

# **Characterizing Multi-way Contingency Tables With Applications to Confidentiality and Statistical Disclosure Limitation**

**Stephen E. Fienberg<sup>1</sup>**

**Aleksandra B. Slavkovic<sup>2</sup>**

**<sup>1, 2</sup>Carnegie Mellon University**

**<sup>2</sup>Pennsylvania State University**

# Synopsis of Presentation

- We give several characterizations of probability distributions for two-way contingency tables using marginals, conditionals, and odds ratios.
  - Technical tools from **algebraic geometry, probability, directed acyclic graphs, and log-linear models.**
- These ideas generalize to higher dimensions.
- Partial specifications involves dropping of components from complete specifications.
- Statistical underpinnings for partial specifications offer insights for developing methodology for disclosure limitation.

# Example: 2<sup>4</sup> Table

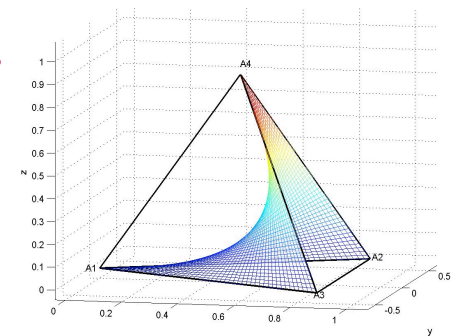
Center	Status	Response	Poor	Moderate	Excellent
		Treatment			
1	1	Active	3	20	5
1	1	Placebo	11	14	8
1	2	Active	3	14	12
1	2	Placebo	6	13	5
2	1	Active	12	12	0
2	1	Placebo	11	10	0
2	2	Active	3	9	4
2	2	Placebo	6	9	3

- **Example from Koch et al. (1982) on results of clinical trial for effectiveness of analgesic drug.**
- **Used in Fienberg and Slavkovic (2004) *Chance*, in press.**

# Contingency Tables

- For contingency tables, sample space is a simplex (of dimension 1 less than no. of cells) and values of r.v. are lattice points; parameter sets  $\Theta$  also lie in related simplex of same dimension.
  - *This is what makes the special link between contingency tables and algebraic geometry.*

For  $2 \times 2$  tables we can look at geometry because simplex  $\Sigma_4(1)$  is 3-dimensional—tetrahedron.



Independence:

$$\alpha = 1 \quad 4$$

# 2×2 Contingency Tables

- **Sample point**

	<i>Y</i>		
	$n_{11}$	$n_{12}$	$n_{1+}$
<i>X</i>	$n_{21}$	$n_{22}$	$n_{2+}$
	$n_{+1}$	$n_{+2}$	$n_{++}$

- **Parameter value**

	<i>Y</i>		
	$p_{11}$	$p_{12}$	$p_{1+}$
<i>X</i>	$p_{21}$	$p_{22}$	$p_{2+}$
	$p_{+1}$	$p_{+2}$	<b>1</b>

- **Characterizing distributions is working with parameter values. We'll assume  $p_{ij} > 0$ .**

$$(n_{11}, n_{12}, n_{21}, n_{22}) \in \Sigma_4(n_{++}) \quad (p_{11}, p_{12}, p_{21}, p_{22}) \in \Sigma_4(\mathbf{1})$$

# Specifying Joint Distributions for Two Binary Variables: I

	$Y$		
	$p_{11}$	$p_{12}$	$p_{1+}$
$X$	$p_{21}$	$p_{22}$	$p_{2+}$
	$p_{+1}$	$p_{+2}$	<b>1</b>

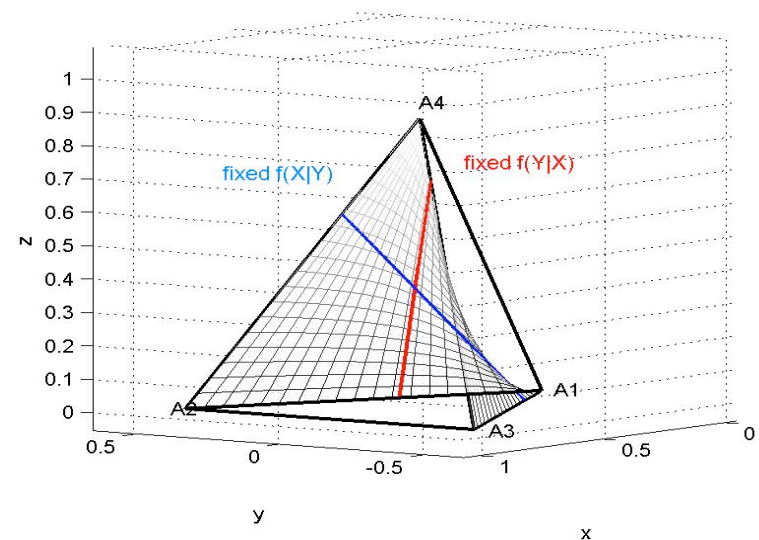
- $\Pr[X \& Y] = \Pr[X] \Pr[Y|X]$   
 $= \Pr[Y] \Pr[X|Y]$

$\implies$  marginal and related conditional in  $2 \times 2$  table specify joint distribution: e.g.,  $\{p_{1+}, p_{2+}\}$  and  $\{p_{11}/p_{1+}, p_{21}/p_{2+}\}$ .

# Specification II

- Can also use two sets of conditionals:  
 $\Pr[Y|X]$  and  $\Pr[X|Y]$ :
  - $\{p_{11}/p_{1+}, p_{21}/p_{2+}\}$  and  $\{p_{11}/p_{+1}, p_{12}/p_{+2}\}$ .
  - Conditionals are different from marginals.
  - Idea of varying  $X$  and considering locus of  $P(Y|X)$ .

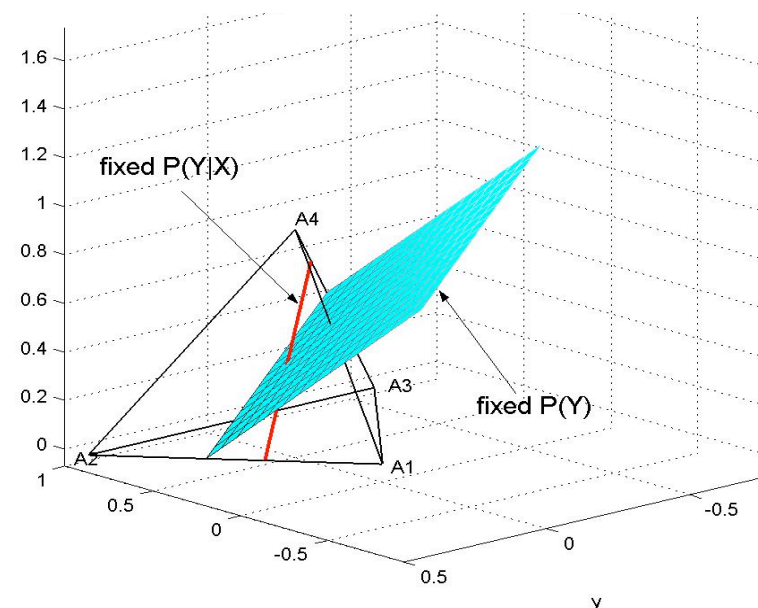
**Hammersley-Clifford  
Theorem!**



# Specification III

- What about conditional  $\{p_{11}/p_{1+}, p_{21}/p_{1+}\}$  and margin  $\{p_{+1}, p_{+2}\}$ , i.e.,  $P(Y|X)$  &  $P(Y)$ ?

Unique for 2x2 table case,  
but not in general!  
(Slavkovic, 2004)

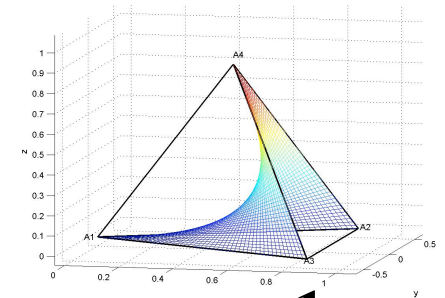


# Log-Linear Model for 2x2 Table

- For  $(i,j)$  cell in table:

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

$$\sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_i u_{12(ij)} = \sum_j u_{12(ij)} = 0.$$



$$\alpha = 1$$

- Independence if  $u_{12(ij)} = 0$  or  $p_{ij} = p_{i+}p_{+j}$ .
- $u_{12(ij)}$  captures something dependence between row and column variables through odds ratio:

$$u_{12(11)} = \frac{1}{4} \log \left[ \frac{p_{11}p_{22}}{p_{12}p_{21}} \right] = \frac{1}{4} \log \alpha$$

- $u_{1(i)}$  and  $u_{2(j)}$  measure margins on centered log scale.

# More Odds Ratios

- Can redefine all log-linear model parameters in terms of new log-odds ratios:

$$u = \frac{1}{4} \log[p_{11} p_{12} p_{21} p_{22}]$$

$$u_{12(11)} = \frac{1}{4} \log \left[ \frac{p_{11} p_{22}}{p_{12} p_{21}} \right] = \frac{1}{4} \log \alpha$$

$$u_{1(1)} = \frac{1}{4} \log \left[ \frac{p_{11} p_{12}}{p_{21} p_{22}} \right] = \frac{1}{4} \log \alpha^*$$

$$u_{2(1)} = \frac{1}{4} \log \left[ \frac{p_{11} p_{21}}{p_{12} p_{22}} \right] = \frac{1}{4} \log \alpha^{**}$$

# Specifications IV and V

- **Can also specify joint distribution using:**
  - IV. Both sets of 1-way marginals,  $\{p_{1+}, p_{2+}\}$  and  $\{p_{+1}, p_{+2}\}$ , and odds-ratio,  $\alpha$ .**
    - Partial specification based solely on marginals, corresponds to line through tetrahedron.**
  - V. The 3 odds ratios:  $\alpha$ ,  $\alpha^*$ , and  $\alpha^{**}$  corresponding to log-linear model parameters.**
- **Statistical models come from restricting values of one or more parameters, e.g., setting  $\alpha=1$  (*independence*), and focusing on subspaces.**
  - **Many (all?) interesting models are algebraic varieties.**

# Algebraic Geometry Representation

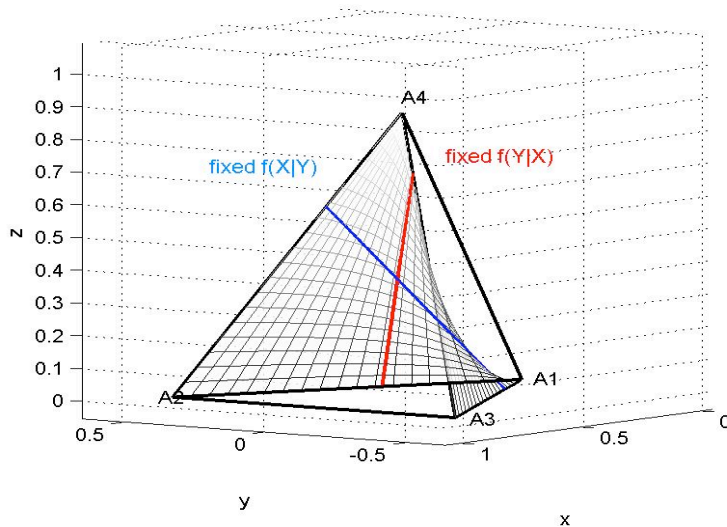
- Consider mapping from  $(p_{11}, p_{12}, p_{21}, p_{22})$  into product and the three odds ratios.
  - 1-1 when restricted to  $\Sigma_4$ .
- Ideal characterized by three polynomials:  
 $\langle p_{11}p_{22} - \lambda p_{12}p_{21}, p_{11}p_{12} - \mu p_{22}p_{21}, p_{11}p_{21} - \nu p_{12}p_{22} \rangle$
- Suppose we fix values of two of  $\alpha$ ,  $\alpha^*$ , and  $\alpha^{**}$ .  
What is locus values for third?
  - Traces out a hyperbola running over values for third.

# Conditionals and Odds Ratios

- Conditionals of form  $\Pr[Y|X]$  include full information on 2 of 3 odds ratios:

$$\frac{(p_{11} / p_{1+})(p_{22} / p_{2+})}{(p_{12} / p_{1+})(p_{21} / p_{2+})} = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \alpha,$$

$$\frac{(p_{11} / p_{1+})(p_{21} / p_{2+})}{(p_{12} / p_{1+})(p_{22} / p_{2+})} = \frac{p_{11}p_{21}}{p_{12}p_{22}} = \alpha^{**}.$$



# Other Issues in Specification

- **Feasibility:** Given components, does there exist joint distribution?
  - Major issue of whether we allow for zero probabilities in conditionals.
    - **Slavkovic & Sullivant (2004)**
- If partial specifications are for an actual table, what are possible tables satisfying them?
- How do we put distributions over possible tables conditional on the other information in partial specification?
  - Markov bases (from algebraic geometry)!

# Partial Specification and Two-Way Fréchet Bounds

- For  $2 \times 2$  tables  $\{p_{ij}\}$  given the marginal totals  $\{p_{1+}, p_{2+}\}$  and  $\{p_{+1}, p_{+2}\}$ :

$$\begin{array}{cc|c} p_{11} & p_{12} & p_{1+} \\ p_{21} & p_{22} & p_{2+} \\ \hline p_{+1} & p_{+2} & 1 \end{array}$$

$$\min(p_{i+}, p_{+j}) \geq p_{ij} \geq \max(p_{i+} + p_{+j} - 1, 0)$$

- Same ideas work for partial specification of table of non-negative counts.
- Interesting multi-way generalizations.

# Generalizations for $I \times J$ Tables

- **Geometric representation works but we move to flats (lines) and linear manifold (ruled surfaces).**
- **Specification results extend, in part (e.g., using  $P(Y|X)$  and  $P(X)$ ), but when we are given margins we need to define multiple odds-ratios.**
- **Result for  $P(Y|X)$  and  $P(Y)$  now involves a set of possible tables and we need to compute bounds or enumerate.**

# Partial Specifications for $I \times J$ Tables

- Given  $P(x|y)$  and  $P(x)$ , unique solution *exists* for  $I \times J$ , if matrix with values  $P(x|y)$  has full rank and  $I \geq J$ .

$a_{11}$	$a_{12}$	$P(x y)$
$a_{21}$	$a_{22}$	
$a_{31}$	$a_{32}$	

- For  $I \times 2$  table:

$$p_{ij} = a_{ij} \frac{p_{i+} - a_{i\{J \setminus j\}}}{a_{ij} - a_{i\{J \setminus j\}}}, \quad i \in I, j \in \{1, 2\}.$$

- If  $I < J$ , can actually get range of values.

**Slavkovic (2004)**

# Bounds given $P(x|y)$ and $P(x)$ When $I < J$

- Example 2x3 table, bounds on  $p_{11}$ :

$$UB = \begin{cases} a_{11} \frac{p_{1+} - \max\{a_{12}, a_{13}\}}{a_{11} - \max\{a_{12}, a_{13}\}} & \text{if } p_{1+} \geq a_{11} \\ a_{11} \frac{p_{1+} - \min\{a_{12}, a_{13}\}}{a_{11} - \min\{a_{12}, a_{13}\}} & \text{if } p_{1+} < a_{11} \end{cases}$$

$p_{11}$	$p_{12}$	$p_{13}$
$p_{21}$	$p_{22}$	$p_{23}$

$a_{11}$	$a_{12}$	$a_{13}$
$a_{21}$	$a_{22}$	$a_{23}$

$P(x|y)$

$$LB = \begin{cases} \max\{0, L \quad s.t. \quad L \leq UB\} & \text{if } p_{1+} \geq a_{11} \\ \max\{0, U \quad s.t. \quad U \leq UB\} & \text{if } p_{1+} < a_{11} \end{cases}$$

**Slavkovic (2004)**

# Log-linear Models for Multi-way Tables

- In  $I \times J \times K$  tables, we model logs of  $\{p_{ijk}\}$ :

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

with summation constraints.

- Can also represent models in terms of **ratios of odds-ratios**, e.g., in  $2 \times 2 \times 2$  table:

$$u_{123(111)} = \frac{1}{8} \log \left( \frac{p_{111} p_{221} p_{122} p_{212}}{p_{121} p_{211} p_{112} p_{222}} \right)$$

and more generally all u-terms are given in Hadamard-like form as

$$\text{for } I \subseteq \{1, 2, 3\}, \quad u_{I(111)} = \frac{1}{8} \log \left( \prod_{i,j,k=1}^2 p_{ijk}^{(-1)^{(i,j,k) \cdot I}} \right)$$

# Some Interesting Questions:

1. What are partial specifications from subset of ratio of odds-ratios?
2. When are subsets of odds ratios implied by conditionals?
3. When do combinations of margins and conditionals reduce to margins?
  - **Wermuth condition!**
4. When combinations don't reduce, how do we get bounds?
  - **Generate Markov basis and traverse tables!**

# Specification of 3-D Discrete Distributions

- **Can get complete characterization of joint distribution of  $(X, Y, Z)$  for 3 binary random variables in terms of**
  - **Marginals — 1-way or 2-way.**
  - **Conditionals — given 1 way or 2-way totals.**
  - **Some subset of 7 ratios of odds-ratios that make up log-linear model.**
- **Partial specifications based on marginals and conditionals produce bounds.**
- **Generalizations to higher dimensions are interesting and have additional features!**

# Links of Margins & Conditionals to Log-linear/Graphical Models

- **Margins are MSS for log-linear models.**
- **Undirected graphical models based on multiple conditional independence relationships are log-linear.**
  - **Decomposable models have triangulated graphs:**
    - Representable as intersections of Segre varieties
    - Explicit formulae for bounds in partial specifications
    - Parallel results on Markov bases: All bases are quadratic.
  - **Regular graphs decompose into pieces for calculating bounds and Markov bases.**
- **Margins and conditionals define Directed Acyclic Graph used for causal modeling.**

# Example With Margins

- Need to include margin for explanatory variables [CST].
- Two interesting well-fitting models:
  - 1. [CST][CSR] and 2. [CST][CSR][TR]
  - $\Delta G^2=5.4$  on 2 d.f.

Releasing[CST],  
[CSR], and [TR]  
seems safe.

Center	Status	Response Treatment	Poor	Moderate	Excellent
1	1	Active	[0,14]	[1,28]	[0,13]
1	1	Placebo	[0,14]	[6,33]	[0,13]
1	2	Active	[0,9]	[3,27]	[1,17]
1	2	Placebo	[0,9]	[0,24]	[0,16]
2	1	Active	[2,21]	[3,22]	[0,0]
2	1	Placebo	[2,21]	[0,19]	[0,0]
2	2	Active	[0,9]	[0,16]	[0,7]
2	2	Placebo	[0,9]	[2,18]	[0,7]

# Example with Conditionals

- **Suppose we try to relax bounds by releasing slightly more information via conditionals [RT] and [RCS]:**
  - **Releasing either [R|T] and [CST] is equivalent to releasing [RT] and [CST].**
  - **Releasing [R|T] and [R] will uniquely identify [RT] because number of levels in R is greater than in T.**
  - **[R|CS] and [R] will not uniquely identify [RCS], yet because of small sample size, space of tables is same?**
  - **Markov basis for [R|CS] includes basis from [RCS] plus extra generator.**
    - **No. tables for [RCS] is 11,081,397,760,00.**
    - **No. tables for [R|CS] is 11,081,579,235,840.**

# Basis for [R|CST]

- Dealing with [R|CST] introduces lots of complexities and likely comes close to uniquely describing [RCST].
- Basis is of size  $27 \times 24$ .

```

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 -1 2 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 -6 -13 -5 12 12 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 -10 -10 -1 11 10 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 0 1 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -8 -8 0 0 0 0 3 9 4 0 0
0 0 0 0 0 0 0 -14 14 29 -6 -13 -5 -2 -2 -1 0 0 0 0 0 0 0 0
0 0 0 0 0 0 -3 -26 0 26 0 0 0 1 1 1 0 0 0 0 0 0 0 0
0 0 0 0 0 0 -3 -12 -14 -3 6 13 5 4 4 0 0 0 0 0 0 0 0 0
-3 -20 -5 0 0 0 3 12 14 3 0 0 0 1 1 0 0 0 0 0 0 0 -2 -4
0 0 0 -11 -14 -8 3 12 14 3 0 0 0 0 0 1 0 0 0 0 0 0 0 0
-3 -20 -5 11 14 8 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 -2 -4
-3 -20 -5 0 0 0 6 52 0 -52 6 13 5 2 2 0 0 0 0 0 0 0 -2 -4
-3 -20 -5 11 14 8 3 26 0 -26 0 0 0 -1 -1 0 0 0 0 0 0 0 -2 -4
0 0 0 -11 -14 -8 6 52 0 -52 6 13 5 1 1 1 0 0 0 0 0 0 0
0 0 0 -11 -14 -8 9 78 0 -78 6 13 5 0 0 0 0 0 0 0 0 0 0 0
-3 -20 -5 0 0 0 12 104 0 -104 0 0 0 11 11 0 0 0 0 0 0 0 -2 -4
0 0 0 -11 -14 -8 12 104 0 -104 0 0 0 10 10 1 0 0 0 0 0 0 0
0 0 0 -11 -14 -8 15 130 0 -130 0 0 0 9 9 0 0 0 0 0 0 0 0 0
-3 -20 -5 0 0 0 18 156 0 -156 0 0 0 8 8 0 0 0 0 0 0 0 -2 -4
0 0 0 -11 -14 -8 18 156 0 -156 0 0 0 7 7 1 0 0 0 0 0 0 0
0 0 0 -11 -14 -8 21 182 0 -182 0 0 0 6 6 0 0 0 0 0 0 0 0
-3 -20 -5 0 0 0 24 208 0 -208 0 0 0 5 5 0 0 0 0 0 0 0 -2 -4
0 0 0 -11 -14 -8 24 208 0 -208 0 0 0 4 4 1 0 0 0 0 0 0 0
0 0 0 -11 -14 -8 27 234 0 -234 0 0 0 3 3 0 0 0 0 0 0 0 0
-3 -20 -5 0 0 0 30 260 0 -260 0 0 0 2 2 0 0 0 0 0 0 0 -2 -4
0 0 0 -11 -14 -8 30 260 0 -260 0 0 0 1 1 1 0 0 0 0 0 0 0
0 0 0 -11 -14 -8 33 286 0 -286 0 0 0 0 0 0 0 0 0 0 0 0 0

```

# Synopsis of Presentation

- We gave several characterizations of probability distributions for two-way contingency tables using marginals, conditionals, and odds ratios.
  - Technical tools from **algebraic geometry, probability, directed acyclic graphs, and log-linear models.**
- Partial specifications involves dropping of components from complete specifications.
- Statistical underpinnings for partial specifications offer insights for developing methodology for disclosure limitation.
- Illustrated selected ideas with  $2^4$  example.

# The End

- **Thanks to the NISS Confidentiality Project, NSF, and funding statistical agencies.**