3

Algebraic geometry of 2×2 contingency tables

Aleksandra B. Slavković

Stephen E. Fienberg

Abstract

Contingency tables represent the joint distribution of categorical variables. In this chapter we use modern algebraic geometry to update the geometric representation of 2×2 contingency tables first explored in (Fienberg 1968) and (Fienberg and Gilbert 1970). Then we use this geometry for a series of new ends including various characterizations of the joint distribution in terms of combinations of margins, conditionals, and odds ratios. We also consider incomplete characterisations of the joint distribution and the link to latent class models and to the phenomenon known as Simpson's paradox. Many of the ideas explored here generalise rather naturally to $I \times J$ and higher-way tables. We end with a brief discussion of generalisations and open problems.

3.1 Introduction

(Pearson 1956) in his presidential address to the Royal Statistical Society was one of the earliest statistical authors to write explicitly about the role of geometric thinking for the theory of statistics, although many authors previously, such as (Edgeworth 1914) and (Fisher 1921), had relied heuristically upon geometric characterisations.

For contingency tables, beginning with (Fienberg 1968) and (Fienberg and Gilbert 1970), several authors have exploited the geometric representation of contingency table models, in terms of quantities such as margins and odds ratios, both for the proof of statistical results and to gain deeper understanding of models used for contingency table representation. For example, see (Fienberg 1970) for the convergence of iterative proportional fitting procedure, (Diaconis 1977) for the geometric representation of exchangeability, and (Kenett 1983) for uses in exploratory data analysis. More recently, (Nelsen 1995, Nelsen 2006) in a discussion of copulas for binary variables points out that two faces of the tetrahedron form the Fréchet upper bound, the other two the lower bound, and the surface of independence is the independence copula.

There has also been considerable recent interest in geometric descriptions of contingency tables models and analytical tools, from highly varying perspectives.

Algebraic and Geometric Methods in Statistics, ed. Paolo Gibilisco, Eva Riccomagno, Maria Piera Rogantin and Henry P. Wynn. Published by Cambridge University Press.

(Erosheva 2005) employed a geometric approach to compare the potential value of using the Grade of Membership, latent class, and Rasch models in representing population heterogeneity for 2^J tables. Similarly, (Heiser 2004, De Rooij and Anderson 2007, De Rooij and Heiser 2005) have given geometric characterisations linked to odds ratios and related models for $I \times J$ tables, (Greenacre and Hastie 1987) focus on the geometric interpretation of correspondence analysis for contingency tables, (Carlini and Rapallo 2005) described some of the links to (Fienberg and Gilbert 1970) as well as the geometric structure of statistical models for case-control studies, and (Flach 2003) linked the geometry to Receiver Operating Characteristic space.

In this chapter we return to the original geometric representation of (Fienberg and Gilbert 1970) and link the geometry to some modern notions from algebraic geometry, e.g., as introduced to statistical audiences in (Diaconis and Sturmfels 1998) and (Pistone *et al.* 2001), to provide a variety of characterisations of the joint distribution of two binary variables, some old and some new. There are numerous ways we can characterise bivariate distributions, e.g., see (Arnold *et al.* 1999, Ramachandran and Lau 1991, Kagan *et al.* 1973). In related work, (Slavkovic and Sullivant 2006) give an algebraic characterisation of compatibility of full conditionals for discrete random variables. In this chapter, however, we are interested in the 'feasibility' question; that is, when do compatible conditionals and/or marginals correspond to an actual table. *Under the assumption that given sets of marginal and conditional binary distributions are compatible, we want to check whether or not they are sufficient to uniquely identify the existing joint distribution.* We are under the assumptions of the uniqueness theorem of (Gelman and Speed 1993) as redefined by (Arnold *et al.* 1999). More specifically, we allow cell entries to be zero as long as we do not condition on an event of zero probability. We draw on a more technical discussion in (Slavkovic 2004), and we note the related discussion in (Luo *et al.* 2004) and in (Carlini and Rapallo 2005).

3.2 Definitions and notation

Contingency tables are arrays of non-negative integers that arise from the cross-classification of a sample or a population of N objects based on a set of categorical variables of interest, see (Bishop *et al.* 1975) and (Lauritzen 1996). We represent the contingency table $\mathbf{n}$ as a vector of non-negative integers, each indicating the number of times a given configuration of classifying criteria has been observed in the sample. We also use the contingency table representation for probabilities $\mathbf{p}$ for the joint occurrence of the set of categorical variables.

We let X and Y be binary random variables and denote by n_{ij} the observed cell counts in a 2×2 table $\mathbf{n}$. When we sum over a subscript we replace it by a '+'. Thus n_{i+} and n_{+j} denote the row and column totals, respectively, and these in turn sum to the grand total n_{++}. See the left-hand panel of Table 3.1. Similarly, we represent the *joint probability distribution* for X and Y as a 2×2 table of cell probabilities $\mathbf{p} = (p_{ij})$, where $p_{ij} = P(X = i, Y = j), i, j = 1, 2$, are non-negative and sum to one. See the right-hand panel of Table 3.1.

Table 3.1 *Notation for* 2×2 *tables: Sample point on the left and parameter value on the right.*

	Y_1	Y_2	Total
X_1	n_{11}	n_{12}	n_{1+}
X_2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n_{++}

	Y_1	Y_2	Total
X_1	p_{11}	p_{12}	p_{1+}
X_2	p_{21}	p_{22}	p_{2+}
Total	p_{+1}	p_{+2}	1

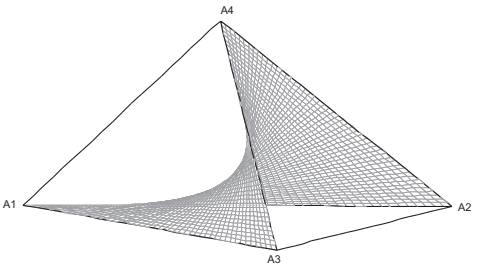

Fig. 3.1 Surface of independence for the 2×2 table. The tetrahedron represents the set of all probability distributions $\mathbf{p} = (p_{11}, p_{12}, p_{21}, p_{22})$ for the 2×2 tables, while the enclosed surface identifies the probability distributions satisfying the equation $p_{11}p_{22} = p_{12}p_{21}$, i.e., the toric variety for the model of independence.

Denote by $\mathbb{R}^4_p$ the four-dimensional real space with coordinates $\mathbf{p} = (p_{11}, p_{12}, p_{21}, p_{22})$. Geometrically, $\mathbf{p}$ is a point lying in a three-dimensional simplex (tetrahedron):

$$\mathbf{p} \in \Delta_3 = \{(p_{11}, p_{12}, p_{21}, p_{22}) : p_{ij} \geq 0, \sum_{i,j} p_{ij} = 1\}.$$

In barycentric coordinates, this tetrahedron of reference has vertices $A_1 = (1, 0, 0, 0)$, $A_2 = (0, 1, 0, 0)$, $A_3 = (0, 0, 1, 0)$, and $A_4 = (0, 0, 0, 1)$; see Figure 3.1. When the observed counts, $\mathbf{n} = \{n_{ij}\}$, come from a multinomial distribution, $Multi(N, \mathbf{p})$, we refer to Δ_3 as a *full parameter space*. If we consider a different parametrisation, the parameter space Θ parametrises a related surface.

The *marginal probability distributions* for X and Y are $\mathbf{p_X} = (p_{1+}, p_{2+}) = (s, 1-s)$ and $\mathbf{p_Y} = (p_{+1}, p_{+2}) = (t, 1-t)$. The lines A_1A_3 and A_2A_4 in the tetrahedron represent the set of all probability distributions, $\mathbf{p} = (s, 0, 1-s, 0)$ and $\mathbf{p} = (0, s, 0, 1-s)$ whose joint distributions are equivalent to the marginal distribution of $\mathbf{p_X} = (s, 1-s)$. Similarly, the lines A_1A_2 and A_3A_4 represent the set of all probability distributions, $\mathbf{p} = (t, 1-t, 0, 0)$ and $\mathbf{p} = (0, 0, t, 1-t)$, whose joint distributions are equivalent to the marginal distribution of $\mathbf{p_Y} = (t, 1-t)$.

We represent the *conditional probability distributions*, $\mathbf{p_{X|Y}}$ and $\mathbf{p_{Y|X}}$, by 2×2 *conditional probability matrices* $C = (c_{ij})$ and $R = (r_{ij})$, and denote by $\mathbb{R}^4_c$ and $\mathbb{R}^4_r$ the four-dimensional real spaces with coordinates $\mathbf{c} = (c_{11}, c_{12}, c_{21}, c_{22})$ and $\mathbf{r} = (r_{11}, r_{12}, r_{21}, r_{22})$, respectively. Given that we have observed $Y = j$, the conditional

probability values are $c_{ij} = P(X = i|Y = j) = p_{ij}/p_{+j}$, such that $\sum_{i=1}^{2} c_{ij} = 1, j = 1, 2$, and

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}.$$

Given that we have observed $X = i$, the conditional probability values are $r_{ij} = P(Y = j|X = i) = p_{ij}/p_{i+}$ such that $\sum_{j=1}^{2} r_{ij} = 1, i = 1, 2$, and

$$R = \begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix}.$$

Defined as such, the conditional probabilities can be considered as two-dimensional linear fractional transformations of either the cell counts or the cell probabilities. Recall that two-dimensional linear fractional transformations take the form $g(x, y) = (axy + cx + ey + g)/(bxy + dx + fy + h)$, e.g., $r_{11} = g(n_{11}, n_{12}) = n_{11}/(n_{11} + n_{12})$. The joint distribution $\mathbf{p}$ has the columns of C and rows of R as its conditional distributions. In the next section we provide a more careful geometric description of these conditionals.

We can now write the *odds ratio* or *cross-product ratio* for a 2×2 table

$$\alpha = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{c_{11}c_{22}}{c_{12}c_{21}} = \frac{r_{11}r_{22}}{r_{12}r_{21}}. \tag{3.1}$$

The odds ratio α is the fundamental quantity that measures the association in the 2×2 table whether we think in terms of probabilities that add to 1 across the entire table or conditional probabilities for rows, or conditional probabilities for columns. We can define two other odds ratios as follows:

$$\alpha^* = \frac{p_{11}p_{12}}{p_{22}p_{21}} = \frac{c_{11}c_{12}}{c_{22}c_{21}}, \tag{3.2}$$

$$\alpha^{**} = \frac{p_{11}p_{21}}{p_{12}p_{22}} = \frac{r_{11}r_{21}}{r_{12}r_{22}}. \tag{3.3}$$

Here α^* is characterised by the column conditionals and α^{**} by the row conditionals.

If we use the usual saturated log-linear model parametrization for the cell probabilities, e.g., see (Bishop *et al.* 1975) or (Fienberg 1980):

$$\log p_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

where $\sum_{i=1}^{2} u_{1(i)} = \sum_{j=1}^{2} u_{2(j)} = \sum_{i=1}^{2} u_{12(ij)} = \sum_{j=1}^{2} u_{12(ij)} = 0$, then it turns out that $u_{1(1)} = \frac{1}{4}\log\alpha^*$, $u_{2(1)} = \frac{1}{4}\log\alpha^{**}$, and $u_{12(11)} = \frac{1}{4}\log\alpha$. Thus we can use the three odds ratios in Equations (3.1), (3.2), and (3.3) to completely characterise the standard saturated log-linear model, and thus the joint distribution $\mathbf{p}$.

3.3 Parameter surfaces and other loci for 2×2 tables

(Fienberg and Gilbert 1970) show that (a) the locus of all points corresponding to tables with independent margins is a hyperbolic paraboloid (Figure 3.1), (b) the locus of all points corresponding to tables with constant degree of association, α, is a hyperboloid of one sheet (Figure 3.2), and (c) the locus of all points corresponding to tables with fixed both margins is a line. Clearly, the other odds ratios

in Equations (3.2) and (3.3) correspond to tables with constant column and row 'effects', respectively, and their surfaces are also hyperboloids of one sheet. All of these surfaces lie within the simplex Δ_3.

Fixing marginals implies imposing sets of linear constraints on the cell counts or the cell probabilities. We can fully specify log-linear models for the vector $\mathbf{p}$ of cell probabilities by a 0-1 *design matrix* $\mathbf{A}$, in the sense that, for each $\mathbf{p}$ in the model, $\log \mathbf{p}$ belongs to the row span of $\mathbf{A}$. The surface of independence, which geometrically represents the *independence model*, corresponds to the *Segre* variety in algebraic geometry (Figure 3.1). If we consider a knowledge of a single marginal, then the vector $\mathbf{p}$ is geometrically described by an intersection of a plane with the simplex, Δ_3. For example, fix the marginal $\mathbf{p_X}$. Then the plane, π_X, is defined by

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \ \mathbf{t} = \begin{pmatrix} s \\ 1-s \end{pmatrix}. \tag{3.4}$$

Similarly, we can define the plane π_Y for the fixed marginal $\mathbf{p_Y}$.

Now consider a set of linear constraints on the cell probabilities imposed by fixing conditional probabilities and clearing the denominators for the values from the matrix R (analogously from C). Then the vector $\mathbf{p}$ can be specified by a *constraint matrix* $\mathbf{A}$ and a vector $\mathbf{t}$ of the following form:

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ r_{12} & -r_{11} & 0 & 0 \\ 0 & 0 & r_{22} & -r_{21} \end{pmatrix}, \ \mathbf{t} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

In the related sample space of integer-valued tables, the constraint matrix $\mathbf{A}$ can also be constructed by using the observed conditional frequencies, or relevant observed cell counts, but adding the parameter N for the sample size as follows:

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ n_{12} & -n_{11} & 0 & 0 \\ 0 & 0 & n_{22} & -n_{21} \end{pmatrix}, \ \mathbf{t} = \begin{pmatrix} N \\ 0 \\ 0 \end{pmatrix}.$$

Hence, any contingency table with fixed marginals and/or conditional probability values is a point in a convex polytope defined by a linear system of equations induced by observed marginals and conditionals. An affine algebraic variety is the common zero set of finitely many polynomials. Thus our problem of finding the loci of all possible tables given an arbitrary set of conditionals and marginals for 2×2 tables translates into an algebraic problem of studying zero sets in $\mathbb{R}^4_p$.

In the next section we derive the geometric description of the parameter space of $\mathbf{p}$ for fixed values of conditional probabilities defined by matrices C and R.

3.3.1 Space of tables for fixed conditional probabilities

Consider a system of linear equations for four unknowns, $p_{11}, p_{12}, p_{21}, p_{22}$, imposed by observing or fixing conditional probabilities defined by the matrix R.

Proposition 3.1 *The locus of probability distributions* $\mathbf{p}$ *for a* 2×2 *table satisfying a set of conditional probability distributions defined by* R *is a ruling of two surfaces of constant associations,* α *and* α^{**}.

Proof Let $f_{p,r} : \mathbb{R}^4_p \setminus W \to \pi_r$ be the map given by $r_{ij} = p_{ij}/p_{i+}$, where W is a union of two varieties, $W = V(\langle p_{11}+p_{12}\rangle) \cup V(\langle p_{21}+p_{22}\rangle)$. Since $\sum_{j=1}^{2} p_{ij}/p_{i+} = 1$, $i = 1, 2$, the image of f is contained in the plane $\pi_r \subset \mathbb{R}^4_r$ of equations $r_{11} + r_{12} = 1, r_{21} + r_{22} = 1$, and we can represent a point $\mathbf{r}$ in this plane by the coordinates $\mathbf{r} = (r_{11}, r_{22})$. Then the preimage of a point $\mathbf{r} \in \pi_r, f^{-1}(\mathbf{r})$, is the plane in $\mathbb{R}^4_p$ of equations $(1 - r_{11})p_{11} - r_{11}p_{12} = 0$ and $-r_{22}p_{21} - (1 - r_{22})p_{22} = 0$.

Since we are interested in $\mathbf{p}$, we restrict the function $f_{p,r}$ on the simplex Δ_3. The intersection $\Delta_3 \cap V(\langle p_{11} + p_{12}\rangle)$ is the face $\overline{12}$, that is the line A_1A_2 consisting of the points of the form $\mathbf{p} = (s, 0, 1 - s, 0)$. Similarly, $\Delta_3 \cap V(\langle p_{21} + p_{22}\rangle)$ is the face $\overline{34}$ consisting of the points of the form $\mathbf{p} = (0, s, 0, 1 - s)$. With $\tilde{W} = \overline{12} \cup \overline{34}$, the map becomes $\tilde{f}_{p,r} : \Delta_3 \setminus \tilde{W} \to \pi_r$. Observe that the condition for the $\mathbf{p}$ to lie in $\Delta_3 \setminus \tilde{W}$ forces $0 \le r_{11} \le 1$ and $0 \le r_{22} \le 1$ such that $\tilde{f}_{p,r} : \Delta_3 \setminus (\tilde{W}) \to \Delta_1 \times \Delta_1$. Thus the preimage of a point $\mathbf{r} \in \pi_r$, $\tilde{f}^{-1}(\mathbf{r})$, is the segment in Δ_3 of equations

$$V_{\Delta_3} := \{(r_{11}s, (1 - r_{11})s, (1 - r_{22})(1 - s), r_{22}(1 - s)) : 0 < s < 1\}.$$

Finally take the closure of V for a given $\mathbf{r}$,

$$\overline{V}_{\Delta_3,r} := \{(r_{11}s, (1 - r_{11})s, (1 - r_{22})(1 - s), r_{22}(1 - s)) : 0 \le s \le 1, \text{fixed } \mathbf{r}\}, \quad (3.5)$$

and parametrise the probability variety by the probability of the margin s we condition upon. □

By taking the closure of V we can understand what is happening with points $\mathbf{p}$ in the closure of the parameter space; that is, the points of $\tilde{W}$. If $s = 0$ we obtain a point $T^* = (0, 0, (1 - r_{22}), r_{22})$ on the line A_3A_4, while if $s = 1$ we obtain a point $T = (r_{11}, 1 - r_{11}, 0, 0)$ on the line A_1A_2. The point T^* is in the closure of the preimage of every point in $\Delta_1 \times \Delta_1$ of the form (t, r_{22}), $0 \le t \le 1$. As t varies, the preimage of (t, r_{22}), that is the segment TT^*, represents a ruling of the surface with different odds ratio; see Figure 3.2. All these rulings pass through the same point (t, r_{22}). Recall from Equations (3.1) and (3.3) that the conditional distributions from R define the association coefficients α and α^{**}. For a fixed value of $\mathbf{r}$-parameter, as we vary the values of s, the segment defined in Equation (3.5) belongs to a family of lines that determine the surface of constant association α, which we denote as S_α. They are also rulings for the surface of constant association defined by α^{**}, that is of $S_{\alpha^{**}}$.

In a similar way, we define the map $f_{p,c} : \mathbb{R}^4_p \setminus W' \to \pi_c$ given by $c_{ij} = p_{ij}/p_{+i}$, where $W' = V(\langle p_{11} + p_{21}\rangle) \cup V(\langle p_{12} + p_{22}\rangle)$ and π_c the plane $\pi_c \subset \mathbb{R}^4_c$ of equations $c_{11} + c_{21} = 1, c_{12} + c_{22} = 1$. The segment with coordinates

$$\overline{V}_{\Delta_3,c} = \{(c_{11}t, (1 - c_{22})(1 - t), (1 - c_{11})t, c_{22}(1 - t)) : 0 \le t \le 1, \text{fixed } \mathbf{c}\}, \quad (3.6)$$

represents an equivalence class with fixed value of the matrix C that is the $\mathbf{c}$-parameter. Thus the lines SS^* are the second set of rulings for the *surface of*

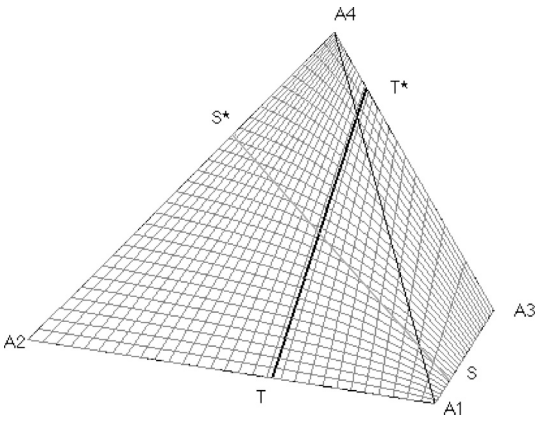

Fig. 3.2 Surface of constant association $\alpha = 6$. The line SS^* represents all probability distributions $\mathbf{p} = (p_{11}, p_{12}, p_{21}, p_{22})$ satisfying fixed **c**-conditional parameter. The line TT^* represent all probability distributions $\mathbf{p} = (p_{11}, p_{12}, p_{21}, p_{22})$ satisfying fixed **r**-conditional parameter.

constant association, α, and also rulings for the surface of association defined by α^*.

If X and Y are independent, then $\mathbf{p_{Y|X}} = \mathbf{p_Y}$ and $\mathbf{p_{X|Y}} = \mathbf{p_X}$. Thus, we confirm the result of (Fienberg and Gilbert 1970), who state that for *surface of independence* ($\alpha = 1$, see Figure 3.1), the rulings are two families of straight lines corresponding to constant column and row margins.

In the following sections we use the above described measures and their geometry, and consider the geometric interpretation of the Uniqueness Theorem, see (Gelman and Speed 1993, Arnold *et al.* 1996, Arnold *et al.* 1999), and complete specification of joint distribution via log-linear models. A geometric interpretation of incomplete specification of the joint distribution **p** is also considered.

3.4 Complete specification of the joint distribution

When we examine observed 2×2 tables, our statistical goal is usually to make inferences about the joint distribution of the underlying categorical variables, e.g., finding estimates of and models for **p**. In this section, we discuss possible complete specifications of the joint distribution and give their geometric interpretations. In Section 3.5, we turn to incomplete specifications, i.e., reduced models.

3.4.1 Specification I

From the definition of conditional probability, we know that the joint distribution for any 2×2 table is uniquely identified by one marginal and the related conditional:

$$P(X, Y) = P(X)P(Y|X) = P(Y)P(X|Y),$$

or equivalently $p_{ij} = p_{i+}r_{ij} = p_{j+}c_{ij}$.

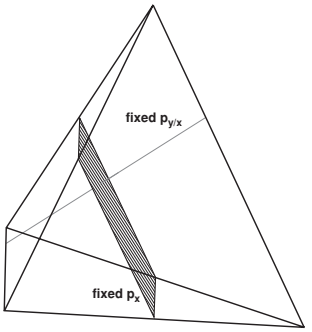

Fig. 3.3 Specification I. The intersection of the simplex Δ_3, the line for fixed $\mathbf{r}$, and the plane π_X, is a fully specified joint distribution $\mathbf{p}$.

We can use the geometric representations in Section 3.3 to demonstrate this uniqueness. For example, consider the locus of points $\mathbf{p}$ for fixed $\mathbf{r}$ as described by $\overline{V}_{\Delta_3,r}$ in Equation (3.5); see the line segment in Figure 3.3. The other locus of points $\mathbf{p}$ is a plane π_X defined by (3.4) observing a specific value of s corresponding to p_{1+}. The intersection of Δ_3 with these two varieties is a unique point representing the joint distribution $\mathbf{p}$. This is a geometric description of the basic factorisation theorem in statistics.

3.4.2 Specification II

The joint distribution for a 2×2 table is also fully specified by knowing two sets of conditionals: $\mathbf{p_{X|Y}}$ and $\mathbf{p_{Y|X}}$, equivalent to Specification I under independence of X and Y. Note that this is the simplest version of the Hammersley–Clifford theorem, see (Besag 1974).

Its geometric representation is the intersection of lines representing $\mathbf{p}$ for fixed $\mathbf{p_{Y|X}}$ and $\mathbf{p_{X|Y}}$ (Figure 3.2). It is an intersection of two varieties defined by Equations (3.5) and (3.6), $\overline{V}_{\Delta_3,r} \cap \overline{V}_{\Delta_3,c}$. Specifically, it is a point on the surface of the constant association, α, identifying the unique table given these conditional distributions.

Lemma 3.1 *The specification of joint distribution* $\mathbf{p}$ *by two sets of conditional parameters,* $\mathbf{r}$ *and* $\mathbf{c}$*, is equivalent to its specification by a saturated log-linear model.*

Proof Based on Proposition 3.1, each conditional includes full information on two out of three odds ratios; $\mathbf{r}$ has full information on α and α^{**}, while $\mathbf{c}$ has information on α and α^{*}. As seen at the end of Section 3.2 all three odds ratios together represent the key parameters of the saturated log-linear model and thus they fully characterise the joint distribution for a 2×2 table. □

This specification is clearly implicit in many treatments of log-linear models and 2×2 tables, e.g., see (Fienberg 1980), but to our knowledge has never been made explicit. We discuss further related specifications with odds ratios in Section 1.4.4.

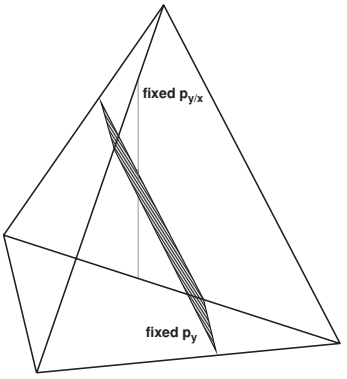

Fig. 3.4 Specification III. The intersection of the simplex Δ_3 with the line segment and the plane is a fully specified joint distribution $\mathbf{p}$.

3.4.3 Specification III

(Arnold *et al.* 1996, Arnold *et al.* 1999) show that sometimes a conditional and the 'wrong' marginal (e.g., $\mathbf{p_{Y|X}}$ and $\mathbf{p_Y}$) also uniquely identify the joint distribution, provided Arnold's *positivity condition*. Here the geometric representation of $\mathbf{p}$ lies in the intersection of simplex Δ_3 with $\overline{V}_{\Delta_3,r}$, see Equation (3.5) and Figure 3.4, and the plane π_Y, see Section 3.3. For 2×2 tables, this result *always* holds and states that for two *dependent* binary random variables, X and Y, either the collection $\{\mathbf{p_{X|Y}}, \mathbf{p_X}\}$ or $\{\mathbf{p_{Y|X}}, \mathbf{p_Y}\}$ uniquely identifies the joint distribution.

If the matrix $\mathbf{p} = (p_{ij})$ has rank 1, X and Y are independent and this implies that common odds ratio $\alpha = 1$. Since conditional distributions also preserve α, this implies that the ranks of matrices $C = (c_{ij})$ and $R = (r_{ij})$ are also both 1. Thus any rank greater than 1 implies a dependence between X and Y. Specifically for 2×2 tables, when the conditional matrices have full rank, X and Y are dependent random variables. We redefine the result on the uniqueness of the joint distribution.

Proposition 3.2 *For two binary discrete random variables,* X *and* Y*, either collection* $\{\mathbf{p_{X|Y}}, \mathbf{p_X}\}$ *or* $\{\mathbf{p_{Y|X}}, \mathbf{p_Y}\}$ *uniquely identifies the joint distribution if the conditional matrices* $C = (c_{ij})$ *and* $R = (r_{ij})$ *have full rank.*

Proof Consider $\mathbf{p_X} = (p_{1+}, p_{2+}) = (s, 1-s)$ and $\mathbf{p_{X|Y}} = (c_{11} = p_{11}/p_{+1}, c_{21} = p_{21}/p_{+1}, c_{12} = p_{12}/p_{+2}, c_{22} = p_{22}/p_{+2})$. Recall that we are assuming that there exists a joint probability distribution $\mathbf{p}$ from which $\mathbf{p_{X|Y}}$ and $\mathbf{p_X}$ are derived, and thus they are compatible. Imposing $p_{ij} \in [0,1]$ requires that either $0 \leq c_{11} \leq s \leq c_{12} \leq 1$ or $0 \leq c_{12} \leq s \leq c_{11}$. If the conditional matrix C has a full rank there are two linearly independent equations from observing $\mathbf{p_{X|Y}}$ that describe relationships on the cell probabilities (p_{ij}). If C has a full rank this implies that the marginal array $\mathbf{p_X}$ also has a full rank, and there are two additional linearly independent constraints describing relationships among the (p_{ij}).

Consider the ideal I generated by the four polynomials obtained after clearing the denominators in the ratios defining relationships between the conditionals c_{ij}'s

Table 3.2 *Representation of the joint distribution* $\mathbf{p}$ *as a function of the* $\mathbf{p_X} = (s, 1-s)$ *and the conditional* $\mathbf{p_{X|Y}} = (c_{11}, c_{12}, c_{21}, c_{22})$.

	Y_1	Y_2
X_1	$\frac{c_{11}(c_{12}-s)}{c_{12}-c_{11}}$	$\frac{-c_{12}(c_{11}-s)}{c_{12}-c_{11}}$
X_2	$\frac{c_{12}+sc_{11}-s-c_{11}c_{12}}{c_{12}-c_{11}}$	$\frac{(c_{11}-s)((c_{12}-1)}{c_{12}-c_{11}}$

and cell probabilities p_{ij}'s, namely $p_{11}+p_{12}-s, p_{21}+p_{22}-1+s, (1-c_{11})p_{11}-c_{11}p_{21}, c_{12}p_{22}-(1-c_{12})p_{12}$. Then a Gröbner basis of I using lexicographic order is $\{p_{21}+p_{22}+s-1, p_{11}+p_{12}-s, p_{12}c_{12}+p_{22}c_{12}-p_{12}, p_{12}c_{11}+p_{22}c_{11}-p_{12}+sc_{11}, p_{22}c_{11}-p_{22}c_{12}-sc_{12}+c_{11}c_{12}+s-c_{11}\}$. Set these polynomials equal to zero. Then, (1) if $c_{11} \neq c_{12}$, matrix C has a full rank, and the equivalent unique solution is given in Table 3.2; and (2) if $c_{11} = c_{12}$, then $c_{11} = 1$ or $c_{11} = s$. When $c_{11} = c_{12} = s$, we have independence of X and Y. However, if $c_{11} = c_{12} = s = 1$ then $\mathbf{p}$ is not identifiable. In this case the matrix C does not have a full rank and conditions of the proposition are not satisfied. Furthermore, $\mathbf{p} = \mathbf{p_Y}$ and solutions would lie on the face A_1A_2 or A_3A_4 of the simplex Δ_3 (see Figure 3.1). □

(Slavkovic 2004) derived a result similar to that in Theorem 4.2. but for $I \times 2$ tables. This characterisation is far more subtle than the previous two and we have not found it in any other setting.

3.4.4 Odds-ratio specification

In Section 3.2 we showed that all three odds ratios, α, α^*, and α^{**} together represent the key parameters of the saturated log-linear model: $\log p_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$. That is $u_{12(11)} = \frac{1}{4}\log\alpha$, $u_{1(1)=}\frac{1}{4}\log\alpha^*$, and $u_{2(1)} = \frac{1}{4}\log\alpha^{**}$, and thus they too specify the joint distribution for 2×2 tables. If we add a representation for the 'constant' term, i.e., $u = \frac{1}{4}\log(p_{11}p_{12}p_{21}p_{22})$, then the implicit representation of the joint distribution is defined by simultaneously solving the equations from

$$V_{\Delta_3} = (p_{11}p_{22} - \alpha p_{12}p_{21}, p_{11}p_{12} - \alpha^* p_{21}p_{22}, p_{11}p_{21} - \alpha^{**} p_{12}p_{22}). \tag{3.7}$$

Let $r_1 = p_{11}/p_{12} = r_{11}/r_{12}$ and $r_2 = p_{21}/p_{22} = r_{21}/r_{22}$ be the row odds. The column odds are $c_1 = p_{11}/p_{21} = c_{11}/c_{21}$ and $c_2 = p_{12}/p_{22} = c_{12}/c_{22}$. (Kadane *et al.* 1999) gave an alternative parametrisation to the one given by Equation (3.7), and showed in the context of capture–recapture type problems that it is sufficient to have α and the odds, r_1 and c_1 to identify the joint distribution. In this setting, r_1 are the odds of a unit being counted twice given that it was counted in the first sample, and c_1 is the odds of a unit being counted twice given that the same unit was counted in the second sample.

Geometrically, the intersection of the probability simplex, Δ_3, with two surfaces of constant associations is a line segment that would be defined by a fixed set of

Table 3.3 *Representation of the joint distribution* $\mathbf{p}$ *as a function of the margins* $\mathbf{p_X} = (s, 1-s)$ *and* $\mathbf{p_Y} = (t, 1-t)$*, and the odds ratios,* α, α^* *and* α^{**}.

	Y_1	Y_2
X_1	$\frac{\sqrt{\alpha\alpha^{**}}}{1+\sqrt{\alpha\alpha^{**}}}s = \frac{\sqrt{\alpha\alpha^{*}}}{1+\sqrt{\alpha\alpha^{*}}}t$	$\frac{1}{1+\sqrt{\alpha\alpha^{**}}}s = \frac{\alpha^{*}}{\alpha^{*}+\sqrt{\alpha\alpha^{*}}}(1-t)$
X_2	$\frac{\alpha^{**}}{\alpha^{**}+\sqrt{\alpha\alpha^{**}}}(1-s) = \frac{1}{1+\sqrt{\alpha\alpha^{*}}}t$	$\frac{\sqrt{\alpha\alpha^{**}}}{\alpha^{**}+\sqrt{\alpha\alpha^{**}}}(1-s) = \frac{\sqrt{\alpha\alpha^{*}}}{\alpha^{*}+\sqrt{\alpha\alpha^{*}}}(-t)$

conditional probabilities as we saw in Section 3.3.1. This line is one of the rulings for each of the respective hyperbolic surfaces for joint distributions $\mathbf{p}$ with constant associations. The observation naturally leads to an equivalence statement about Specification I and the following two sets of parameters: (1) $\{\mathbf{p_X}, \alpha, \alpha^{**}\}$ and (2) $\{\mathbf{p_Y}, \alpha, \alpha^{*}\}$. Let $\{\mathbf{p_X}, \mathbf{p_{Y|X}}\}$ and $\{\mathbf{p_Y}, \mathbf{p_{X|Y}}\}$ uniquely identify the joint distribution $\mathbf{p}$. Then the following lemma holds:

Lemma 3.2 *For a* 2×2 *table, the specification of* $\mathbf{p}$ *by* $\{\mathbf{p_X}, \mathbf{p_{Y|X}}\}$ *is equivalent to characterisation by* $\{\mathbf{p_X}, \alpha, \alpha^{**}\}$*, and* $\{\mathbf{p_Y}, \mathbf{p_{X|Y}}\}$ *is equivalent to characterisation by* $\{\mathbf{p_Y}, \alpha, \alpha^{*}\}$.

Proof The two odds ratios will completely specify the missing conditional distributions on the probability simplex (cf. Section 3.4), and thus completely specify the joint distribution. Consider the two ideals generated by

$$p_{11} + p_{12} - s, p_{21} + p_{22} - 1 + s, p_{11}p_{22} - \alpha p_{12}p_{21}, p_{11}p_{12} - \alpha^{*} p_{21}p_{22}$$

and

$$p_{11} + p_{21} - t, p_{12} + p_{22} - 1 + t, p_{11}p_{22} - \alpha p_{12}p_{21}, p_{11}p_{21} - \alpha^{**} p_{12}p_{22}.$$

Finding the Gröbner basis, and setting the defining polynomials equal to zero results in the solution in Table 3.3. More specifically, the probabilities $p_{ij} = g(\alpha, \alpha^{**})\mathbf{p_X} = h(\alpha, \alpha^{*})\mathbf{p_Y}$ where g, and h are functions of the three odds ratios given in Table 3.3. □

If $\alpha = 1$, $\mathbf{p} = \{\frac{\sqrt{\alpha^{**}}}{1+\sqrt{\alpha^{**}}}s, \frac{1}{1+\sqrt{\alpha^{**}}}s, \frac{\alpha^{**}}{\alpha^{**}+\sqrt{\alpha^{**}}}(1-s), \frac{\sqrt{\alpha^{**}}}{\alpha^{**}+\sqrt{\alpha^{**}}}(1-s)\}$. Clearly $\mathbf{p_{X|Y}} = \mathbf{p_X}$, and $\mathbf{p_Y} = \{\frac{\sqrt{\alpha^{**}}}{1+\sqrt{\alpha^{**}}}, \frac{1}{1+\sqrt{\alpha^{**}}}\}$ and we have independence of X and Y. If $\alpha = \alpha^{**} = 1$ then the joint distribution $\mathbf{p}$ is identified as $\{\frac{1}{2}s, \frac{1}{2}s, \frac{1}{2}(1-s), \frac{1}{2}(1-s)\}$. Notice that if $s = 1$ then $c_{11} = c_{12} = s = 1$ and $\mathbf{p}$ is not identifiable. Furthermore, $\mathbf{p} = \mathbf{p_Y}$ and potential solutions would lie on the face A_1A_2 or A_3A_4 of the simplex Δ_3. Similar considerations can be made for t, α, and α^{*}.

This specification is related to the parametrisation given by (Kadane *et al.* 1999). Then the following sets of parameters will also uniquely identify the joint distribution: (3) $\{\mathbf{p_X}, \alpha, r_1\}$ and (4) $\{\mathbf{p_Y}, \alpha, c_1\}$. These characterisations are different from any previously described in the literature and may be of special interest to those attempting to elicit joint distributions via components in a Bayesian context.

3.4.5 Specification via the non-central hypergeometric distribution

Finally we point out a well-established fact in statistical literature that both sets of one-way marginals, $\mathbf{p_X}$ and $\mathbf{p_Y}$, and the odds-ratio, α give a complete specification of the joint probability distribution $\mathbf{p}$ via the non-central hypergeometric distribution. Within Δ_3, as shown in (Fienberg and Gilbert 1970), the locus of joint probability distributions $\mathbf{p}$ given $\{\mathbf{p_X}, \mathbf{p_Y}\}$ is a line segment. This line segment intersects the hyperboloid specified by α in a unique point $V_{\Delta_3,s,t,\alpha}$ with coordinates

$$\left\{ \left(st, s(1-t), \frac{(1-s)t}{\alpha(1-t)+t}, \frac{\alpha(1-s)(1-t)}{\alpha(1-t)+t} \right) : \text{fixed } s, t, \alpha \right\}.$$

3.5 Incomplete specification of the joint distribution

Statistical models come from restricting values of one or more parameters and focusing on subspaces. A natural question arises as to the specification of the joint distribution if one of the parameters from the complete specification is set to zero or missing. For example, setting $\alpha = 1$ in Equation (3.7) defines the model of independence which corresponds to a hyperbolic paraboloid surface and the *Segre variety* in Figure 3.1.

3.5.1 Space of tables for a fixed marginal and odds-ratio

As noted in Section 3.4.5, both sets of one-way marginals and the odds-ratio, $\{\mathbf{p_X}, \mathbf{p_Y}, \alpha\}$ give a complete specification of $\mathbf{p}$ via the non-central hypergeometric distribution. In this section we consider the specification if one of the margins is missing.

Partial specification of the joint probability distribution $\mathbf{p}$ based solely on one odds-ratio, e.g., α, is an intersection of a hyperbolic surface with the probability simplex Δ_3, see (Fienberg and Gilbert 1970); knowledge of odds-ratio also specifies the locus of conditional distributions (see Section 1.5.2). Partial specification via one margin and α yields points lying on the intersection of a hyperbola and the probability simplex Δ_3:

$$V_{\Delta_3,s,\alpha} = \left\{ \left(st, s(1-t), \frac{(1-s)t}{\alpha(1-t)+t}, \frac{\alpha(1-s)(1-t)}{\alpha(1-t)+t} \right) : 0 \le t \le 1, \text{fixed } s, \alpha \right\} \quad (3.8)$$

as shown in Figure 3.5. This is a *rational parametric representation* requiring that $\alpha(1-t)+t \neq 0$ and it implies not conditioning on the event of probability zero.

3.5.2 Space of conditional tables

Proposition 3.3 *The locus of conditional distributions* $\mathbf{r}$ *or* $\mathbf{c}$*, given a fixed odds-ratio lies in the intersection of a quadric with the plane* π_r *or* π_c*, respectively.*

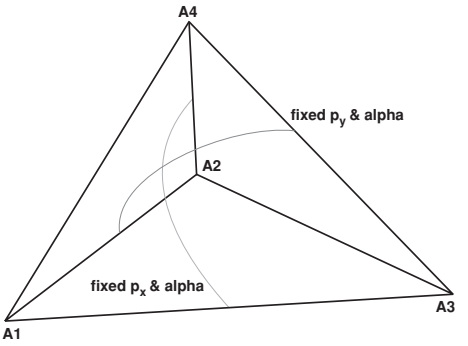

Fig. 3.5 Incomplete specification of the joint distribution $\mathbf{p}$ is given by the intersection of the simplex Δ_3 with the curve defined by one marginal and odds-ratio.

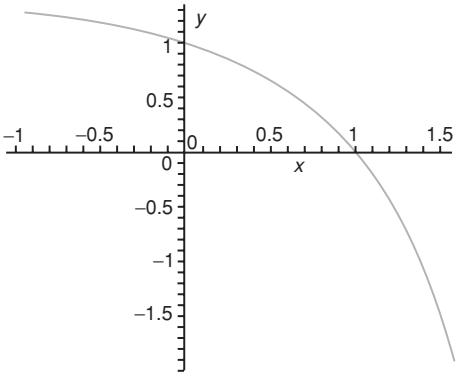

Fig. 3.6 Specification of the conditional distribution $\mathbf{p}_{\mathbf{Y}|\mathbf{X}}$ lies in the intersection of a quadric and π_r.

We treat the case of α and $\mathbf{r}$ and $\mathbf{c}$, but the α^{**} or α^{*} with either $\mathbf{r}$ or $\mathbf{c}$ would work in a similar way.

Proof Fix the odds-ratio α. Recall that the joint probabilities $\mathbf{p}$ satisfying the odds-ratio lie on the intersection of the hyperbolic surface S_α and Δ_3 where $S_\alpha := V(\langle p_{11}p_{22} - \alpha p_{12}p_{21}\rangle)$ and $\alpha = p_{11}p_{22}/p_{12}p_{21} = r_{11}r_{22}/r_{12}r_{21}$. Restrict our attention on the plane π_X. A bijection $\tilde{f}_{\pi_X} : \pi_X \to \pi_r$ given by

$$\begin{pmatrix} r_{11} \\ r_{22} \end{pmatrix} = \begin{pmatrix} \frac{1}{s} & 0 \\ 0 & \frac{1}{1-s} \end{pmatrix} \begin{pmatrix} p_{11} \\ p_{22} \end{pmatrix}$$

is the restriction of $\tilde{f}$ to the plane π_X. The image of surface S_α under the map $\tilde{f}$ is the curve

$$C_{r,\alpha} := V(\langle \alpha(1 - r_{11})(1 - r_{22}) - r_{11}r_{22}\rangle)$$

which is clearly the intersection of a quadric with the plane π_r. Similar derivation can be done for the intersection of a quadric and the plane π_c defined by the equation $\alpha(1 - c_{11})(1 - c_{22}) = c_{11}c_{22}$. □

Once we fix a plane π_X, the curve $C_{r,\alpha}$ is in the bijection with the curve $S_\alpha \cap \pi_X$. Note that this bijection exists only when you fixed a specific plane π_X which is needed to define a conditional distribution. In fact, a point $\mathbf{r}$ on the curve $C_{r,\alpha}$ has as preimage the segment $\overline{V}_{\Delta_3,r}$ defined by Equation (3.5). Once we fix a plane π_X, the preimage of $\mathbf{r}$ is exactly the point determined by the intersection $\overline{V}_{\Delta_3,r} \cap \pi_X$. If we fix another plane π'_X, the preimage of $\mathbf{r}$ will be another point in $\overline{V}_{\Delta_3,r}$ but given by the intersection $\overline{V}_{\Delta_3,r} \cap \pi'_X$. This corresponds with the fact that, given a conditional distribution $\mathbf{p_{Y|X}}$ (i.e., a point $\mathbf{r}$) and a marginal $\mathbf{p_X}$ (i.e., a plane π_X) the probabilities of $\mathbf{p}$ are uniquely determined (the point in the intersection $\overline{V}_{\Delta_3,r} \cap \pi_X$).

From the above we directly derived the corresponding probability variety given in Equation (3.8).

3.5.3 Margins

If we are given the row and column totals, then the well-known Fréchet bounds for the individual cell counts are:

$$\min(n_{i+}, n_{+j}) \geq n_{ij} \geq \max(n_{i+} + n_{+j} - n, 0) \text{ for } i = 1, 2, j = 1, 2.$$

The extra lower bound component comes from the upper bounds on the cells complementary to (i, j). These bounds have been widely exploited in the disclosure limitation literature and have served as the basis for the development of statistical theory on copulas (Nelsen 2006). The link to statistical theory comes from recognizing that the minimum component $n_{i+} + n_{+j} - n$ corresponds to the MLE of the expected cell value under independence, $n_{i+}n_{+j}/n$. For further details see (Dobra 2001, Dobra 2003) and Chapter 8 in this volume.

Geometric interpretation corresponds to fixing $\mathbf{p_X}$ and $\mathbf{p_Y}$, that is restricting the parameter space to the intersection of Δ_3 with π_X and π_Y, respectively (see Section 1.3). The points $\mathbf{p}$ then lie in intersection of Δ_3 with the segment $\pi_X \cap \pi_Y$ given by $C_{s,t} := V(\langle p_{11} - p_{22} - (s + t - 1)\rangle)$.

3.5.4 Two odds-ratios

In this section we address the question of specification of the joint probability distribution $\mathbf{p}$ when we have two odds ratios, e.g. α and α^*. This is the case when we are missing the marginal from the log-linear model specification, e.g., non-hierarchical log-linear model. We treat the case with α and α^{**}, but α^* would work in a similar way. This characterisation is related to the specifications of $\mathbf{p}$ discussed in Section 1.4.4, and results in Table 1.2. (Carlini and Rapallo 2005) describe an analogous question but with application to case-control studies.

Lemma 3.3 *The points* $\mathbf{p}$ *with given* α *and* α^{**} *lie in the intersection of* Δ_3 *with the line segment defined by*

$$V_{\alpha,\alpha^{**}} := \left\{ \frac{s\sqrt{\alpha\alpha^{**}}}{\sqrt{\alpha\alpha^{**}}+1}, \frac{s}{\sqrt{\alpha\alpha^{**}}+1}, \frac{\sqrt{\alpha^{**}}(1-s)}{\sqrt{\alpha_1}+\sqrt{\alpha^{**}}}, \frac{\sqrt{\alpha}(1-s)}{\sqrt{\alpha}+\sqrt{\alpha^{**}}} | 0 < s < 1 \right\}. \quad (3.9)$$

We first note that the partial specification based solely on two odds ratios uniquely specifies the missing conditional. We used this result in the proof of Lemma 2 in Section 1.4.4.

Proof The points in the plane π_r with the given odds ratio lie on two curves, $C_{r,\alpha} := V(\langle \alpha(1-r_{11})(1-r_{22}) - r_{11}r_{22} \rangle)$ and $C_{r,\alpha^{**}} := V(\langle \alpha^{**}(1-r_{11})r_{22} - r_{11}(1-r_{22}) \rangle)$ (see Section 1.5.2), whose intersection, $C_{r,\alpha} \cap C_{r,\alpha^{**}}$, consists of two points:

$$\begin{array}{ll} r_{11} = \frac{\sqrt{\alpha\alpha^{**}}}{1+\sqrt{\alpha\alpha^{**}}} & r_{12} = \frac{1}{1+\sqrt{\alpha\alpha^{**}}} \\ r_{21} = \frac{\sqrt{\alpha^{**}}}{\sqrt{\alpha}+\sqrt{\alpha^{**}}} & r_{22} = \frac{\sqrt{\alpha}}{\sqrt{\alpha}+\sqrt{\alpha^{**}}} \end{array}$$

or

$$\begin{array}{ll} r_{11} = \frac{\sqrt{\alpha\alpha^{**}}}{-1+\sqrt{\alpha\alpha^{**}}} & r_{12} = -\frac{1}{-1+\sqrt{\alpha\alpha^{**}}} \\ r_{21} = -\frac{\sqrt{\alpha^{**}}}{\sqrt{\alpha}-\sqrt{\alpha^{**}}} & r_{22} = \frac{\sqrt{\alpha}}{\sqrt{\alpha}-\sqrt{\alpha^{**}}} \end{array}$$

The second point does not represent conditional probabilities since it has two negative coordinates. The preimage of the other point is the segment given by Equation (3.9) which consists of points $\mathbf{p}$ in the intersection of the surfaces (in Δ_3) $S_\alpha := V(\langle p_{11}p_{22} - \alpha p_{12}p_{21} \rangle)$ and $S_{\alpha^{**}} := V(\langle p_{11}p_{21} - \alpha^{**}p_{12}p_{22} \rangle)$; that is, points $\mathbf{p}$ with given odds ratios α and α^{**}. The set $V_{\alpha,\alpha^{**}}$ corresponds to points on a ruling for each surface S_i. □

These line segments are the rulings discussed in Section 3.3.1, and thus describe the equivalent segments as when we fix the conditional, in this case, the $\mathbf{r}$-conditional (see Figure 3.2).

3.6 Extensions and discussion

The geometric representation described in Section 1.3.1 about the space of tables given fixed conditionals extend to $I \times J$ tables via linear manifolds. The specification results on $\mathbf{p}$ also generalise, in part (e.g., using $\mathbf{p}_{\mathbf{Y}|\mathbf{X}}$ and $\mathbf{p}_{\mathbf{X}}$), but when we are given margins we need to define multiple odds ratios. The bounds are also directly applicable to $I \times J$ tables and essentially a related argument can be used to derive exact sharp bounds for multi-way tables whenever the marginal totals that are fixed correspond to the minimal sufficient statistics of a log-linear model that is *decomposable*.

The natural extension to k-way tables is via log-linear models and understanding the specifications via fixed margins and combinations of margins and odds ratios,

and ratios of odds ratios. For $I \times J \times K$ tables, we use a triple subscript notation and we model the logarithms of the cell probabilities as

$$\log(p_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)} \quad (3.10)$$

where we set the summation of a u-term over any subscript equal to 0 for identification. There is a one-to-one correspondence between the u terms and odds ratio. For example, for $2 \times 2 \times 2$ tables, we can rewrite the parameters as a function of the logarithm of the cell probabilities

$$u_{123(111)} = \frac{1}{8} \log \left(\frac{\alpha^{(1)}}{\alpha^{(2)}} \right) \quad (3.11)$$

where $\alpha^{(k)} = p_{11k}p_{22k}/p_{12k}p_{21k}$. See (Bishop *et al.* 1975, Chapter 2) for further details. The toric variety corresponding to the model of no second-order interaction, i.e., $u_{123(ijk)} = 0$ for $i, j, k = 1, 2$, is a hyper-surface with three sets of generators corresponding to the first-order interactions, $p_{11k}p_{22k} - \alpha^{(k)}p_{12k}p_{21k}$, $p_{1j1}p_{2j2} - \alpha^{(j)}p_{1j2}p_{2j1}$, $p_{i11}p_{i22} - \alpha^{(i)}p_{i12}p_{i21}$, such that $\alpha^{(i=1)} = \alpha^{(i=2)}$, $\alpha^{(j=1)} = \alpha^{(j=2)}$, $\alpha^{(k=1)} = \alpha^{(k=2)}$. Each of the other subscripted u-terms in the log-linear model of Equation (3.10) can also be represented in terms of a ratio of odds ratios of the form of Equation (3.11).

3.6.1 Simpson's paradox

For three events A, B, and C, (Simpson 1951) observed that it was possible that $P(A|B) < P(A|\bar{B})$ (where $\bar{B}$ is the complementary set of B) but that $P(A|BC) > P(A|\bar{B}C)$ and $P(A|B\bar{C}) > P(A|\bar{B}\bar{C})$. This became known as Simpson's paradox although (Yule 1903) had made a similar observation 50 years earlier. For an extensive discussion of related aggregation phenomena, see (Good and Mittal 1987) and for an early geometrical treatment see (Shapiro 1982). As many authors have observed, another way to think about Simpson's paradox is as the reversal of the direction of an association when data from several groups are combined to form a single group. Thus for a $2 \times 2 \times 2$ table we are looking at three sets of 2×2 tables, one for each level of the third variable and another for the marginal table, and we can display all three within the same simplex Δ_3.

Consider the model of complete independence for a $2 \times 2 \times 2$ table:

$$\log p_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)}$$

where $u_{12(ij)} = u_{13(ik)} = u_{23(jk)} = u_{123(ijk)} = 0$, for $i, j, k = 1, 2$ that is the corresponding odds ratios and ratios of odds ratios are all equal to 1. Now consider the marginal 2×2 table with vector of probabilities $\mathbf{p} = (p_{ij+})$. The complete independence model implies marginal independence, i.e., $\log p_{ij+} = v + v_{1(i)} + v_{2(j)}$, so that the marginal odds ratios α_{12}=1, and $\mathbf{p}$ would be a point on the surface of independence.

Next suppose that variables 1 and 2 are conditionally independent given 3, i.e., $\log p_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)} + u_{23(jk)}$. The marginal odds ratio $\alpha_{12} \neq 1$, but the two conditional odds ratios for each level of the third variable equal one,

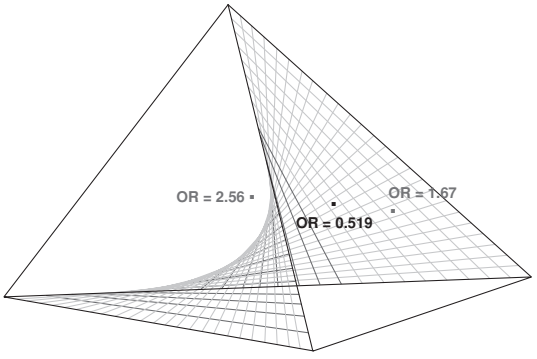

Fig. 3.7 An example of Simpson's paradox. Two dots with odds ratios (OR) > 1 are conditional 2×2 tables and on the same side of surface of independence. The $\mathbf{p}$ with odds-ratio (OR) < 1 is the marginal 2×2 table.

that is $\alpha_{12(3)} = 1$, and $\mathbf{p_{12|3}}$ would be two points on the surface of independence. When we connect such two points on the surface of independence, the line segment corresponds to tables with either positive association or negative association. This is the boundary for the occurrence of Simpson's paradox.

Simpson's paradox occurs when we have two tables corresponding to points lying on one side of the surface of independence, but the line segment connecting them cuts the surface and includes points on the 'other side'. Figure 3.7 gives one such example. If we put a probability measure over the simplex, we could begin to discuss 'the probability of the occurrence of Simpson's paradox,' cf. (Hadjicostas 1998).

When we connect two points lying on the surface of independence then we are combining two *different* independence models and the line connecting them will either consists of all weighted combinations of the two tables, or in the sense described above, all possible marginal tables. These will either all have values of $\alpha > 1$ or values of $\alpha < 1$ unless the two original tables being combined share either row or column margins, in which case $\alpha = 1$. The locus of all possible such lines corresponds to the $k = 2$ latent class model described in Chapter 2 in this volume and it consists of the entire simplex Δ_3.

3.7 Generalisations and questions

In this chapter we have employed an algebraic geometric approach to describe a variety of characterisations, both complete and incomplete, of bivariate distributions for two categorical variables. We have updated some older geometric representations of 2×2 contingency tables, e.g., from (Fienberg and Gilbert 1970), and we have described a series of new characterisations of the joint distribution using arbitrary sets of margins, conditionals, and odds ratios. We also considered incomplete characterisations of the joint distribution, and their links to latent class models and to Simpson's paradox. Many of the ideas explored here generalise rather naturally to $I \times J$ and higher-way tables. For higher-way tables, the usual characterisations corresponding to log-linear models come in terms of specifying marginal totals

(minimal sufficient statistics) and setting various sets of generalised odds ratios equal to zero. The number of such specifications grows dramatically with the dimensionality of the table.

Many questions remain to be explored; e.g. (i) What are the partial specifications arising from subset of ratio of odds ratios? (ii) When are subsets of odds ratios implied by conditionals? (iii) When do combinations of margins and conditionals reduce to higher-order margins? (iv) What are the implications of such results for bounds in contingency tables? About question (iv), see also Chapter 8 in this volume.

Acknowledgements

We thank Cristiano Bocci and Eva Riccomagno for helpful suggestions regarding some proofs. This research was supported in part by NSF Grant SES-0532407 to the Department of Statistics, Penn State University, NSF grants EIA9876619 and IIS0131884 to the National Institute of Statistical Sciences, NSF Grant DMS-0439734 to the Institute for Mathematics and Its Application at the University of Minnesota, and NSF Grant DMS-0631589 to Carnegie Mellon University.

References

Arnold, B., Castillo, E. and Sarabia, J. M. (1996). Specification of distributions by combinations of marginal and conditional distributions, *Statistics and Probability Letters* **26**, 153–57.

Arnold, B., Castillo, E. and Sarabia, J. M. (1999). *Conditional Specification of Statistical Models*, (New York, Springer-Verlag).

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion), *Journal of the Royal Statistical Society, Series B* **36**, 192–236.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice* (Cambridge, MA, MIT Press). Reprinted (2007) (New York, Springer-Verlag).

Carlini, E. and Rapallo, F. (2005). The geometry of statistical models for two-way contingency tables with fixed odds ratios, *Rendiconti dell'Istituto di Matematica dell'Università di Trieste* **37**, 71–84.

De Rooij, M. and Anderson, C.J. (2007). Visualizing, summarizing, and comparing odds ratio structures, *Methodology* **3**, 139–48.

De Rooij, M., and Heiser, W. J. (2005). Graphical representations and odds ratios in a distance-association model for the analysis of cross-classified data, *Psychometrika* **70**, 99–123.

Diaconis, P. (1977). Finite forms of de Finetti's theorem on exchangeability, *Synthese* **36**, 271–81.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions, *Annals of Statistics* **26**(1), 363–97.

Dobra, A. (2001). Statistical tools for disclosure limitation in multi-way contingency tables. PhD thesis, Department of Statistics, Carnegie Mellon University.

Dobra, A. (2003). Markov bases for decomposable graphical models, *Bernoulli* **9**(6), 1–16.

Edgeworth, F. Y. (1914). On the use of analytical geometry to represent certain kinds of statistics, *Journal of the Royal Statistical Society* **77**, 838–52.

Erosheva, E. A. (2005). Comparing latent structures of the grade of membership, Rasch, and latent class models, *Psychometrika* **70**, 619–28.

Fienberg, S. E. (1968). The geometry of an $r \times c$ contingency table, *Annals of Mathematical Statistics* **39**, 1186–90.

Fienberg, S. E. (1970). An iterative procedure for estimation in contingency tables, *Annals of Mathematical Statistics* **41**, 907–17. Corrigenda **42**, 1778.

Fienberg, S. E. and Gilbert, J. P. (1970). The geometry of a two by two contingency table, *Journal of the American Statistical Association* **65**, 694–701.

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data* 2nd edn (Cambridge, MA, MIT Press). Reprinted (2007) (New York, Springer-Verlag).

Fisher, R. A. (1921). On the interpretation of χ^2 from contingency tables, and the calculation of P, *Journal of the Royal Statistical Society* **85**, 87–94.

Flach, P. A. (2003). The geometry of ROC space: understanding machine learning metrics through ROC isometrics, In *Proc. ICML-2003*, Washington DC, 194–201.

Gelman, A. and Speed, T. P. (1993). Characterizing a joint probability distribution by conditionals, *Journal of the Royal Statistical Society. Series B* **55**, 185–8. Corrigendum **6**, 483 (1993).

Good, I. J. and Mittal, Y. (1987). The amalgamation and geometry of two-by-two contingency tables, *Annals of Statistics* **15**, 694–711. Addendum **17**, 947 (1989).

Greenacre, M. and Hastie, T. (1987). The geometric interpretation of correspondence analysis, *Journal of the American Statistical Association* **82**, 437–47.

Hadjicostas, P. (1998). The asymptotic proportion of subdivisions of a 2×2 table that result in Simpson's paradox, *Combinatorics, Probability and Computing* **7**, 387–96.

Heiser, W. J. (2004). Geometric representation of association between categories, *Psychometrika* **69**, 513–45.

Kadane, J. B., Meyer, M. M. and Tukey, J. W. (1999). Yule's association paradox and ignored stratum heterogeneity in capture-recapture studies, *Journal of the American Statistical Association* **94**, 855–9.

Kagan, A. M., Linnik, Y. V. and Rao, C. R. (1973). *Characterization Problems in Mathematical Statistics* (New York, John Wiley & Sons).

Kenett, R. S. (1983). On an exploratory analysis of contingency tables, *The Statistician* **32**, 395–403.

Lauritzen, S. L. (1996). *Graphical Models* (New York, Oxford University Press).

Luo, D., Wood, G. and Jones, G. (2004). Visualising contingency table data, *Australian Mathematical Society Gazette* **31**, 258–62.

Nelsen, R. B. (2006). *An Introduction to Copulas* 2nd edn (New York, Springer-Verlag).

Nelsen, R. B. (1995). Copulas, characterization, correlation, and counterexamples, *Mathematics Magazine* **68**, 193–8.

Pearson, E. S. (1956). Some aspects of the geometry of statistics, *Journal of the Royal Statistical Society. Series A* **119**, 125–46.

Pistone, G., Riccomagno, E. and Wynn, H. P. (2001). *Algebraic Statistics* (Boca Raton, Chapman & Hall/CRC).

Ramachandran, B. and Lau, K. S. (1991). *Functional Equations in Probability Theory* (New York, Academic Press).

Shapiro, S. H. (1982). Collapsing contingency tables – A geometric approach, *American Statistician* **36**, 43–6.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables, *Journal of the Royal Statistical Society. Series B* **13**, 238–41.

Slavkovic, A. B. (2004). Statistical disclosure limitation beyond the margins: characterization of joint distributions for contingency tables. PhD thesis, Department of Statistics, Carnegie Mellon University.

Slavkovic, A. B. and Sullivant, S. (2004). The space of compatible full conditionals is a unimodular toric variety, *Journal of Symbolic Computing* **46**, 196–209.

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics, *Biometrika* **2**, 121–34.