Characterizing Joint Distributions for Contingency Tables with Applications to Statistical Disclosure Limitation

Aleksandra B. Slavkovic
Department of Statistics, Penn State University
sesa@stat.psu.edu

(Joint work with Steve Fienberg)
IMA Minneapolis, MN --- Tue., March 6, 2007
Outline: Two parts

- Outline of several characterizations of probability distributions for two-way contingency tables using marginals, conditionals, and odds ratios.
  - Tools from algebraic geometry, probability, and log-linear models.
  - Partial specifications involves dropping of components from complete specifications.
  - 2x2 table, but these ideas generalize to higher dimensions.
  - Simpson’s paradox

- Algebraic Statistics & Statistical Disclosure Limitation
  - Partial specifications offer insights for developing methodology for disclosure limitation.
Statistical Disclosure Limitation

- Agencies seek to release confidential data to the public, and to improve analyses, but such that
  - do not reveal identities or sensitive attributes,
  - are useful for a wide range of analyses,
  - are easy for analysts and agencies to use
- The release of partial information

- Data Utility: Are released data (often tables) useful for statistical inference?

- Risk of Disclosure: Are the results of their analysis safe to be released?

Privacy cartoon by Chris Slane at:
cagle.msnbc.com/news/ PrivacyCartoons/main.asp
Problem Statement

- Consider \( K \) random variables \( X = (X_1, ..., X_K) \) each taking values on a finite set \([d_k] = \{1,2,\ldots,d_k\}\)

- A \( K \)-way contingency table of counts \( n = n(i), i \in D, D = [d_1] \times \cdots \times [d_K] \) is a point in a simplex of dimension \( D-1 \); values of \( X_i \) are lattice points. Parameter sets \( \Theta \) also lie in related simplex of same dimension.
  - *Link between contingency tables and algebraic geometry.*

- \( n \in \mathbb{R}^\Theta \) is an element of the vector space of real functions such that
  - joint array: \( p_{i_1,\ldots,i_k}(x_K) = \Pr(X_1 = i_1, ..., X_k = i_k) \)
  - marginal array: \( p(x_A) = \sum_{K \setminus A} p(x_K) \)
  - conditional array: \( p(x_A \mid x_B) = \frac{p(x_{AB})}{p(x_B)} \)
Problem Statement

- Characterization (complete & incomplete) of discrete joint distributions for contingency tables given an arbitrary collections, \( T \), of conditionals, marginals, and/or odds-ratios

- Problem 1 (Uniqueness): Characterize conditions for which \( T \) uniquely identifies the joint distributions \( p(x) \) via different factorizations

- Problem 2 (Bounds): Find the upper and lower bounds on each cell entry

\[
U(i) = \max \{ n(i) : n \in \mathcal{F}_i \} \quad \text{and} \quad L(i) = \min \{ n(i) : n \in \mathcal{F}_i \}
\]

where fiber

\[
\mathcal{F}_i = \{ n \in \mathbb{N}^D : t = (\text{conditional arrays}(n), \text{marginal arrays}(n)) \}
\]

- Problem 3 (Uniqueness & Bounds): Find a Markov basis for studying the set of tables satisfying \( T \)

- Data confidentiality: Determine risk and utility in terms of arbitrary set \( T \)
Two sides of Algebraic Statistics

1. **Representation of a Statistical Model**: alternative description of the parameter space.
   - Log-Linear Models and in general Exponential Families of Discrete Distributions are toric varieties (defined by binomial equations).
   - Affect the quality of inference
   - The margins are sufficient statistics and inference is **conditional** on the observed margins and MLE-based.

2. **Conditional Inference**: study and characterization of portions of the sample space and, in particular, of all datasets (i.e., tables) having the observed margins and/or conditionals
   - **counting** and/or **sampling** over the possible space of tables
   - **compute** sharp upper and lower **bounds** on cell entries
   - Important for assessment of risk of disclosure.
   - “exact distributions” of the sample statistics
Based on Koch (1983) data on effectiveness of an analgesic drug

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Response</th>
<th>Poor</th>
<th>Good</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td></td>
<td>21</td>
<td>76</td>
</tr>
<tr>
<td>Placebo</td>
<td></td>
<td>34</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>55</td>
<td>139</td>
</tr>
</tbody>
</table>

To estimate the parameters, we have to first consider how the data were collected!

\[ \hat{p}_{ij} = \frac{n_{ij}}{n} \]

\[ n_{11} \sim \text{Binomial } (n_{1+}, P(R|T=\text{active})) \]
\[ n_{21} \sim \text{Binomial } (n_{2+}, P(R|T=\text{placebo})) \]

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Response</th>
<th>Poor</th>
<th>Good</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td></td>
<td>0.11</td>
<td>0.39</td>
</tr>
<tr>
<td>Placebo</td>
<td></td>
<td>0.18</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.29</td>
<td>0.71</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Response</th>
<th>Poor</th>
<th>Good</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td></td>
<td>0.22</td>
<td>0.78</td>
</tr>
<tr>
<td>Placebo</td>
<td></td>
<td>0.35</td>
<td>0.65</td>
</tr>
</tbody>
</table>
### 2x2 Contingency Table

- For 2x2 tables we can look at geometry in a probability simplex $\Delta_3(1)$

- **Sample point** $(n_{11}, n_{12}, n_{21}, n_{22}) \in \Delta_3(n_{++})$

- **Parameter point** $P= (p_{11}, p_{12}, p_{21}, p_{22}) \in \Delta_3(1)$

<table>
<thead>
<tr>
<th></th>
<th>Y=1</th>
<th>Y=2</th>
</tr>
</thead>
<tbody>
<tr>
<td>X=1</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
</tr>
<tr>
<td>X=2</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
</tr>
<tr>
<td></td>
<td>$n_{+1}$</td>
<td>$n_{+2}$</td>
</tr>
</tbody>
</table>

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

- $p_{ij} = P(X=i, Y=j)$ for all $i, j$
2x2 Contingency Table: Conditionals

- \( d_{ij} = P(Y = j | X = i) = \frac{p_{ij}}{p_{i+}} \)
- \( c_{ij} = P(X = i | Y = j) = \frac{p_{ij}}{p_{j+}} \)

\[
\begin{array}{ccc}
  & d_{11} & d_{12} \leftarrow 1 \\
 d_{21} & d_{22} \leftarrow 1 \\
\end{array}
\]

\((d_{11}, d_{12}, d_{21}, d_{22}) \in \Pi_d\)

\[
\begin{array}{ccc}
  & c_{11} & c_{12} \\
 c_{21} & c_{22} \leftarrow 1 \\
\end{array}
\]

\((c_{11}, c_{12}, c_{21}, c_{22}) \in \Pi_c\)

- \( f_{p,d} : R^4_p \setminus V \rightarrow \Pi_d \subset R^4_d \) where \( \Pi_d \) is defined by equations: \( d_{11} + d_{12} = 1, d_{21} + d_{22} = 1 \)
- Represent points \( d \) in \( \Pi_d \) by the set of coordinates \( d = (d_{11}, d_{22}) \).
- \( f^{-1}(d) \), is the plane in \( R^4_p \) of equations \(<(1 - d_{11})p_{11} - d_{11}p_{12} = 0, -d_{22}p_{21} - (1 - d_{22})p_{22} = 0>\).

- Fix now the marginal probabilities \( p_{1+}, p_{2+} \).
- Plane \( \Pi_X \) of equations \(<p_{11} + p_{12} = p_{1+}, p_{21} + p_{22} = p_{2+}>\)
- The restriction of \( f \) at the plane \( \Pi_X \) is a bijection \( f: \Pi_X \rightarrow \Pi_d \) given by

\[
\begin{pmatrix}
  d_{11} \\
  d_{22}
\end{pmatrix} = \begin{pmatrix}
  \frac{1}{p_{1+}} & 0 \\
  0 & \frac{1}{p_{2+}}
\end{pmatrix}\begin{pmatrix}
  p_{11} \\
  p_{22}
\end{pmatrix}
\]
Uniqueness: Complete specification of the joint

- Full disclosure; ability to completely reconstruct the original table

- Uniqueness in two-way tables:
  - $P(x), P(y \mid x)$
  - $P(x \mid y), P(y \mid x)$
  - $P(y), P(y \mid x)$

- $W := \{ p_{11} + p_{12} = p_{1+}, p_{21} + p_{22} = 1 - p_{1+} \}$
- $V := \{ (d_{11}s, (1 - d_{11})s, (1 - d_{22})(1 - s), d_{22}(1 - s) \mid 0 < s < 1 \}$
- $V \cap W := \{ (d_{11}p_{1+}, (1 - d_{11})p_{1+}, (1 - d_{22})(1 - p_{1+}), d_{22}(1 - p_{1+}) \}$

- Interpretation: prospective sampling

Log-Linear Model for 2×2 Table

- For \((i,j)\) cell in table:
  \[
  \log p_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}
  \]
  \[
  \sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_i u_{12(i)} = \sum_j u_{12(j)} = 0.
  \]

- Odds ratio
  \[
  \alpha = \frac{p_{11}p_{22}}{p_{12}p_{21}} \quad \text{or} \quad p_{11}p_{22} - \alpha p_{12}p_{21} = 0
  \]

- \(u_{12(ij)}\) 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.

- Statistical models come from restricting values of one or more parameters and focusing on subspaces.
  - Statistical models are algebraic varieties.

- Model of independence (Segre variety) if
  \[
  u_{12(ij)} = 0 \quad \text{or} \quad p_{ij} = p_i p_j \quad \text{or} \quad p_{11}p_{22} - p_{12}p_{21} = 0
  \]

\((0.11, 0.39, 0.18, 0.32)\)
Log-linear models & more Odds Ratios

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

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

- Can also specify joint distribution using the 3 odds ratios: \( \alpha, \alpha^*, \) and \( \alpha^{**} \) corresponding to log-linear model parameters.

\[
\langle p_{11}p_{22} - \lambda p_{12}p_{21}, p_{11}p_{12} - \mu p_{22}p_{21}, p_{11}p_{21} - \nu p_{12}p_{22} \rangle
\]
Conditionals and Odds Ratios

- Conditionals of form $p[y|x]$ and $p[x|y]$ include full information on 2 of 3 odds ratios:

$$\frac{d_{11}d_{22}}{d_{12}d_{21}} = \frac{c_{11}c_{22}}{c_{12}c_{21}} = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \alpha,$$

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

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

- Claim: conditionals are the rulings for the surface of constant associations, $\alpha$
- Interpretation: Gibbs sampling & capture-recapture problems
Specification of the joint

- Both sets of 1-way marginals, \( \{p_{1+}, p_{2+}\} \) and \( \{p_{+1}, p_{+2}\} \), and odds-ratio, \( \alpha \).
- Partial specification based solely on one odds ratio, e.g., \( \alpha \), gives a quadric in \( \pi_d \) and a hyperbolic surface in the \( \Delta_3(1) \).
- Partial specification via one margin and \( \alpha \) give a hyperbola in \( \Delta_3(1) \).

\[
V_{\Delta}(I) = (st, s(1-t), \frac{(1-s)t}{\alpha(1-t) + t} \frac{\alpha(1-s)(1-t)}{\alpha(1-t) + t} : 0 \leq t \leq 1, \text{fixed } s, \alpha)
\]

- Interpretation: prospective and retrospective sampling
Specification of the joint

- Both sets of 1-way marginals, \( \{p_{1+}, p_{2+}\} \) and \( \{p_{+1}, p_{+2}\} \), and odds-ratio, \( \alpha \) via non-central hypergeometric distribution

- Partial specification based solely on marginals, corresponds to line through tetrahedron.

\[
\begin{array}{ccc}
p_{11} & p_{12} & p_{1+} \\
p_{21} & p_{22} & p_{2+} \\
p_{+1} & p_{+2} & 1
\end{array}
\]

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

- Interpretation: Two-Way Fréchet Bounds
Specification of the joint

- The 2 odds ratios: \( \alpha \), and \((\alpha^*, \text{ or } \alpha^{**})\) and one margin corresponding to log-linear model parameters.

\[
V_{\alpha_1, \alpha_3} := \left\{ \frac{\sqrt{\alpha_1 \alpha_3}}{1 + \sqrt{\alpha_1 \alpha_3}} s, \frac{1}{1 + \sqrt{\alpha_1 \alpha_3}} s, \frac{\alpha_3}{\alpha_3 + \sqrt{\alpha_1 \alpha_3}} (1 - s), \frac{\sqrt{\alpha_1 \alpha_3}}{\alpha_3 + \sqrt{\alpha_1 \alpha_3}} (1 - s) \right\}
\]

- Partial specification based solely on two odds ratios, e.g., \( \alpha \) and \( \alpha^{**} \) uniquely specifies the missing conditional, and in \( \Lambda_3(1) \) the points \( (p_{ij}) \) lie in the intersection of

\[
S_{\alpha_i} := V(p_{11}p_{22} - \alpha_1 p_{12} p_{21}) \quad \text{and} \quad S_{\alpha_3} := V(p_{11}p_{21} - \alpha_3 p_{12} p_{22})
\]

- corresponds to points in a ruling in each surface

- Interpretation: sensitivity and specificity parameters, non-hierarchical models
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 $P(Y)$ in general now involves a set of possible tables and we need to compute bounds or enumerate.
  - Uniqueness Theorem for a special cases (Slavkovic 2004)
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
\[
\text{for } I \subseteq \{1,2,3\}, \quad u_{I(111)} = \frac{1}{8} \log \left( \prod_{i,j,k=1}^{2} p_{ijk}^{(-1)}(i,j,k)^I \right)
\]
Simpson’s paradox

- *When marginal association have different direction from each conditional association*

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Response</th>
<th>Poor</th>
<th>Good</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>0.11</td>
<td>0.39</td>
<td>0.5</td>
</tr>
<tr>
<td>Placebo</td>
<td>0.18</td>
<td>0.32</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.29</td>
<td>0.71</td>
<td></td>
</tr>
</tbody>
</table>

- OR=0.519

- **Health status = bad**  OR=2.5

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Response</th>
<th>Poor</th>
<th>Good</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>0.06</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>Placebo</td>
<td>0.01</td>
<td>0.15</td>
<td></td>
</tr>
</tbody>
</table>

- **Health status = good**  OR=1.67

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Response</th>
<th>Poor</th>
<th>Good</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>0.05</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Placebo</td>
<td>0.17</td>
<td>0.17</td>
<td></td>
</tr>
</tbody>
</table>
Red points are on the opposite side of the surface of independence relative to the marginal table.

$OR = 2.56$

$OR = 1.67$

$OR = 0.519$
Some Interesting Questions

- Generalization of multi-way tables
- What are partial specifications from subset of ratio of odds-ratios?
- When are subsets of odds ratios implied by conditionals?
- When do combinations of margins and conditionals reduce to margins?
  - Wermuth condition!
- When combinations don’t reduce, how do we get bounds?
  - Generate Markov basis and tranverse tables!
- What are geometric interpretations of other statistics
  - E.g., specificity, sensitivity, difference in proportions, relative risk, etc..
Example: Clinical trial data (Koch (1983))

- Effectiveness of an analgesic drug measured at two different centers, and two different health conditions, with two treatments (1=Active, 2=Placebo), and responses (1=Poor, 2=Not Poor).

<table>
<thead>
<tr>
<th>Center</th>
<th>Status</th>
<th>Treatment</th>
<th>Poor</th>
<th>Moderate</th>
<th>Excellent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Active</td>
<td>3</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>11</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Active</td>
<td>3</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>6</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Active</td>
<td>12</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>11</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Active</td>
<td>3</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>6</td>
<td>9</td>
<td>3</td>
</tr>
</tbody>
</table>

- Possible margins release for well-fitted models:
  \([\text{CST}][\text{CRT}][\text{CTR}]\)  \([\text{CST}][\text{CRT}][\text{TR}]\)  \([\text{CST}][\text{CRT}]\)
Conditional Inference: sampling with Markov bases

- It is possible to perform a random walk on the space of all the tables with a given set of margins (or conditionals).
  - It requires the identification of moves: integer valued vectors in the kernel of $A$ that, added to the current table, will produce a table with same margins.

- **Markov Bases:** minimal set of moves that preserve connectedness in the fiber.
  - Computed with algebraic software, but this amounts to finding the minimal generators of a set of polynomials defined by $A$:
    \[ I = \langle x^u_+ - x^u_-, \forall u \in \text{kernel}(A) \cap \mathbb{N}^l \rangle \]

- Using Markov Bases, it is possible to build a Gibbs sampler that can be used to explore the fiber and estimate the posterior distribution of the tables given the margins and the distribution of statistics over the fiber (usually Likelihood Ratio, Pearson’s $\chi^2$).
  - Known for the margins generalized hypergeometric distribution
  - For the conditionals, no generalization yet.
For the [CST][CSR] release there are 12 elements in the Markov Basis.

<table>
<thead>
<tr>
<th>Center</th>
<th>Status</th>
<th>Treatment</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>Active</td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Active</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>Active</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Active</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

**Table 1:**

<table>
<thead>
<tr>
<th></th>
<th>-1</th>
<th>+1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

**Table 2:**

<table>
<thead>
<tr>
<th></th>
<th>-1</th>
<th>+1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

**Table 3:**

<table>
<thead>
<tr>
<th></th>
<th>-1</th>
<th>+1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

**Table 4:**

<table>
<thead>
<tr>
<th></th>
<th>-1</th>
<th>+1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
Conditional Inference: Optimizing

- **Table entry data security problem:** compute sharp lower and upper bounds for the entries in a table with given margins, conditionals and/or odds-ratios.
  - This gives a way of assessing potential risk of disclosure.
  - It is an Linear Integer Program and Non-linear Integer program

- Algebraic techniques are being developed to solve Integer Programs.
  - Available symbolic software packages can be used to solve (small) problems: **Latte**, **4ti2**.

- Another active area of research is the study of the **integer gap**: difference from the solution obtained using LP-relaxations. Serkan and Sturmfels (2003). Results indicate that the gap can be considerable.
Conditional inference given the margins: counting & optimizing

- Need to include margin for explanatory variables [CST].
- Two interesting well-fitting models with $\Delta G^2 = 5.4$ on 2 d.f.:
  - 1. [CST][CRS] 65,419,200 tables and 2. [CST][CRS][RT] 108,490 tables

<table>
<thead>
<tr>
<th>Center</th>
<th>Status</th>
<th>Treatment</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Poor</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Active</td>
<td>3 [0,14]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>11 [0,14]</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Active</td>
<td>3 [0,9]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>6 [0,9]</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Active</td>
<td>12 [2,21]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>11 [2,21]</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Active</td>
<td>3 [0,9]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Placebo</td>
<td>6 [0,9]</td>
</tr>
</tbody>
</table>

- Safe to release

Software: Latte, 4ti2
Conditional inference given the conditionals: counting & optimizing

- Release full conditional $[R | CST]$ and sample size

<table>
<thead>
<tr>
<th>Center</th>
<th>Status</th>
<th>Response Treatment</th>
<th>Poor</th>
<th>Moderate</th>
<th>Excellent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Active</td>
<td>3 [1, 17.03]</td>
<td>20 [6.67,113.55]</td>
<td>5 [1.67, 28.39]</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Placebo</td>
<td>11 [1.38, 51.26]</td>
<td>14 [1.75, 65.23]</td>
<td>8 [1, 37.28]</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Active</td>
<td>3 [1, 16.48]</td>
<td>14 [4.67, 76.91]</td>
<td>12 [4, 65.92]</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Placebo</td>
<td>6 [1.2, 38.61]</td>
<td>13 [2.60, 83.66]</td>
<td>5 [1, 32.18]</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Active</td>
<td>12 [1.10, 79.44]</td>
<td>12 [1, 72.26]</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Placebo</td>
<td>11 [1.10, 79.48]</td>
<td>10 [1, 72.26]</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Active</td>
<td>3 [1, 29.06]</td>
<td>9 [3, 87.17]</td>
<td>4 [1, 38.74]</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Placebo</td>
<td>6 [2, 51.89]</td>
<td>9 [3, 77.83]</td>
<td>3 [1, 25.94]</td>
</tr>
</tbody>
</table>

- There are 7,703,002 tables
- These are LP relaxation bounds
- For data privacy, it is safe to release this conditional

Results of clinical trial for effectiveness of analgesic drug
Data source: Koch et al. (1982)
Bounds from the posterior distribution of $R|CST$

<table>
<thead>
<tr>
<th>Cell</th>
<th>True Value</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Mean</th>
<th>Median</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1,1)</td>
<td>3</td>
<td>0</td>
<td>6</td>
<td>2.7009</td>
<td>3</td>
<td>0.9176728</td>
</tr>
</tbody>
</table>

Compared to LP: [1, 17.03]

Presence of integer gaps which can strongly influence the disclosure risk and utility.
Bounds given \([R|CS]\) vs. \([CSR]\)

- Number (tables | conditional) \(\geq\) Number (tables | corresponding margin)
  - 31,081,579,235,840 vs. 31,081,397,760,000

- Markov basis for \([R|CS]\) includes elements from \([CRS]\) plus
  
  \[
  \begin{pmatrix}
  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 & -1 & 0 & 0 & 1 & 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 & -1 & 0 & 0 & 1 & 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 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  -14 & -34 & -34 & 0 & 0 & 0 & -9 & -27 & -17 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  -14 & -34 & -34 & 0 & 0 & 0 & 18 & 34 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  -42 & -102 & -39 & 0 & 0 & 0 & -9 & -27 & -17 & 0 & 0 & 0 & 138 & 132 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  -59 & -156 & -52 & 0 & 0 & 0 & 9 & 27 & 17 & 0 & 0 & 0 & 115 & 110 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
  -70 & -170 & -65 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 
  \end{pmatrix}
  
  - LP relaxation bounds wider than for the margin
  - But, sharp bounds are the same!
    - Last cell \([3], [0, 39.74], [0,7]\)
Posterior distribution (CRS,T) vs (R|CS,T)

<table>
<thead>
<tr>
<th>Cell</th>
<th>True Value</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>Mean</th>
<th>Median</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1,1,1)</td>
<td>3</td>
<td>3</td>
<td>14</td>
<td>8.7557</td>
<td>9</td>
<td>1.8847</td>
</tr>
</tbody>
</table>

Histogram of $x[,1]$

Histogram of $x[,1]$
Practical Implications

- Partial specifications involves dropping of components from complete specifications.
- Statistical underpinnings for partial specifications offer insights for developing methodology for disclosure limitation.

- Agencies already release conditionals in 2-way and 3-way tables, and higher k-way
  - Releasing full conditionals too risky
  - Small conditionals may release less information (less disclosure) than corresponding marginals

- Algebraic geometry useful for exploring the space of tables
  - Size of the move may determine uniqueness
  - Number of tables as a measure for disclosure evaluation
  - Computing sharp bounds
  - Implication for distributions
  - Synthetic data
  - Works well for smaller tables
Open problems

- **Q1:** Bounds given the odds-ratios
  - non-linear programming in probability simplex
  - non-linear integer programming in the sample space
  - Or maybe another way?

- **Q2:** Integer gap in the bounds given the conditionals & odds-ratios?

- **Q3:** Can we characterize the difference?
  - \([R | CS]\) and \([R]\) will not uniquely identify \([RCS]\), yet because of the small sample size, the bounds are the same but the space of tables is the different.

- **Q4:** Markov bases for k-way tables given fixed conditionals
  - Markov bases for tables larger than \(2^6\)
  - Markov bases and approximations: rounding of conditionals give different moves
Open problems (contd.)

- Algebraic risk for doubly random swapping
- 2x2x2 table and that we preserve one-way margins, [X], [Y], [Z].
- What is the distance between the Original data (OD) and Masked data (MD)
  - Q5: Finding the set of all tables, MD, that are at distance k moves from the OD
  - Q6: Finding the minimal Markov distance
Acknowledgments

- Joint work with Stephen E. Fienberg

- Juyoun Lee, Cristiano Bocci, Edoardo Airoldi, Alessandro Rinaldo

- NSF grant SES-0532407 to the Department of Statistics, Penn State

- References:
  - [http://www.stat.psu.edu/~sesa](http://www.stat.psu.edu/~sesa)
  - [http://www.stat.cmu.edu/~fienberg](http://www.stat.cmu.edu/~fienberg)
  - [http://www.niss.org/dgii/techreports.html](http://www.niss.org/dgii/techreports.html)