordinal contingency tables. J. Stat. Plann. Infer. **33**, 205–212 (1992)

Roverato, A.: A unified approach to the characterisation of Markov equivalence classes of DAGs, chain graphs with no flags and chain graphs. Scand. J. Stat. **32**, 295–312 (2005)

Royall, R.: On the probability of observing misleading statistical evidence (with discussion). J. Am. Stat. Assoc. **95**, 760–780 (2000)

Royall, R., Tsou, T.S.: Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. J. Roy. Stat. Soc., Ser. B **65**, 391–404 (2003)

Rudas, T.: Odds Ratios in the Analysis of Contingency Tables. Series: Quantitative Applications in the Social Sciences. Sage Publications, Thousand Oaks (1998)

Rudas, T., Bergsma, W.: Letter to the editor: Reconsidering the odds ratio as a measure of 2×2 association in a population, by H. C. Kraemer (Stat. Med. 2004; 23, 257–270). Stat. Med. **23**, 3545–3548 (2004)

Rudas, T., Bergsma, W., Nemeth, R.: Marginal log-linear parameterization of conditional independence models. Biometrika **97**, 1006–1012; Faulbaum, F. (ed.) Softstat '93: Advances in Statistical Software, pp. 293–299. Gustav Fischer, New York (2010)

Samuels, M.L.: Simpson's paradox and related phenomena. J. Am. Stat. Assoc. **88**, 81–88 (1993)

Santner, T.J., Duffy, D.E.: The Statistical Analysis of Discrete Data. Springer, New York (1989)

Sarkar, S.K.: Quasi-independence in ordinal triangular contingency tables. J. Am. Stat. Assoc. **84**, 592–597 (1989)

Sauermann, W.: Bootstrapping the maximum likelihood estimator in high-dimensional log-linear models. Ann. Stat. **17**, 1198–1216 (1989)

Savage, I.R., Deutsch, K.A.: A statistical model of the gross analysis of transaction flows. Econometrica **28**, 551–572 (1960)
Schriever, B.F.: Scaling of order dependent categorical variables with correspondence analysis. Int. Stat. Rev. **51**, 225–238 (1983)
Schuster, C.: A mixture model approach to indexing rater agreement. Br. J. Math. Stat. Psychol. **55**, 289–303 (2002)
Schuster, C.: A note on the interpretation of weighted kappa and its relation to other rater agreement statistics for metric scales. Educ. Psychol. Meas. **55**, 243–253 (2004)
Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
Scott, D.W., Tapia, R.A., Thompson, J.R.: Nonparametric probability density estimation by discrete maximum penalized-likelihood criteria. Ann. Stat. **8**, 820–832 (1980)
Semenya, K., Koch, G.G., Stokes, M.E., Forthofer, R.N.: Linear models methods for some rank function analyses of ordinal categorical data. Commun. Stat. Part A **12**, 1277–1298 (1983)
Shaked, M., Shanthikumar, J.G.: Stochastic Orders. Springer, New York (2007)
Shapiro, S.H.: Collapsing contingency tables: a geometric approach. Am. Stat. **36**, 43–46 (1982)
Siciliano, R., Mooijaart, A.: Three-factor association models for three-way contingency tables. Comput. Stat. Data Anal. **24**, 337–356 (1997)
Silva Mato, A., Martín Andrés, A.: Simplifying the calculation of the *P*-value for Barnard's test and its derivatives. Stat. Comput. **7**, 137–143 (1997)
Silvapulle, M.J., Sen, P.K.: Constrained Statistical Inference. Wiley, New York (2005)
Simon, G.A.: Alternative analysis for the singly ordered contingency table. J. Am. Stat. Assoc. **69**, 971–976 (1974)
Simon, G.A.: Efficacies of measures of association for ordinal contingency tables. J. Am. Stat. Assoc. **73**, 545–551 (1978)
Simon, S.D.: Understanding the odds ratio and the relative risk. J. Androl. **22**, 533–536 (2001)
Simonoff, J.S.: A penalty function approach to smoothing large sparse contingency tables. Ann. Stat. **11**, 208–218 (1983)
Simonoff, J.S.: Detecting outlying cells in two-way contingency tables via backward-stepping. Technometrics **30**, 339–345 (1988)
Simonoff, J.S.: Smoothing categorical data. J. Stat. Plann. Infer. **47**, 41–69 (1995)
Simonoff, J.S.: Three sides of smoothing: categorical data smoothing, nonparametric regression, and density estimation. Int. Stat. Rev. **66**, 137–156 (1998)
Simonoff, J.S.: Analyzing Categorical Data. Springer, New York (2003)
Simpson, E.H.: The interpretation of interaction in contingency tables. J. Roy. Stat. Soc. Ser. B **13**, 238–241 (1951)
Sobel, M.E.: Some Models for the multiway contingency table with a one-to-one correspondence among categories. Socio. Meth. **43**, 165–192 (1988)
Sobel, M.E., Becker, M.P., Minick, S.P.: Origins, destinations, and association in occupational mobility. Am. J. Socio. **104**, 687–721 (1998)
Somers, R.H.: A new asymmetric measure of association for ordinal variables. Am. Socio. Rev. **27**, 799–811 (1962)
Spall, J.C.: Monte Carlo computation of the Fisher information matrix in nonstandard settings. J. Comput. Graph. Stat. **14**, 889–909 (2005)
Spiegelhalter, D.J., Smith, A.F.M.: Bayes factors for linear and log-linear models with vague prior information. J. Roy. Stat. Soc. Ser. B **44**, 377–387 (1982)
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A.: Bayesian measures of model complexity and fit (with discussion). J. Roy. Stat. Soc. Ser. B **64**, 583–639 (2002)
Spilerman, S.: Extensions of the mover-stayer model. Am. J. Socio. **78**, 599–626 (1972)
Spitzer, R.L., Cohen, J., Fleiss, J.L., Endicott, J.: Quantification of agreement in psychiatric diagnosis. Arch. Gen. Psychiatry **17**, 83–87 (1967)
Stigler, S.M.: Studies in the history of probability and statistics XLIII. Karl Pearson and quasi-independence. Biometrika **79**, 563–575 (1992)
Stigler, S.M.: Karl Pearson's theoretical errors and the advances they inspired. Stat. Sci. **23**, 261–271 (2008)

Stokes, M.E., Davis, C.S., Koch, G.G.: Categorical Data Analysis Using SAS, 3rd edn. SAS Institute Inc., Cary (2012)
Streitberg, B.: Exploring interactions in high-dimensional tables: a bootstrap alternative to log-linear models. Ann. Stat. **2**, 405–413 (1999)
Stuart, A.: The estimation and comparison of strengths of association in contingency tables. Biometrika **40**, 105–110 (1953)
Stuart, A.: A test for homogeneity of the marginal distributions in a two-way classification. Biometrika **42**, 412–416 (1955)
Stukel, T.A.: Generalized logistic models. J. Am. Stat. Assoc. **83**, 426–431 (1988)
Su, Y., Zhou, M.: On a connection between the Bradley-Terry model and the Cox proportional hazards model. Stat. Probab. Lett. **76**, 698–702 (2006)
Suissa, S., Shuster, J.J.: Exact unconditional samples sizes for the 2 by 2 binomial trial. J. Roy. Stat. Soc. Ser. A **148**, 317–327 (1985)
Tanner, M.A., Young, M.A.: Modeling agreement among raters. J. Am. Stat. Assoc. **80**, 175–180 (1985)
Tarantola, C., Consonni, G., Dellaportas, P.: Bayesian clustering of row effects models. J. Stat. Plann. Infer. **138**, 2223–2235 (2008)
Tarone, R.E.: On heterogeneity tests based on efficient scores. Biometrika **72**, 91–95 (1985)
Ten Berge, J.M.F.: Simplicity and typical rank results for three-way arrays. Psychometrika **76**, 3–12 (2011)
Tenenhaus, M., Young, F.W.: An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. Psychometrika **50**, 91–119 (1985)
Theus, M., Lauer, S.R.W.: Visualizing loglinear models. J. Comput. Graph. Stat. **8**, 396–412 (1999)
Titterington, D.M., Bowman, A.W.: A comparative study of smoothing procedures for ordered categorical data. J. Stat. Comput. Simul. **21**, 291–312 (1985)
Touloumis, A.: `multgee`: GEE solver for correlated nominal or ordinal multinomial responses. `R` package version 1.1 (2012)
Touloumis, A., Agresti, A., Kateri, M.: GEE for multinomial responses using a local odds ratios parameterization. Biometrics **69**, 633–640 (2013)
Train, K.: Discrete Choice Methods with Simulation. Cambridge University Press, New York (2009)
Tsai, M.T., Sen, P.K.: A test of quasi-independence in ordinal triangular contingency tables. Statistica Sinica **5**, 767–780 (1995)
Tukey, J.W.: One degree of freedom for non-additivity. Biometrics **5**, 232–242 (1949)
Tunaru, R.: Models of association versus causal models for contingency tables. J. Roy. Stat. Soc. Ser. D **50**, 257–269 (2001)
Turner, H., Firth, D.: gnm: a package for generalized nonlinear models. Newsletter R Proj. **7/2**, 8–12 (2007). http://www.r-project.org/doc/Rnews/Rnews-2007-2.pdf
Turner, H., Firth, D.: Generalized nonlinear models in R: an overview of the `gnm` package. http://cran.r-project.org/web/packages/gnm/vignettes/gnmOverview.pdf (2012a)
Turner, H., Firth, D.: Bradley-Terry models in `R`: the `BradleyTerry2` package. J. Stat. Software **48**, 1–21 (2012b)
Tutz, G.: Bradley-Terry-Luce models with an ordered response. J. Math. Psychol. **30**, 306–316 (1986)
Tutz, G.: Regression for Categorical Data. Cambridge University Press, New York (2012)
Uebersax, J.S.: Statistical modeling of expert ratings on medical treatment appropriateness. J. Am. Stat. Assoc. **88**, 421–427 (1993)
Uebersax, J.S., Grove, W.M.: Latent class analysis of diagnostic agreement. Stat. Med. **9**, 559–572 (1990)
Uebersax, J.S., Grove, W.M.: A latent trait finite mixture model for the analysis of rating agreement. Biometrics **49**, 823–835 (1993)
Upton, G.J.G.: The Analysis of Cross-Tabulated Data. Wiley, New York (1978)

Upton, G.J.G.: A comparison of alternative tests for the 2×2 comparative trial. J. Roy. Stat. Soc. Ser. A **145**, 86–105 (1982)
Upton, G.J.G.: Fisher's exact test. J. Roy. Stat. Soc. Ser. A **155**, 395–402 (1992)
Valet, F., Guinot, C., Yves Mary, J.: Log-linear non-uniform association models for agreement between two ratings on an ordinal scale. Stat. Med. **26**, 647–662 (2007)
van de Geer, S.A.: High-dimensional generalized linear models and the lasso. Ann. Stat. **36**, 614–645 (2008)
van de Velden, M., Kiers, H.A.L.: Rotation in correspondence analysis. J. Classification **22**, 251–271 (2005)
van den Hout, A., van der Heijden, P.G.M.: Randomized response, statistical disclosure control and misclassification: a review. Int. Stat. Rev. **70**, 269–288 (2002)
van der Heijden, P., de Leeuw, J.: Correspondence analysis used complimentary to log-linear analysis. Psychometrika **50**, 429–447 (1985)
van der Heijden, P.G.M., de Falguerolles, A., de Leeuw, J.: A combined approach to contingency table analysis using correspondence analysis and log-linear analysis. Appl. Stat. **38**, 249–292 (1989)
van Mechelen, I., Bock, H.H., de Boeck, P.: Two-mode clustering methods: a structured overview. Stat. Meth. Med. Res. **13**, 363–394 (2004)
Vellaisamy, P., Vijay, V.: Some collapsibility results for n-dimensional contingency tables. Ann. Inst. Stat. Math. **59**, 557–576 (2007)
Vellaisamy, P., Vijay, V.: Collapsibility of contingency tables based on conditional models. J. Stat. Plann. Infer. **140**, 1243–1255 (2010)
Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S, 4th edn. Springer, New York (2002)
Vermunt, J.K, Rodrigo, M.F., Ato-Garcia, M.: Modeling joint and marginal distributions in the analysis of categorical panel data. Socio. Meth. Res. **30**, 170–196 (2001)
Viele, K., Srinivasan, C.: Parsimonious Estimation of multiplicative interaction in analysis of variance using Kullback-Leibler information. J. Stat. Plann. Infer. **84**, 201–219 (2000)
von Davier, M., Carstensen, C.H.: Multivariate and Mixture Distribution Rasch Models: Extensions and Applications. Springer, New York (2007)
Wahrendorf, J.: Inference in contingency tables with ordered categories using Plackett's coefficient of association for bivariate distributions. Biometrika **67**, 15–21 (1980)
Waite, H.: Association of finger-prints. Biometrika **10**, 421–478 (1915)
Walker, S.H., Dunkan, D.B.: Estimation of the probability of an event as a function of several independent variables. Biometrika **54**, 167–179 (1967)
Walter, S.D.: Small-sample estimation of log odds ratios from logistic regression and fourfold tables. Stat. Med. **4**, 437–444 (1985)
Walter, S.D., Cook, R.J.: A comparison of several point estimators of the odds ratio in a single 2×2 contingency table. Biometrics **47**, 795–811 (1991)
Wang, Y.J.: The probability integrals of bivariate normal distributions: a contingency table approach. Biometrika **74**, 1985–1990 (1987)
Warrens, M.J.: On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions. Psychometrika **73**, 777–789 (2008)
Warrens, M.J.: Conditional inequalities between Cohen's kappa and weighted kappas. Stat. Meth. **10**, 14–22 (2013)
Webb, E.L., Forster, J.J.: Bayesian model determination for multivariate ordinal and binary data. Comput. Stat. Data Anal. **52**, 2632–2649 (2008)
Weller, S., Romney, A.K.: Metric Scaling Correspondence Analysis. Quantitative Applications in the Social Sciences. Sage University, Newbury Park (1990)
Wermuth, N.: Parametric collapsibility and the lack of moderating effects in contingency tables with a dichotomous response variable. J. Roy. Stat. Soc. Ser. B **49**, 353–364 (1987)
Wermuth, N., Cox, D.R.: On the application of conditional independence to ordinal data. Int. Stat. Rev. **66**, 181–199 (1998)

Wermuth, N., Lauritzen, S.L.: Graphical and recursive models for contingency tables. Biometrika **70**, 537–552 (1983)
Wermuth, N., Lauritzen, S.L.: On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). J. Roy. Stat. Soc. Ser. B **52**, 21–72 (1990)
Whaley, F.S.: Comparison of different maximum likelihood estimators in a small sample logistic regression with two independent binary variables. Stat. Med. **10**, 723–731 (1991)
White, H.C.: Cause and effect in social mobility tables. Behav. Sci. **7**, 14–27 (1963)
Whittaker, J.: Graphical Models in Applied Multivariate Statistics. Wiley, New York (1990)
Whittemore, A.S.: Collapsibility of multidimensional contingency tables. J. Roy. Stat. Soc. Ser. B **40**, 328–340 (1978)
Wilks, S.S.: The likelihood test of independence in contingency tables. Ann. Math. Stat. **6**, 190–196 (1935)
Williams, E.J.: The interpretations of interactions in factorial experiments. Biometrika **39**, 65–81 (1952)
Williams, O.D., Grizzle, J.E.: Analysis of contingency tables having ordered response categories. J. Am. Stat. Assoc. **67**, 55–63 (1972)
Williamson, J.M., Kim, K.M.: A global odds ratio regression model for bivariate ordinal categorical data from ophthalmologic studies. Stat. Med. **15**, 1507–1518 (1996)
Williamson, J.M., Kim, K.M., Lipsitz, S.R.: Analyzing bivariate ordinal data using a global odds ratio. J. Am. Stat. Assoc. **90**, 1432–1437 (1995)
Wong, R.S.K.: Multidimensional association models: a multilinear approach. Socio. Meth. Res. **30**, 197–240 (2001)
Wong, R.S.K.: Association models. Quantitative Applications in the Social Sciences. Sage Publications,Thousand Oaks (2010)
Woolf, B.: On estimating the relation between blood group and disease. Ann. Hum. Genet. **19**, 251–253 (1955)S
Xie, Y.: The log-multiplicative layer effect model for comparing mobility tables. Am. Socio. Rev. **57**, 380–95 (1992)
Yang, I., Becker, M.P: Latent variable modeling of diagnostic accuracy. Biometrics **53**, 948–958 (1997)
Yates, F.: Contingency tables involving small numbers and the χ^2 test. J. Roy. Stat. Soc. Suppl. **1**, 217–235 (1934)
Yates, F.: The analysis of contingency tables with groupings based on quantitative characters. Biometrika **35**, 176–181 (1948)
Yates, F.: Tests of significance for 2×2 contingency tables. J. Roy. Stat. Soc. Ser. A **147**, 426–463 (1984)
Yee, T.W.: The VGAM package. Newsletter R Proj. **8/2**, 28–39 (2008). http://CRAN.R-project.org/doc/Rnews/Rnews-2008-2.pdf
Yee, T.W., Wild, C.J.: Vector generalized additive models. J. Roy. Stat. Soc. Ser. B **58**, 481–493 (1996)
Yelland, P.M.: An introduction to correspondence analysis. Math. J. **12**, 1–23 (2010)
Yule, G.U.: On the association of attributes in statistics: with illustrations from the material of the Childhood Society & c. Phil. Trans. Ser. A **194**, 257–319 (1900)
Yule, G.U.: Notes on the theory of association of attributes in statistics. Biometrika **2**, 121–134 (1903)
Yule, G.U.: On the methods of measuring association between two attributes. J. Roy. Stat. Soc. **75**, 579–642 (1912)
Zeileis, A., Meyer, D., Hornik, K.: Residual-based shadings for visualizing (conditional) independence. J. Comput. Graph. Stat. **16**, 507–525 (2007)
Zelterman, D.: Goodness-of-fit tests for large sparse multinomial distributions. J. Am. Stat. Assoc. **82**, 624–629 (1987)
Zelterman, D.: Models for Discrete Data. Oxford University Press, New York (2006)

Index

M. Kateri, *Contingency Table Analysis: Methods and Implementation Using R*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4811-4,