Springer Series in Statistics

Markov Bases in Algebraic Statistics

Springer

Springer Series in Statistics

For further volumes:
http://www.springer.com/series/692

Satoshi Aoki • Hisayuki Hara • Akimichi Takemura

Markov Bases in Algebraic Statistics

Springer

Satoshi Aoki
Department of Mathematics
and Computer Science
Graduate School of Science
and Engineering
Kagoshima University
Kagoshima, Japan

Hisayuki Hara
Faculty of Economics
Niigata University
Niigata, Japan

Akimichi Takemura
Department of Mathematical Informatics
Graduate School of Information Science
and Technology
University of Tokyo
Tokyo, Japan

ISSN 0172-7397
ISBN 978-1-4614-3718-5 ISBN 978-1-4614-3719-2 (eBook)
DOI 10.1007/978-1-4614-3719-2
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012938710

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Algebraic statistics is a rapidly developing field, where ideas from statistics and algebra meet and stimulate new research directions. Statistics has been relying on classical asymptotic theory as a basis for statistical inferences. This classical basis is still very useful. However, when the validity of asymptotic theory is in doubt, for example, when the sample size is small, statisticians rely more and more on various computational methods. Similarly, algebra has long been considered as the purest field of mathematics, far apart from practical computations. However, due mainly to the development of Gröbner basis technology, algebra is now becoming a field where computations for practical applications are feasible. It is an interesting trend, because historically algebra was invented to speed up various calculations.

These two trends meet in the field of algebraic statistics. Algebraic algorithms are now very useful and essential for some practical statistical computations such as Markov chain Monte Carlo tests for discrete exponential families, which is the main topic of this book. On the other hand algebraic structures and computational needs of statistical models provide new challenging problems to algebraists. Some algebraic structures are naturally motivated from statistical modeling, but not necessarily from pure mathematical considerations.

Algebraic statistics has two origins. One origin is the work by Pistone and Wynn in 1996 on the use of Gröbner bases for studying confounding relations in factorial designs of experiments. Another origin is the work by Diaconis and Sturmfels in 1998 on the use of Gröbner bases for constructing a connected Markov chain for performing conditional tests of a discrete exponential family. These two works opened up the whole new field of algebraic statistics. In this book we take up the second topic. We give a detailed treatment of results following the seminal work of Diaconis and Sturmfels. We also briefly consider the first topic in Chap. 15 of this book.

As a general reference to the first origin of algebraic statistics we mention Pistone et al. [118]. For the second origin we mention Drton et al. [55], Pachter and Sturmfels [116], and our review paper [15]. For Japanese people the following two books are very useful: Hibi [86], and JST CREST Hibi team [93]. The Markov bases database (`http://markov-bases.de/`) provides very useful online material for studying Markov bases.

Algebraic statistics gave us some exciting opportunities for research and collaboration. In particular we enjoyed working with Takayuki Hibi and Hidefumi Ohsugi, who are the leading researchers on Gröbner bases in Japan. Since 2008 Takayuki Hibi has a project, "Harmony of Gröbner Bases and the Modern Industrial Society," in the mathematics program of the Japan Science and Technology Agency. Algebraic statistics offers a rare ground where algebraists and statisticians can talk about the same problems, albeit often with different terminologies. This book is intended for statisticians with minimal backgrounds in algebra. As we ourselves learned algebraic notions through working on statistical problems, we hope that this book with many practical statistical problems is useful for statisticians to start working on algebraic statistics.

In preparing this book we very much benefited from comments of Takayuki Hibi, Hidehiko Kamiya, Kei Kobayashi, Satoshi Kuriki, Mitsunori Ogawa, Hidefumi Ohsugi, Toshio Sakata, Tomonari Sei, Kentaro Tanaka, and Ruriko Yoshida.

Finally we acknowledge great editorial help from John Kimmel.

Kagoshima, Japan Satoshi Aoki
Niigata, Japan Hisayuki Hara
Tokyo, Japan Akimichi Takemura

Contents

Part I Introduction and Some Relevant Preliminary Material

1 Exact Tests for Contingency Tables and Discrete Exponential Families 3
- 1.1 Independence Model of 2×2 Two-Way Contingency Tables 3
- 1.2 2×2 Contingency Table Models as Discrete Exponential Family 8
- 1.3 Independence Model of General Two-Way Contingency Tables 10
- 1.4 Conditional Independence Model of Three-Way Contingency Tables 14
  - 1.4.1 Normalizing Constant of Hypergeometric Distribution for the Conditional Independence Model 18
- 1.5 Notation of Hierarchical Models for *m*-Way Contingency Tables 19

2 Markov Chain Monte Carlo Methods over Discrete Sample Space 23
- 2.1 Constructing a Connected Markov Chain over a Conditional Sample Space: Markov Basis 23
- 2.2 Adjusting Transition Probabilities by Metropolis–Hastings Algorithm 27

3 Toric Ideals and Their Gröbner Bases 33
- 3.1 Polynomial Ring 33
- 3.2 Term Order and Gröbner Basis 35
- 3.3 Buchberger's Algorithm 38
- 3.4 Elimination Theory 39
- 3.5 Toric Ideals 39

Part II Properties of Markov Bases

4 Definition of Markov Bases and Other Bases 47
4.1 Discrete Exponential Family 47
4.2 Definition of Markov Basis 50
4.3 Properties of Moves and the Lattice Basis 51
4.4 The Fundamental Theorem of Markov Basis 54
4.5 Gröbner Basis from the Viewpoint of Markov Basis 59
4.6 Graver Basis, Lawrence Lifting, and Logistic Regression 60

5 Structure of Minimal Markov Bases 65
5.1 Accessibility by a Set of Moves 65
5.2 Structure of Minimal Markov Basis and Indispensable Moves 66
5.3 Minimum Fiber Markov Basis 71
5.4 Examples of Minimal Markov Bases 72
5.4.1 One-Way Contingency Tables 72
5.4.2 Independence Model of Two-Way Contingency Tables 73
5.4.3 The Unique Minimal Markov Basis for the Lawrence Lifting 73
5.5 Indispensable Monomials 75

6 Method of Distance Reduction 79
6.1 Distance Reducing Markov Bases 79
6.2 Examples of Distance-Reducing Proofs 81
6.2.1 The Complete Independence Model of Three-Way Contingency Tables 81
6.2.2 Hardy–Weinberg Model 83
6.3 Graver Basis and 1-Norm Reducing Markov Bases 85
6.4 Some Results on Minimality of 1-Norm Reducing Markov Bases 86

7 Symmetry of Markov Bases 91
7.1 Motivations for Invariance of Markov Bases 91
7.2 Examples of Invariant Markov Bases 92
7.3 Action of Symmetric Group on the Set of Cells 93
7.4 Symmetry of a Toric Model and the Largest Group of Invariance 96
7.5 The Largest Group of Invariance for the Independence Model of Two-Way Tables 98
7.6 Characterizations of a Minimal Invariant Markov Basis 100

Part III Markov Bases for Specific Models

8 Decomposable Models of Contingency Tables 109
8.1 Chordal Graphs and Decomposable Models 109
8.2 Markov Bases for Decomposable Models 111

8.3 Structure of Degree 2 Fibers 113
8.4 Minimal Markov Bases for Decomposable Models 115
8.5 Minimal Invariant Markov Bases 119
8.6 The Relation Between Minimal and Minimal Invariant Markov Bases 127

9 Markov Basis for No-Three-Factor Interaction Models and Some Other Hierarchical Models 129
9.1 No-Three-Factor Interaction Models for $3 \times 3 \times K$ Contingency Tables 129
9.2 Unique Minimal Markov Basis for $3 \times 3 \times 3$ Tables 130
9.3 Unique Minimal Markov Basis for $3 \times 3 \times 4$ Tables 139
9.4 Unique Minimal Markov Basis for $3 \times 3 \times 5$ and $3 \times 3 \times K$ Tables for $K > 5$ 142
9.5 Indispensable Moves for Larger Tables 145
9.6 Reducible Models 149
9.7 Markov Basis for Reducible Models 150
9.8 Markov Complexity and Graver Complexity 153
9.9 Markov Width for Some Hierarchical Models 156

10 Two-Way Tables with Structural Zeros and Fixed Subtable Sums 159
10.1 Markov Bases for Two-Way Tables with Structural Zeros 159
10.1.1 Quasi-Independence Model in Two-Way Incomplete Contingency Tables 159
10.1.2 Unique Minimal Markov Basis for Two-Way Quasi-Independence Model 161
10.1.3 Enumerating Elements of the Minimal Markov Basis 164
10.1.4 Numerical Example of a Quasi-Independence Model 167
10.2 Markov Bases for Subtable Sum Problem 168
10.2.1 Introduction of Subtable Sum Problem 168
10.2.2 Markov Bases Consisting of Basic Moves 169
10.2.3 Markov Bases for Common Diagonal Effect Models 172
10.2.4 Numerical Examples of Common Diagonal Effect Models 176

11 Regular Factorial Designs with Discrete Response Variables 181
11.1 Conditional Tests for Designed Experiments with Discrete Observations 181
11.1.1 Conditional Tests for Log-Linear Models of Poisson Observations 181
11.1.2 Models and Aliasing Relations 184
11.1.3 Conditional Tests for Logistic Models of Binomial Observations 191
11.1.4 Example: Wave-Soldering Data 193

11.2 Markov Bases and Corresponding Models for Contingency Tables 194
11.2.1 Rewriting Observations as Frequencies of a Contingency Table 194
11.2.2 Models for the Two-Level Regular Fractional Factorial Designs with 16 Runs 200
11.2.3 Three-Level Regular Fractional Factorial Designs and 3^{s-k} Continent Tables 203

12 Groupwise Selection Models 209
12.1 Examples of Groupwise Selections 209
12.1.1 The Case of National Center Test in Japan 209
12.1.2 The Case of Hardy–Weinberg Models for Allele Frequency Data 212
12.2 Conditional Tests for Groupwise Selection Models 213
12.2.1 Models for NCT Data 214
12.2.2 Models for Allele Frequency Data 215
12.3 Gröbner Basis for Segre–Veronese Configuration 217
12.4 Sampling from the Gröbner Basis for the Segre–Veronese Configuration 219
12.5 Numerical Examples 219
12.5.1 The Analysis of NCT Data 219
12.5.2 The Analysis of Allele Frequency Data 221

13 The Set of Moves Connecting Specific Fibers 229
13.1 Discrete Logistic Regression Model with One Covariate 229
13.2 Discrete Logistic Regression Model with More than One Covariate 231
13.3 Numerical Examples 238
13.3.1 Exact Tests of Logistic Regression Model 238
13.4 Connecting Zero-One Tables with Graver Basis 240
13.5 Rasch Model 241
13.6 Many-Facet Rasch Model 242
13.7 Latin Squares and Zero-One Tables for No-Three-Factor Interaction Models 245

Part IV Some Other Topics of Algebraic Statistics

14 Disclosure Limitation Problem and Markov Basis 251
14.1 Swapping with Some Marginals Fixed 251
14.2 *E*-Swapping 252
14.3 Equivalence of Degree-Two Square-Free Move of Markov Bases and Swapping of Two Records 253
14.4 Swappability Between Two Records 254
14.5 Searching for Another Record for Swapping 257

15 Gröbner Basis Techniques for Design of Experiments 261
15.1 Design Ideals 261
15.2 Identifiability of Polynomial Models and the Quotient with Respect to the Design Ideal 262
15.3 Regular Two-Level Designs 267
15.4 Indicator Functions 269

16 Running Markov Chain Without Markov Bases 275
16.1 Performing Conditional Tests When a Markov Basis Is Not Available 275
16.2 Sampling Contingency Tables with a Lattice Basis 275
16.3 A Lattice Basis for Higher Lawrence Configuration 277
16.4 Numerical Experiments 278
16.4.1 No-Three-Factor Interaction Model 278
16.4.2 Discrete Logistic Regression Model 282

References 287

Index 295

Part I
Introduction and Some Relevant Preliminary Material

In Part I of this book we give introductory material on performing exact tests using Markov basis and a short survey on Gröbner basis.

In Chap. 1, using the example of Fisher's exact test for the independence model in two-way contingency tables, we give an introduction to exact tests. We also discuss conditional independence model for three-way contingency tables.

In Chap. 2 we discuss basic notions of Markov chain and Markov bases. In particular we explain the Metropolis-Hastings procedure for adjusting transition probabilities to achieve a desired stationary distribution.

Chapter 3 is a brief summary of results in the theory of Gröbner basis. In this chapter we collect relevant facts on ideals in polynomial rings and their Gröbner bases, which are often needed for discussion of Markov bases.

In this book, $\mathbb{R}, \mathbb{Q}, \mathbb{Z}, \mathbb{N} = \{0, 1, \dots\}$ stand for the set of reals, rationals, integers and nonnegative integers, respectively. For a positive integer n, we denote the set of n-dimensional vectors of elements from $\mathbb{R}, \mathbb{Q}, \mathbb{Z}, \mathbb{N}$, by $\mathbb{R}^n, \mathbb{Q}^n, \mathbb{Z}^n, \mathbb{N}^n$, respectively.

Chapter 1
Exact Tests for Contingency Tables and Discrete Exponential Families

1.1 Independence Model of 2×2 Two-Way Contingency Tables

The theory of exact tests for discrete exponential families is best explained by Fisher's exact test of homogeneity of two binomial populations and the independence model of 2×2 contingency tables. We begin with the test of homogeneity of two binomial populations. An excellent introduction to contingency tables is given in [59]. We also refer to Agresti [3] as a survey paper of the exact methods.

Fisher's exact test can be applied to three different sampling schemes: (i) test of homogeneity of two binomial populations, (ii) test of independence in multinomial sampling for 2×2 tables, (iii) the main effect model for logarithms of mean parameters of independent Poisson random variables in 2×2 tables. We discuss these three sampling schemes in this order. With this example we confirm that the same Markov basis can be used for different sampling schemes.

Let X be distributed according to a binomial distribution $\mathrm{Bin}(n_1, p_1)$, where n_1 is the number of trials and p_1 is the success probability. Let Y be distributed according to the binomial distribution $\mathrm{Bin}(n_2, p_2)$. Suppose that X and Y are independent. We can display X and Y in the following 2×2 contingency table:

$$\begin{array}{|c|c|c} \hline X & n_1 - X & n_1 \\ \hline Y & n_2 - Y & n_2 \\ \hline \multicolumn{1}{c}{t} & \multicolumn{1}{c}{n-t} & n \end{array}$$

where $t = X + Y$ and $n = n_1 + n_2$. The hypothesis of homogeneity of two binomial populations is specified as

$$H : p_1 = p_2.$$

S. Aoki et al., *Markov Bases in Algebraic Statistics*, Springer Series in Statistics 199, DOI 10.1007/978-1-4614-3719-2_1,

The joint probability function of X and Y is written as

$$p(x,y) = \binom{n_1}{x} p_1^x (1-p_1)^{n_1-x} \binom{n_2}{y} p_2^y (1-p_2)^{n_2-y}.$$

Note that here we are using the conventional notational distinction between random variables X,Y in capital letters and their values x,y in lower-case letters. However, for the rest of this book for notational simplicity we do not necessarily stick to this convention.

Under the null hypothesis H, the joint probability is written as

$$p(x,y) = \binom{n_1}{x} \binom{n_2}{y} p_1^{x+y} (1-p_1)^{n-(x+y)}. \tag{1.1}$$

This joint probability depends on (x,y) through $t = x+y$. Therefore from the factorization theorem for sufficient statistics (see Sect. 2.6 of Lehmann and Romano [98]), $T = X+Y$ is a sufficient statistic under the null hypothesis H. Given $T = t$, the conditional distribution of X does not depend on the value of $p_1 = p_2$. Hence by using X as the test statistic, we obtain a testing procedure, whose level does not depend on the value of $p_1 = p_2$; that is, we obtain a *similar test* (Sect. 4.3 of [98]).

Under H the distribution of $T = X+Y$ is the binomial distribution $\mathrm{Bin}(n, p_1)$. Therefore the conditional distribution of X given $T = t$ is calculated as

$$\begin{aligned} P(X = x \mid T = t) &= \frac{\binom{n_1}{x}\binom{n_2}{t-x} p_1^t (1-p_1)^{n-t}}{\binom{n_1+n_2}{t} p_1^t (1-p_1)^{n-t}} = \frac{\binom{n_1}{x}\binom{n_2}{t-x}}{\binom{n}{t}} \\ &= \frac{n_1!n_2!t!(n-t)!}{n!x!(n_1-x)!(t-x)!(n_2-t+x)!}. \end{aligned} \tag{1.2}$$

This is a *hypergeometric distribution*. Indeed the conditional distribution does not depend on the value of $p_1 = p_2$.

The null hypothesis H is rejected if the value of X is too large or too small. Because the distribution of X is not symmetric when $n_1 \neq n_2$, the rejection region is usually determined by unbiasedness consideration. For optimality of similar unbiased test see Sect. 4.4 of [98]. This testing procedure is called Fisher's exact test. It is an exact test because the significance level is computed from the hypergeometric distribution. It is also called a *conditional test* because we use the conditional null distribution given $T = t$. In contrast, the usual large-sample test is based on the large-sample normal approximation to the following "z-statistic":

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}, \qquad \hat{p}_1 = \frac{X}{n_1},\ \hat{p}_2 = \frac{Y}{n_2}.$$

Table 1.1
Cross-classification of belief in afterlife by gender

	Belief in Afterlife	
Gender	Yes	No or Undecided
Females	509	116
Males	398	104

The test based on z is an unconditional test. However, when the sample size is small, it is desirable to use the exact test (Haberman [68]).

In the case of homogeneity of two binomial populations, we saw that $X+Y$ (total number of successes) is a sufficient statistic. We could also take $n-X-Y$ (total number of failures) or even the pair $(X+Y, n-X-Y)$ as a sufficient statistic. Note that the pair contains redundancy, but it is still a sufficient statistic, because fixing $(x+y, n-x-y)$ is equivalent to fixing $x+y$. Furthermore we could also include n_1 and n_2 into the sufficient statistic, although these values are fixed in the case of homogeneity of two binomial populations. Indeed $T = (X+Y, n-X-Y, n_1, n_2)$ is a sufficient statistic, because given the value of the vector T the conditional distribution of X is the hypergeometric distribution in (1.2) and it does not depend on $p_1 = p_2$.

Next we discuss the multinomial sampling scheme. Let x_{ij}, $i = 1,2$, $j = 1,2$, be frequencies of four *cells* of a 2×2 contingency table. The row sums and the column sums (i.e., the marginal frequencies) are denoted as x_{i+}, x_{+j}, $i, j = 1,2$. The total sample size is $n = x_{11} + x_{12} + x_{21} + x_{22}$. The data are displayed as follows.

$$
\begin{array}{|c|c|c}
\hline
x_{11} & x_{12} & x_{1+} \\
\hline
x_{21} & x_{22} & x_{2+} \\
\hline
\multicolumn{1}{c}{x_{+1}} & \multicolumn{1}{c}{x_{+2}} & n
\end{array}
\tag{1.3}
$$

At this point we mention some customary terminology of contingency tables. We look at the frequencies in (1.3) as the frequencies of a two-dimensional random variable $Y = (Y_1, Y_2)$, such that both Y_1 and Y_2 take the values 1 or 2. For example, in Table 1.1 taken from Chap. 2 of [5], Y_1 is the gender and Y_2 is the belief in afterlife. The values taken by a variable are often called *levels* of the variable. For example, in Table 1.1 two levels of the variable "gender" are "female" and "male". In this terminology x_{ij} is the joint frequency such that Y_1 takes the level i and Y_2 takes the level j. The row and the column of the contingency table are sometimes called *axes* of the table. Then Y_1 is the random variable for the first axis and Y_2 is the random variable for the second axis.

Let

$$
p_{ij} \geq 0, \quad i = 1,2, \quad j = 1,2, \quad \sum_{i,j=1}^{2} p_{ij} = 1
$$

be the probabilities of the cells. In a single multinomial trial, we observe one of the four cells according to the probabilities. With n independent and identical multinomial trials, the joint probability function of $\boldsymbol{X} = (X_{11}, X_{12}, X_{21}, X_{22})$ is given as

$$p(\boldsymbol{x}) = \binom{n}{x_{11}, x_{12}, x_{21}, x_{22}} p_{11}^{x_{11}} p_{12}^{x_{12}} p_{21}^{x_{21}} p_{22}^{x_{22}}. \tag{1.4}$$

As in this example, we use the boldface letter $\boldsymbol{x}$ for the vector of frequencies and call $\boldsymbol{x}$ the *frequency vector*. When necessary, we make the notational distinction between column vector and row vector. For example, $\boldsymbol{x}$ is meant as a column vector when we write $\boldsymbol{x} = (x_{11}, x_{12}, x_{21}, x_{22})'$. We use $'$ for denoting the transpose of a vector or a matrix in this book.

Let $p_{i+} = p_{i1} + p_{i2}$, $i = 1,2$, denote the marginal probability of the first variable of the contingency table and similarly let $p_{+j} = p_{1j} + p_{2j}$, $j = 1,2$, denote the marginal probability of the second variable. The hypothesis of independence H in the multinomial sampling scheme is specified as follows:

$$H: \ p_{ij} = p_{i+}p_{+j}, \quad i = 1,2, \quad j = 1,2. \tag{1.5}$$

On the other hand, if there is no restriction on the probability vector $\boldsymbol{p} = (p_{11}, p_{12}, p_{21}, p_{22})$, except that the elements of $\boldsymbol{p}$ are nonnegative and sum to one, we call the model *saturated*.

Write $r_i = p_{i+}$ and $c_j = p_{+j}$. Then $p_{ij} = r_i c_j$ under H. Note that in (1.5),

$$1 = \sum_{i=1}^{2} p_{i+} = \sum_{j=1}^{2} p_{+j}.$$

However, when we write $r_i = p_{i+}$ and $c_j = p_{+j}$, we can remove the restriction $1 = r_1 + r_2 = c_1 + c_2$ and only assume that r_i and c_j are nonnegative such that the total probability is 1:

$$1 = \sum_{i,j=1}^{2} r_i c_j = (r_1 + r_2)(c_1 + c_2).$$

Furthermore we can incorporate the total probability into the normalizing constant and write the probability as

$$p_{ij} = \frac{1}{(r_1 + r_2)(c_1 + c_2)} r_i c_j, \qquad i,j = 1,2, \tag{1.6}$$

where we only assume that r_i and c_j are nonnegative without any further restrictions. In this example of 2×2 tables, the normalizing constant is obvious and the above discussion may be pedantic. However, for more general models of contingency tables, it is best to consider the joint probability in the form of (1.6).

Under H, with the normalization $1 = (r_1 + r_2)(c_1 + c_2)$, the joint probability function $p(\boldsymbol{x})$ is written as

$$
\begin{aligned}
p(\boldsymbol{x}) &= \binom{n}{x_{11},x_{12},x_{21},x_{22}} (r_1c_1)^{x_{11}} (r_1c_2)^{x_{12}} (r_2c_1)^{x_{21}} (r_2c_2)^{x_{22}} \\
&= \binom{n}{x_{11},x_{12},x_{21},x_{22}} r_1^{x_{1+}} r_2^{x_{2+}} c_1^{x_{+1}} c_2^{x_{+2}} \\
&= \binom{n}{x_{11},x_{12},x_{21},x_{22}} p_{1+}^{x_{1+}} p_{2+}^{x_{2+}} p_{+1}^{x_{+1}} p_{+2}^{x_{+2}}. \qquad (1.7)
\end{aligned}
$$

Hence the sufficient statistic under H is given as

$$
T = (x_{1+}, x_{2+}, x_{+1}, x_{+2}).
$$

Given T, in the case of the 2×2 table, there is only one degree of freedom in $\boldsymbol{x}$. Namely, if x_{11} is given, then the other values x_{12}, x_{21}, x_{22} are automatically determined as

$$
x_{12} = x_{1+} - x_{11}, \qquad x_{21} = x_{+1} - x_{11}, \qquad x_{22} = n - x_{1+} - x_{+1} + x_{11}.
$$

As mentioned above, let us consider (i, j) as the pair of levels of two random variables Y_1 and Y_2. Under the null hypothesis H of independence in (1.5), Y_1 and Y_2 are independent. Suppose that we observe n independent realizations $(y_1^1, y_2^1), \dots, (y_1^n, y_2^n)$ of (Y_1, Y_2). Then x_{i+} is the number of times that Y_1 takes the value i. Hence x_{1+} is distributed according to the binomial distribution $\mathrm{Bin}(n, p_{1+})$. Similarly x_{+1} is distributed according to the binomial distribution $\mathrm{Bin}(n, p_{+1})$. Furthermore they are independent. Therefore the joint distribution of x_{1+} and x_{+1} is written as

$$
p(x_{1+}, x_{+1}) = \binom{n}{x_{1+}} p_{1+}^{x_{1+}} p_{2+}^{x_{2+}} \binom{n}{x_{+1}} p_{+1}^{x_{+1}} p_{+2}^{x_{+2}}. \qquad (1.8)
$$

From (1.7) and (1.8) it follows that the conditional distribution of X_{11} given the sufficient statistic is computed as follows.

$$
\begin{aligned}
p(x_{11} \mid x_{1+}, x_{2+}, x_{+1}, x_{+2}) &= \frac{\binom{n}{x_{11},x_{12},x_{21},x_{22}} p_{1+}^{x_{1+}} p_{2+}^{x_{2+}} p_{+1}^{x_{+1}} p_{+2}^{x_{+2}}}{\binom{n}{x_{1+}} p_{1+}^{x_{1+}} p_{2+}^{x_{2+}} \binom{n}{x_{+1}} p_{+1}^{x_{+1}} p_{+2}^{x_{+2}}} \\
&= \frac{\binom{n}{x_{11},x_{12},x_{21},x_{22}}}{\binom{n}{x_{1+}}\binom{n}{x_{+1}}} = \frac{x_{1+}!\,x_{2+}!\,x_{+1}!\,x_{+2}!}{n!\,x_{11}!\,x_{12}!\,x_{21}!\,x_{22}!}. \qquad (1.9)
\end{aligned}
$$

This is again a hypergeometric distribution. Equation (1.9) is clearly the same as (1.2) if we write the row sums and the column sums as $n_1 = x_{1+}$, $n_2 = x_{2+}$, $t = x_{+1}$, $n - t = x_{+2}$. Therefore Fisher's exact test is the same in this multinomial sampling scheme as in the case of testing the homogeneity of two binomial populations.

Note that in this scheme n is fixed and $x_{2+} = n - x_{1+}$ and $x_{+2} = n - x_{+1}$ can be omitted from the sufficient statistic $T = (x_{1+}, x_{2+}, x_{+1}, x_{+2})$. However, as in the first scheme we can allow the redundancy in the sufficient statistic.

Finally we consider the sampling scheme of Poisson random variables. Let X_{ij}, $i,j = 1,2$, be independently distributed according to the Poisson distribution with mean λ_{ij}. The joint probability of $\boldsymbol{X}$ is written as

$$p(\boldsymbol{x}) = \prod_{i,j=1}^{2} \frac{\lambda_{ij}^{x_{ij}}}{x_{ij}!} e^{-\lambda_{ij}}.$$

Consider the null hypothesis H that λ_{ij} can be factored as

$$H : \lambda_{ij} = r_i c_j, \qquad i,j = 1,2,$$

where r_i, c_j are nonnegative. Again by writing down the joint probability under the null hypothesis H, we can easily check that a sufficient statistic under H is given by $T = (x_{1+}, x_{2+}, x_{+1}, x_{+2})$, where now the redundancy is only in $x_{+2} = x_{1+} + x_{2+} - x_{+1}$. Instead of writing out the joint probability, we use the following property of independent Poisson random variables for verifying that T is a sufficient statistic under H. Let $n = X_{11} + X_{12} + X_{21} + X_{22}$. Then n is distributed as the Poisson random variable with mean $\mu = \sum_{i,j=1}^{2} \lambda_{ij}$. Under H, $\mu = (r_1 + r_2)(c_1 + c_2)$. Given n, the conditional distribution of $(X_{11}, X_{12}, X_{21}, X_{22})$ is the multinomial distribution with cell probabilities $p_{ij} = \lambda_{ij}/\mu$. Under H, the cell probability is written as

$$p_{ij} = \frac{1}{(r_1 + r_2)(c_1 + c_2)} r_i c_j, \qquad i,j = 1,2,$$

which is the same as (1.6). From this fact we see that $T = (x_{1+}, x_{2+}, x_{+1}, x_{+2})$ is a sufficient statistic under H. Given T, the conditional distribution of $\boldsymbol{x}$ is the same as the multinomial case; that is, X_{11} follows the hypergeometric distribution in (1.9).

We now note the relation between the cell frequencies and the sufficient statistic. The column vector of cell frequencies $\boldsymbol{x} = (x_{11}, x_{12}, x_{21}, x_{22})'$ and the column vector of the sufficient statistic $(x_{1+}, x_{2+}, x_{+1}, x_{+2})'$ are related as follows:

$$\begin{pmatrix} x_{1+} \\ x_{2+} \\ x_{+1} \\ x_{+2} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{12} \\ x_{21} \\ x_{22} \end{pmatrix}. \tag{1.10}$$

We write this as $\boldsymbol{t} = A\boldsymbol{x}$ and call the matrix A the *configuration* for the above three models.

1.2 2×2 Contingency Table Models as Discrete Exponential Family

In the previous section we explained three sampling schemes for 2×2 contingency tables and pointed out that they share the same sufficient statistic when redundancies are allowed. In this section we present the standard formulation of the sampling schemes as discrete exponential family models. We confirm that the sufficient statistics under the null hypothesis correspond to nuisance parameters. Hence fixing the sufficient statistic has the effect of eliminating the nuisance parameters and the resulting conditional test is a similar test. Here we only consider the multinomial scheme of the previous section, because the other cases can be treated in a similar manner.

A family of joint probability functions $p(\boldsymbol{x}) = p(\boldsymbol{x};\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, is said to form an *exponential family* (see Sect. 2.7 of [98]) if $p(\boldsymbol{x},\boldsymbol{\theta})$ is written in the following form.

$$p(\boldsymbol{x};\boldsymbol{\theta}) = h(\boldsymbol{x})\exp\left(\sum_{j=1}^{k} T_j(\boldsymbol{x})\phi_j(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta})\right). \tag{1.11}$$

By the factorization theorem (Sect. 2.6 of [98]), $T = (T_1(\boldsymbol{x}),\ldots,T_k(\boldsymbol{x}))$ is a sufficient statistic of this family. Note that $p(\boldsymbol{x};\boldsymbol{\theta})$ and $\psi(\boldsymbol{\theta})$ depend on $\boldsymbol{\theta}$ only through $\boldsymbol{\phi} = (\phi_1,\ldots,\phi_k)$ and we can write $\psi(\boldsymbol{\phi})$ instead of $\psi(\boldsymbol{\theta})$. In Chap. 4 we simply denote $\phi_j(\boldsymbol{\theta})$ itself as θ_j.

Let p_{ij}, $i,j = 1,2$, denote the cell probabilities in the multinomial sampling of a 2×2 contingency table. Now consider the following transformation:

$$\phi_1 = \log\frac{p_{12}}{p_{22}}, \qquad \phi_2 = \log\frac{p_{21}}{p_{22}}, \qquad \lambda = \log\frac{p_{11}p_{22}}{p_{12}p_{21}}. \tag{1.12}$$

In the region where the elements of the probability vector $\boldsymbol{p} = (p_{11},p_{12},p_{21},p_{22})$ are positive, the transformation is one-to-one and the inverse transformation is written as

$$\begin{aligned}
p_{11} &= \frac{e^{\phi_1+\phi_2+\lambda}}{1+e^{\phi_1}+e^{\phi_2}+e^{\phi_1+\phi_2+\lambda}},\\
p_{12} &= \frac{e^{\phi_1}}{1+e^{\phi_1}+e^{\phi_2}+e^{\phi_1+\phi_2+\lambda}},\\
p_{21} &= \frac{e^{\phi_2}}{1+e^{\phi_1}+e^{\phi_2}+e^{\phi_1+\phi_2+\lambda}},\\
p_{22} &= \frac{1}{1+e^{\phi_1}+e^{\phi_2}+e^{\phi_1+\phi_2+\lambda}}.
\end{aligned} \tag{1.13}$$

Substituting this into (1.4) we can write the joint probability function of $\boldsymbol{x}$ as

$$\begin{aligned}
p(\boldsymbol{x}) = \binom{n}{x_{11},x_{12},x_{21},x_{22}}\exp\big((x_{11}+x_{12})\phi_1 + (x_{11}+x_{21})\phi_2 + x_{11}\lambda \\
-n\log(1+e^{\phi_1}+e^{\phi_2}+e^{\phi_1+\phi_2+\lambda})\big).
\end{aligned} \tag{1.14}$$

This is written in the form (1.11) and hence the family of $p(\boldsymbol{x})$ forms an exponential family. By putting $r_1 = e^{\phi_1}, r_2 = 1, c_1 = e^{\phi_2}, c_2 = 1$ we see that the null hypothesis of the independence (1.5) is equivalently written as

$$H : \lambda = 0.$$

Note that λ is the parameter of interest for the null hypothesis and ϕ_1, ϕ_2 are the nuisance parameters under the null hypothesis. Under the null hypothesis, $\lambda = 0$ is no longer a parameter of the family of distributions and the distributions under the null hypothesis are parametrized by the nuisance parameters ϕ_1, ϕ_2. In (1.14) the sufficient statistic corresponding to (ϕ_1, ϕ_2) is

$$x_{1+} = x_{11} + x_{12}, \quad x_{+1} = x_{11} + x_{21}.$$

In (1.11) and (1.14) we considered the joint probability of the frequency vector. In fact, when we consider a single observation $n = 1$, then the cell probabilities are already in the exponential family form. Write

$$\log \boldsymbol{p} = (\log p_{11}, \log p_{12}, \log p_{21}, \log p_{22}),$$
$$\psi(\phi_1, \phi_2) = \log(1 + e^{\phi_1} + e^{\phi_2} + e^{\phi_1 + \phi_2}).$$

Taking the logarithms of p_{ij} in (1.13) with $\lambda = 0$, in a matrix form we can write

$$\log \boldsymbol{p} = (\phi_1, 0, \phi_2, 0) \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} - \psi(\phi_1, \phi_2) \times (1, 1, 1, 1). \tag{1.15}$$

Note that the matrix on the right-hand side is the configuration A appearing in the right-hand side of (1.10).

1.3 Independence Model of General Two-Way Contingency Tables

Generalizing the discussion of the previous section we now consider the independence model of general $I \times J$ two-way contingency tables. The discussion on three sampling schemes is entirely the same as in the case of 2×2 tables. Therefore we only discuss the multinomial sampling.

Let p_{ij}, $i = 1, \dots, I$, $j = 1, \dots, J$, denote the cell probabilities of an $I \times J$ contingency table. Let p_{i+} and p_{+j} denote the marginal probabilities. The null hypothesis of independence is written as

$$H : p_{ij} = p_{i+} p_{+j}, \quad i = 1, \dots, I, \qquad j = 1, \dots, J.$$

We can also write $p_{ij} = r_i c_j$ without requiring that r_is and c_js correspond to probabilities. Let x_{ij} denote the frequency of the cell (i, j). A sufficient statistic

T under the null hypothesis H is the set of the row sums x_{i+}, $i = 1,\dots,I$ and the column sums x_{+j}, $j = 1,\dots,J$. Let n denote the total sample size.

Under the null hypothesis the joint probability of $\boldsymbol{x} = \{x_{ij}\}$ is written as

$$\begin{aligned} p(\boldsymbol{x}) &= \binom{n}{x_{11},\dots,x_{IJ}} \prod_{i=1}^{I}\prod_{j=1}^{J}(p_{i+}p_{+j})^{x_{ij}} \\ &= \binom{n}{x_{11},\dots,x_{IJ}} \prod_{i=1}^{I} p_{i+}^{x_{i+}} \prod_{j=1}^{J} p_{+j}^{x_{+j}}. \end{aligned}$$

Also, under the null hypothesis, as in the case of 2×2 tables, the vector of row sums $\{x_{i+}\}$ and the vector of column sums $\{x_{+j}\}$ are independently distributed according to multinomial distributions:

$$\begin{aligned} p(\{x_{i+}\}) &= \binom{n}{x_{1+},\dots,x_{I+}} p_{1+}^{x_{1+}}\cdots p_{I+}^{x_{I+}}, \\ p(\{x_{+j}\}) &= \binom{n}{x_{+1},\dots,x_{+J}} p_{+1}^{x_{+1}}\cdots p_{+J}^{x_{+J}}. \end{aligned}$$

From this fact, the conditional distribution of $\boldsymbol{x} = \{x_{ij}\}$ given the sufficient statistic $\boldsymbol{t}$ is written as

$$\begin{aligned} p(\boldsymbol{x} \mid T = \boldsymbol{t}) &= \frac{p(\{x_{ij}\})}{p(\{x_{i+}\})p(\{x_{+j}\})} = \frac{\binom{n}{x_{11},\dots,x_{IJ}}}{\binom{n}{x_{1+},\dots,x_{I+}}\binom{n}{x_{+1},\dots,x_{+J}}} \\ &= \frac{\prod_{i=1}^{I} x_{i+}!\,\prod_{j=1}^{J} x_{+j}!}{n!\prod_{i,j} x_{ij}!}. \end{aligned} \tag{1.16}$$

This distribution is often called the *multivariate hypergeometric distribution*. However in this book we show many variations of distributions of this type and we often refer to them simply as hypergeometric distributions.

Given the row sums and the column sums, the degrees of freedom in the frequency vector $\boldsymbol{x}$ is $(I-1)\times(J-1)$ because the elements of the last row and the last column are determined uniquely from the other elements. This degrees of freedom is also the dimension of the parameter of interest when the joint probability distribution is written in the exponential family form. More precisely let

$$\begin{aligned} \phi_{1i} &= \log\frac{p_{iJ}}{p_{IJ}}, && i = 1,\dots,I-1, \\ \phi_{2j} &= \log\frac{p_{Ij}}{p_{IJ}}, && j = 1,\dots,J-1, \\ \lambda_{ij} &= \log\frac{p_{ij}p_{IJ}}{p_{iJ}p_{Ij}}, && i = 1,\dots,I-1,\ j = 1,\dots,J-1. \end{aligned} \tag{1.17}$$

Then the null hypothesis is written as

$$H : \lambda_{ij} = 0, \qquad i = 1, \dots, I-1, \qquad j = 1, \dots, J-1.$$

One consequence of the multidimensionality of the parameter of interest is that there is no unique best choice for a test statistic, even under the requirement of similarity and unbiasedness.

Let

$$\hat{m}_{ij} = n\hat{p}_{ij} = \frac{x_{i+}x_{+j}}{n}$$

denote the "expected frequency" of the cell (i, j), where $\hat{p}_{ij}$ is the maximum likelihood estimate (MLE) of p_{ij}. For testing the null hypothesis of independence, popular test statistics are *Pearson's chi-square test*

$$\chi^2(\boldsymbol{x}) = \sum_i \sum_j \frac{(x_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \geq c_\alpha \Rightarrow \text{reject } H$$

and the (twice log) *likelihood ratio test*

$$G^2(\boldsymbol{x}) = 2\sum_i \sum_j x_{ij} \log \frac{x_{ij}}{\hat{m}_{ij}} \geq c_\alpha \Rightarrow \text{reject } H,$$

where c_α is the critical value for the respective test statistic. $G^2(\boldsymbol{x})$ is actually twice the logarithm of the likelihood ratio. In the usual asymptotic theory, c_α is approximated by the upper α-quantile of the chi-square distribution with $(I-1)(J-1)$ degrees of freedom. In this book we denote the chi-square distribution with m degrees of freedom by χ^2_m.

These two statistics are "omnibus test statistics" in the sense that all possible alternative hypotheses are roughly equally treated. When some specific deviations from the null hypothesis are expected, then a more suitable test statistic, which is sensitive against the deviation, can be used. For performing a test of H, once a test statistic is chosen, it only remains to evaluate its null distribution. As in the previous section, in this book we consider exact tests; that is, we are interested in the distribution of a test statistic under the hypergeometric distribution (1.16).

At this point we investigate the conditional sample space; that is, the set of contingency tables given the sufficient statistic for $I \times J$ case. As in the 2×2 case, the relation between the sufficient statistic and the frequency vector is written in a matrix form. Let $\boldsymbol{t} = (x_{1+}, \dots, x_{I+}, x_{+1}, \dots, x_{+J})'$ denote the (column) vector of the sufficient statistic and let $\boldsymbol{x} = (x_{11}, x_{12}, \dots, x_{1J}, x_{21}, \dots, x_{IJ})'$ denote the frequency vector. Then

$$\boldsymbol{t} = A\boldsymbol{x}, \tag{1.18}$$

where the configuration A is an $(I+J) \times IJ$ matrix consisting of 0s and 1s as in (1.10).

An explicit form of A can be given using the Kronecker product notation. For two matrices, $C = \{c_{ij}\} : m_1 \times n_1$ and $D : m_2 \times n_2$, their Kronecker product $C \otimes D$ is an $m_1 m_2 \times n_1 n_2$ matrix of the following block form

$$C \otimes D = \begin{pmatrix} c_{11}D & \dots & c_{1n_1}D \\ \vdots & & \vdots \\ c_{m_1 1}D & \dots & c_{m_1 n_1}D \end{pmatrix}. \tag{1.19}$$

Let $\mathbf{1}_n = (1,\dots,1)'$ denote the n-dimensional vector consisting of 1s and let E_m denote an $m \times m$ identity matrix. Then A in (1.18) is written as

$$A = \begin{pmatrix} E_I \otimes \mathbf{1}'_J \\ \mathbf{1}'_I \otimes E_J \end{pmatrix}.$$

Alternatively let $\boldsymbol{e}_{j,n} = (0,\dots,0,1,0,\dots,0)' \in \mathbb{R}^n$ denote the jth standard basis vector of $\mathbb{R}^n$. When the dimension n is clear from the context, we simply write the standard basis vector as $\boldsymbol{e}_j$ instead of $\boldsymbol{e}_{j,n}$. Then the columns of A are of the form

$$\begin{pmatrix} \boldsymbol{e}_{i,I} \\ \boldsymbol{e}_{j,J} \end{pmatrix}, \qquad i = 1,\dots,I, \qquad j = 1,\dots,J. \tag{1.20}$$

We sometimes denote the stacked vector in (1.20) as

$$\boldsymbol{e}_{i,I} \oplus \boldsymbol{e}_{j,J} = \begin{pmatrix} \boldsymbol{e}_{i,I} \\ \boldsymbol{e}_{j,J} \end{pmatrix}. \tag{1.21}$$

It is easily checked that the rank of A is

$$\operatorname{rank} A = I + J - 1.$$

Hence the dimension of the kernel of A is given as

$$\dim \ker A = IJ - (I + J - 1) = (I-1)(J-1).$$

As mentioned above, this dimension corresponds to the fact that, if we ignore the requirement of nonnegativity, we can choose the elements of the first $I-1$ rows and the first $J-1$ columns freely. With the additional requirement of nonnegativity, the *conditional sample space* given the sufficient statistic is defined as

$$\mathscr{F}_{\boldsymbol{t}} = \{\boldsymbol{x} \in \mathbb{Z}^{IJ} \mid \boldsymbol{x} \geq \mathbf{0}, \boldsymbol{t} = A\boldsymbol{x}\}, \tag{1.22}$$

where $\boldsymbol{x} \geq \mathbf{0}$ means that the elements of $\boldsymbol{x}$ are nonnegative. We call $\mathscr{F}_{\boldsymbol{t}}$ the *fiber* of $\boldsymbol{t}$ (or also call it the t-fiber). The hypergeometric distribution in (1.16) is a probability distribution over the fiber $\mathscr{F}_{\boldsymbol{t}}$. When a test statistic $\phi(\boldsymbol{x})$ is given, we want to evaluate the distribution of $\phi(\boldsymbol{x})$, where $\boldsymbol{x}$ is distributed according to the hypergeometric distribution over $\mathscr{F}_{\boldsymbol{t}}$.

Suppose that ϕ is chosen such that a larger value of ϕ indicates more deviation from the null hypothesis, as in Pearson's chi-square statistic or the likelihood ratio statistic. Then testing can be conveniently performed via *p-value*. Let $\boldsymbol{x}^o$ denote the observed contingency table. The p-value of $\boldsymbol{x}^o$ is defined as

$$p = P(\phi(\boldsymbol{x}) \geq \phi(\boldsymbol{x}^o) \mid H) = \sum_{\boldsymbol{x} \in \mathscr{F}_{\boldsymbol{t}}, \phi(\boldsymbol{x}) \geq \phi(\boldsymbol{x}^o)} p(\boldsymbol{x} \mid \boldsymbol{t} = A\boldsymbol{x}^o, H), \tag{1.23}$$

which is the probability under the hypergeometric distribution of observing the value $\phi(\boldsymbol{x})$ which is larger than or equal to $\phi(\boldsymbol{x}^o)$. Given the level of significance α, we reject H if $p \leq \alpha$.

There are three methods to evaluate the p-value in (1.23).

1. By enumerating $\mathscr{F}_{\boldsymbol{t}}$, $\boldsymbol{t} = A\boldsymbol{x}^o$, and performing the sum in (1.23) for all $\boldsymbol{x} \in \mathscr{F}_{\boldsymbol{t}}$ such that $\phi(\boldsymbol{x}) \geq \phi(\boldsymbol{x}^o)$.
2. Directly sampling $\boldsymbol{x}$ from the hypergeometric distribution and approximating (1.23) by Monte Carlo simulation.
3. By sampling $\boldsymbol{x}$ by a Markov chain whose stationary distribution is the hypergeometric distribution, that is, by a Markov chain Monte Carlo method.

Clearly the enumeration is the best if it is feasible. However, when the row sums and the column sums become large, the size of the fiber $\mathscr{F}_{\boldsymbol{t}}$ becomes large and the enumeration becomes infeasible. In the case of the independence model of this section, direct sampling of a frequency vector from the hypergeometric distribution is easy to carry out. In more complicated models treated later in the book, though, direct sampling is not easy. On the other hand, there exists a general theory of constructing a Markov chain having the hypergeometric distribution as the stationary distribution. Hence the subject of this book is the Markov chain sampling from the fiber $\mathscr{F}_{\boldsymbol{t}}$.

In the next chapter, again employing the independence model of $I \times J$ contingency tables, we discuss how to perform Markov chain sampling from the fiber $\mathscr{F}_{\boldsymbol{t}}$.

1.4 Conditional Independence Model of Three-Way Contingency Tables

In this section we discuss the conditional independence model for three-way contingency tables. It is a relatively simple model in the sense that for each level of the conditioning variable, the problem reduces to the case of an independence model of two-way contingency tables for the other variables. However, it is a convenient model for introducing a notation for general m-way contingency tables in the next section.

Consider an $I_1 \times I_2 \times I_3$ three-way contingency table $\boldsymbol{x}$. We denote each cell of the table by a multi-index $\boldsymbol{i} = (i_1, i_2, i_3)$. For a positive integer J write

$$[J] = \{1, \dots, J\}.$$

The set of the cells is the following direct product

$$\mathscr{I} = \{\boldsymbol{i} = (i_1, i_2, i_3) \mid i_1 \in [I_1], i_2 \in [I_2], i_3 \in [I_3]\} = [I_1] \times [I_2] \times [I_3].$$

With this notation the three-way contingency table, or the frequency vector, is denoted as

$$\boldsymbol{x} = \{x(\boldsymbol{i}) \mid \boldsymbol{i} \in \mathscr{I}\}.$$

Note that this notation is somewhat heavy and in fact for three-way tables we prefer to use subscripts i, j, k. The merit of this notation is that it can be used for general m-way tables.

For a subset $D \subset \{1, 2, 3\}$ of the variables, let $\boldsymbol{i}_D$ denote the set of indices in D. For example,

$$\boldsymbol{i}_{\{1,2\}} = (i_1, i_2).$$

Note that $\boldsymbol{i}_D$ corresponds to the D-marginal cell of the contingency table. The set of D-marginal cells is denoted by

$$\mathscr{I}_D = \prod_{k \in D} [I_k]. \tag{1.24}$$

For example $\mathscr{I}_{\{1,2\}} = \{(i_1, i_2) \mid i_1 \in [I_1], i_2 \in [I_2]\}$. The D-marginal frequencies of $\boldsymbol{x}$ are written as

$$x_D(\boldsymbol{i}_D) = \sum_{\boldsymbol{i}_{D^C} \in \mathscr{I}_{D^C}} x(\boldsymbol{i}_D, \boldsymbol{i}_{D^C}), \tag{1.25}$$

where D^C denotes the complement of D. Note that in $x(\boldsymbol{i}_D, \boldsymbol{i}_{D^C})$, for notational simplicity, the indices in $\mathscr{I}_D$ are collected to the left. Also we are writing $x(\boldsymbol{i}_D, \boldsymbol{i}_{D^C})$ instead of $x((\boldsymbol{i}_D, \boldsymbol{i}_{D^C}))$. In the two-way case

$$x_{i+} = x_{\{1\}}(i) = \sum_j x_{ij}.$$

For a probability distribution $\{p(\boldsymbol{i}), \boldsymbol{i} \in \mathscr{I}\}$, we denote the D-marginal probability as $p_D(\boldsymbol{i}_D)$. Note that in $x_D(\boldsymbol{i}_D)$ and $p_D(\boldsymbol{i}_D)$, the subset D is indicated twice. If there is no notational confusion we alternatively write

$$x(\boldsymbol{i}_D), x_D(\boldsymbol{i}), p(\boldsymbol{i}_D) \quad \text{or} \quad p_D(\boldsymbol{i}) \tag{1.26}$$

for simplicity.

We call a D-marginal probability distribution *saturated* if there is no restriction on the probability vector $\{p_D(\boldsymbol{i}_D), \boldsymbol{i}_D \in \mathscr{I}_D\}$.

Let Y_1, Y_2, Y_3 be random variables corresponding to the three axes of the contingency table. We consider the model that Y_1 and Y_3 are conditionally independent given the level i_2 of Y_2. The relevant conditional probabilities are written as

$$p(i_1, i_3 \mid i_2) = \frac{p(\boldsymbol{i})}{p_{\{2\}}(i_2)}, \quad p(i_1 \mid i_2) = \frac{p_{\{1,2\}}(i_1, i_2)}{p_{\{2\}}(i_2)}, \quad p(i_3 \mid i_2) = \frac{p_{\{2,3\}}(i_2, i_3)}{p_{\{2\}}(i_2)}.$$

In the following we omit subscripts to p and write, for example, $p(i_1, i_2)$ instead of $p_{\{1,2\}}(i_1, i_2)$. Similarly we write $x(i_1, i_2)$ instead of $x_{\{1,2\}}(i_1, i_2)$. The null hypothesis of conditional independence is written as

$$H : \frac{p(\boldsymbol{i})}{p(i_2)} = \frac{p(i_1, i_2)}{p(i_2)} \times \frac{p(i_2, i_3)}{p(i_2)}, \qquad \forall \boldsymbol{i} \in \mathscr{I}, \tag{1.27}$$

or equivalently as

$$H : p(\boldsymbol{i}) = \frac{1}{p(i_2)} p(i_1, i_2) p(i_2, i_3), \qquad \forall \boldsymbol{i} \in \mathscr{I}. \tag{1.28}$$

Here we are assuming $p(i_2) > 0$. In the case $p(i_2) = 0$ for a particular level i_2, we have $p(\boldsymbol{i}) = p(i_1, i_2) = p(i_2, i_3) = 0$ for indices containing this level i_2 of Y_2. Hence in this case we understand (1.28) as $0 = 0 \times 0/0$. Let

$$\alpha(i_1, i_2) = \frac{p(i_1, i_2)}{p(i_2)}, \quad \beta(i_2, i_3) = p(i_2, i_3).$$

Then the conditional independence model is written as

$$H : p(\boldsymbol{i}) = \alpha(i_1, i_2)\beta(i_2, i_3). \tag{1.29}$$

Note that there is some indeterminacy in specifying α and β. For example we can include the factor $1/p(i_2)$ into $\beta(i_2, i_3)$ instead of into $\alpha(i_1, i_2)$.

We can show that (1.27), (1.28), and (1.29) are in fact equivalent. Suppose that $p(\boldsymbol{i}) = p(i_1, i_2, i_3)$ can be written as $p(\boldsymbol{i}) = \alpha(i_1, i_2)\beta(i_2, i_3)$. Then

$$\begin{aligned}
p(i_2) &= \sum_{i_1, i_3} p(i_1, i_2, i_3) = \sum_{i_1, i_3} \alpha(i_1, i_2)\beta(i_2, i_3) = \left(\sum_{i_1} \alpha(i_1, i_2)\right)\left(\sum_{i_3} \beta(i_2, i_3)\right), \\
p(i_1, i_2) &= \sum_{i_3} p(i_1, i_2, i_3) = \alpha(i_1, i_2) \sum_{i_3} \beta(i_2, i_3), \\
p(i_2, i_3) &= \sum_{i_1} p(i_1, i_2, i_3) = \left(\sum_{i_1} \alpha(i_1, i_2)\right) \beta(i_2, i_3).
\end{aligned}$$

Therefore

$$\begin{aligned}\frac{p(i_1,i_2)p(i_2,i_3)}{p(i_2)} &= \frac{\alpha(i_1,i_2)\beta(i_2,i_3)(\sum_{i'_1}\alpha(i'_1,i_2))(\sum_{i'_3}\beta(i_2,i'_3))}{(\sum_{i'_1}\alpha(i'_1,i_2))(\sum_{i'_3}\beta(i_2,i'_3))}\\ &= \alpha(i_1,i_2)\beta(i_2,i_3)\\ &= p(\boldsymbol{i})\end{aligned}$$

and hence (1.28) holds. This shows that the null hypothesis of conditional independence can be written in any one of (1.27), (1.28), and (1.29).

Now suppose that we observe a contingency table $\boldsymbol{x}$ of sample size n from the conditional independence model. The joint probability function is written as

$$\begin{aligned}p(\boldsymbol{x}) &= \frac{n!}{\prod_{\boldsymbol{i}\in\mathscr{I}} x(\boldsymbol{i})!}\prod_{\boldsymbol{i}\in\mathscr{I}}(\alpha(i_1,i_2)\beta(i_2,i_3))^{x(\boldsymbol{i})}\\ &= \frac{n!}{\prod_{\boldsymbol{i}\in\mathscr{I}} x(\boldsymbol{i})!}\prod_{\boldsymbol{i}_{\{1,2\}}\in\mathscr{I}_{\{1,2\}}}\alpha(i_1,i_2)^{x(i_1,i_2)}\prod_{\boldsymbol{i}_{\{2,3\}}\in\mathscr{I}_{\{2,3\}}}\beta(i_2,i_3)^{x(i_2,i_3)}.\end{aligned} \tag{1.30}$$

Hence a sufficient statistic T is the set of $\{1,2\}$-marginals and $\{2,3\}$-marginals of $\boldsymbol{x}$:

$$T = (\{x(\boldsymbol{i}_{\{1,2\}}) \mid \boldsymbol{i}_{\{1,2\}} \in \mathscr{I}_{\{1,2\}}\},\ \{x(\boldsymbol{i}_{\{2,3\}}) \mid \boldsymbol{i}_{\{2,3\}} \in \mathscr{I}_{\{2,3\}}\}).$$

In this case the marginal distribution of T is not immediately clear and hence the conditional probability of $\boldsymbol{x}$ given $T = \boldsymbol{t}$ is also not immediately clear. However, without worrying about the marginal distribution of T at this point, we can proceed as follows. Let A be the configuration relating the frequency vector to the sufficient statistic: $\boldsymbol{t} = A\boldsymbol{x}$. Define $\mathscr{F}_{\boldsymbol{t}} = \{\boldsymbol{x} \geq 0 \mid \boldsymbol{t} = A\boldsymbol{x}\}$ as in (1.22). The terms containing the parameters α, β on the right-hand side of (1.30) are fixed by the sufficient statistic, therefore these terms do not appear in the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{t}$. It follows that the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{t}$ is written as

$$p(\boldsymbol{x} \mid \boldsymbol{t}) = c \times \frac{1}{\prod_{\boldsymbol{i}\in\mathscr{I}} x(\boldsymbol{i})!}, \qquad c = \left[\sum_{\boldsymbol{x}\in\mathscr{F}_{\boldsymbol{t}}}\frac{1}{\prod_{\boldsymbol{i}\in\mathscr{I}} x(\boldsymbol{i})!}\right]^{-1}. \tag{1.31}$$

As in the previous examples, an exact test of the null hypothesis H of conditional independence can be performed if either we can enumerate the elements of $\mathscr{F}_{\boldsymbol{t}}$ or if we can sample from this distribution. Note that we often call (1.31) the hypergeometric distribution over $\mathscr{F}_{\boldsymbol{t}}$.

In general, the normalizing constant c cannot be written explicitly. The Markov chain sampling discussed in the next chapter can be performed without knowing the explicit form of the normalizing constant. This is one of the major advantages of Markov chain Monte Carlo methods.

It turns out that for the conditional independence model the marginal distribution of the sufficient statistic T and the normalizing constant c can be written down explicitly. This is a special case of the result of Sundberg [140] for decomposable models, which is studied in Chap. 8. In the following section, we explain the marginal distribution of T. The following section can be skipped, because the normalizing constant c is not needed for performing Markov chain Monte Carlo methods.

1.4.1 Normalizing Constant of Hypergeometric Distribution for the Conditional Independence Model

For illustration let us explicitly write out the configuration for relating the frequency vector to the sufficient statistic for the case of $2 \times 2 \times 2$ tables. We order the elements of T according to the level of Y_2. Then $\boldsymbol{t} = A\boldsymbol{x}$ is written as

$$\begin{pmatrix} x_{\{1,2\}}(1,1) \\ x_{\{1,2\}}(2,1) \\ x_{\{2,3\}}(1,1) \\ x_{\{2,3\}}(1,2) \\ x_{\{1,2\}}(1,2) \\ x_{\{1,2\}}(2,2) \\ x_{\{2,3\}}(2,1) \\ x_{\{2,3\}}(2,2) \end{pmatrix} = \begin{pmatrix} 1\ 1\ 0\ 0 & \\ 0\ 0\ 1\ 1 & \mathbf{0} \\ 1\ 0\ 1\ 0 & \\ 0\ 1\ 0\ 1 & \\ & 1\ 1\ 0\ 0 \\ \mathbf{0} & 0\ 0\ 1\ 1 \\ & 1\ 0\ 1\ 0 \\ & 0\ 1\ 0\ 1 \end{pmatrix} \begin{pmatrix} x(1,1,1) \\ x(1,1,2) \\ x(2,1,1) \\ x(2,1,2) \\ x(1,2,1) \\ x(1,2,2) \\ x(2,2,1) \\ x(2,2,2) \end{pmatrix}, \tag{1.32}$$

where the big 0 is the 4×4 zero matrix. Note that the 8×8 matrix on the right-hand side is a block diagonal with identical blocks. Furthermore, the diagonal block is the same as on the right-hand side of (1.10). Partition $\boldsymbol{x}$ on the right-hand side of (1.32) into two 4-dimensional subvectors $\boldsymbol{x}_1, \boldsymbol{x}_2$. We call each $\boldsymbol{x}_{i_2}$, $i_2 = 1, 2$, the *slice* of the contingency table $\boldsymbol{x}$ by fixing the level i_2 of the second variable. Similarly we partition $\boldsymbol{t}$ on the left-hand side of (1.32) into two 4-dimensional subvectors $\boldsymbol{t}_1, \boldsymbol{t}_2$. Then clearly

$$\boldsymbol{x} \in \mathscr{F}_{\boldsymbol{t}} \quad \Leftrightarrow \quad \boldsymbol{x}_1 \in \mathscr{F}_{\boldsymbol{t}_1} \quad \text{and} \quad \boldsymbol{x}_2 \in \mathscr{F}_{\boldsymbol{t}_2}, \tag{1.33}$$

where $\mathscr{F}_{\boldsymbol{t}_1}$ and $\mathscr{F}_{\boldsymbol{t}_2}$ are fibers in (1.22) for the independence model of 2×2 contingency tables.

We have thus far looked at the $2 \times 2 \times 2$ case. However, it is clear that a similar result holds for the general $I_1 \times I_2 \times I_3$ case. Namely, when we sort the cells according to the levels of Y_2, then the configuration is in a block diagonal form with identical blocks, which correspond to the configuration of the independence model for $I_1 \times I_3$ contingency tables.