Tabular Data: From Margins to Margins and Conditionals

NCHS, Data Confidentiality --- Fri., May 2, 2008

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

Tabular Data -- Frequency data

- Contingency tables: cross-cross-classifies individuals by attributes
- Publicly available data as **marginal** and **conditional** tables

- Strike a balance between *data utility* and *disclosure risk*
  - *Utility* tied to usefulness of *marginal totals* & *log-linear models*
  - *Risk measure* is ability to identify small cell counts (e.g. via bounds, Dobra (2002))

	Education Level of Head of Household				
County	Low	Medium	High	Very High	Total
Alpha	15	1	3	1	20
Beta	20	10	10	15	55
Gamma	3	10	10	2	25
Delta	12	14	7	2	35
Total	50	35	30	20	135

- What can we release about this data to achive the balance? How can we protect it?

Delinquent Children by County & Education Level
Source: OMB Statistical Policy Working Paper 22

Statistical Disclosure Limitation (Control) methods

- Apply to microdata and/or tabulated data before release after identifying sensitive data
- **Data masking:** Transform the original data (matrix *X*) to the disseminated data (*Y*)
  - Y=AXB + C
  - A=record transformation, B=attribute transformation, C=noise addition
- Traditional approaches
  - Aggregation: Rounding, Topcoding & Tresholding
  - Suppression, e.g., cell suppression
  - Data Perturbations
  - Data Swapping
- Modern approaches: Sampling and Simulation techniques
  - Synthetic data
  - Remote access servers
  - Secure computation
  - Partial information releases

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

Recent methods: Simulation & Sampling

- Digital Governemnt Project I & II at NISS
- World "without original micordata"

- Synthetic & Partially synthetic data uses Bayesian methodology
  - Raghunathan, Reiter, and Rubin (2003, *JOS*)
  - Reiter (2003, *Surv. Meth.;* 2005, *JRSSA*)

- Partial data releases for tabular data
  - Dobra et al. (2003)
  - Slavkovic (2004), Fienberg & Slavkovic (2005), Fienberg et al. (2006)

- Remote access servers
  - Rowland (2003, NAS Panel on Data Access).
  - Gomatam, Karr, Reiter, Sanil (2005, *Stat. Science*)

- "Secure" statistical analysis/computation
  - Benaloh (1987, *CRYPTO86*)
  - Karr, Lin, Sanil, and Reiter (2005, NISS tech. rep.)

Problem Statement

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

- A K-way contingency table of counts $\mathbf{n} = \mathbf{n}(i)$, $i \in \mathcal{D}, \mathcal{D} = [d_1]\text{x}\ldots\text{x}[d_K]$ is a point in a simplex of dimension $\mathcal{D}$*-1; v*alues of X_i are lattice points. Parameter sets Θ also lie in related simplex of same dimension.
  - *Link between contingency tables and algebraic geometry.*

- $\mathbf{n} \in \mathbf{R}^{\mathcal{D}}$ is an element of the vector space of real functions such that

$$\text{joint array}: \quad p_{i_1 \ldots i_k}(x_K) = \Pr(X_1 = i_1,..., X_k = i_k)$$

$$\text{marginal array}: \quad p(x_A) = \sum_{K \setminus A} p(x_K)$$

$$\text{conditional array}: \quad p(x_A \mid x_B) = \frac{p(x_{AB})}{p(x_B)}$$

Partial data releases for tabular data

- Goal: Determine safe releases in terms of arbitrary set of marginals and/or conditionals
  - Assume data reported without error, compatible margins and conditionals, and unweighted counts
  - Currently we are exploring extensions to odds ratios, and to these assumptions
  - Non-interactive (but could potentially be used for interactive)

- Given observed information
  - Assessing potential risk of disclosure
  - if we can uniquely identify the joint distribution, that is a full disclosure
  - if we have a partial specification of the joint distribution, we may use bounds and/or distributions over the space of possible solutions to assess the risk of disclosure and the data utility

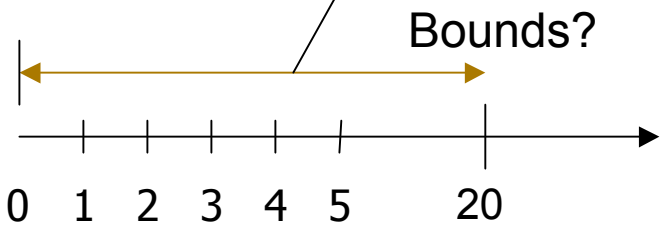
- Using tools from linear/integer programming, specification of joint distributions and algebraic statistics

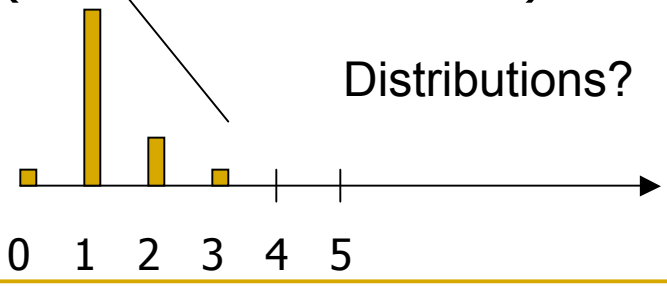
- Close links to perturbation, data swapping, synthetic data and remote access servers
- Dobra et al. (2003), Slavkovic (2004, 2005)

Delinquent Children by County & Education Level

	Education Level of Head of Household				
County	Low	Medium	High	Very High	Total
Alpha	15	1	3	1	20
Beta	20	10	10	15	55
Gamma	3	10	10	2	25
Delta	12	14	7	2	35
Total	50	35	30	20	135

18,272,363,056 tables have our margins (De Loera & Sturmfels).





Data Source: OMB Statistical Policy Working Paper 22 & S. Roehrig

What we know: Margins

- Optimization: Linear & Integer programming

Consider an $I \times J$ table

$$\begin{aligned} &\textit{Min } \mathbf{cn} \\ &\textit{s.t. } \mathbf{An} = \mathbf{b} \\ &\quad \mathbf{n} \geq \mathbf{0} \end{aligned}$$

c is a row vector of lenght d

n is a column vector of lenght d

A is a $m \times d$ matrix

b is a column vector of lenght m

Consider a 2x2 table

$$\mathbf{c} = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} n_{11} \\ n_{12} \\ n_{21} \\ n_{22} \end{pmatrix} = \begin{pmatrix} m1 \\ m2 \\ m3 \\ m4 \end{pmatrix}$$

- Fréchet bounds for IxJ tables: $\min\{n_{i+}, n_{+j}\} \geq n_{ij} \geq \max\{0, n_{i+} + n_{+j} - n_{++}\}$
- Dobra and Feinberg (2001, 2003) have made extensions to *k*-way tables
  - Explicit formula for decomposable models; closed form solutions for MLEs
  - Reducible models
  - Sharp bounds via "Shuttle Algorithm" exploits hierarchical structure within table, and sequentially updates bounds for cells

What we know: Margins

- 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:

$$I = \left\langle x^{u+} - x^{u-}, \forall u \in \text{kernel}(A) \cap \mathbb{N}^I \right\rangle$$

- Using Markov Bases, it is possible to build a Gibbs sampler 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 χ^2).
  - Sampling from generalized hypergeometric distribution

Delinquent Children by County & Education Level

- Release observed conditional frequencies

$$P(Education \mid County) = \begin{pmatrix} 0.750 & 0.050 & 0.150 & 0.050 \\ 0.364 & 0.182 & 0.182 & 0.273 \\ 0.120 & 0.400 & 0.400 & 0.080 \\ 0.343 & 0.400 & 0.200 & 0.057 \end{pmatrix}$$

Link to: ***Support*** and ***confidence*** for "***association rules***" in data mining!

- IP: no feasible solution unless using original counts
- LP relaxation bounds:

County	Low	Medium	High	Very High
Alpha	15 [0.75, 99]	1 [0.05, 6.6]	3 [0.15, 19.8]	1 [0.05, 6.6]
Beta	20 [0.36, 48]	10 [0.18, 24]	10 [0.18, 24]	15 [0.27, 36]
Gamma	3 [0.12, 15.84]	10 [0.4, 52.8]	10 [0.4, 52..8]	2 [0.08, 10.56]
Delta	12 [0.34, 45.26]	14 [0.4, 52.8]	7 [0.2, 26.4]	2 [0.06, 7.54]

Release margins [0, 20]

- Is it safe to release this conditional?

NO, only 1 table!

IP=Integer Programming

LP=Linear Programming

What we know: Conditionals

- Let X and Y be two random variables and $O = \{o_{ij}\}$ be the $I \times J$ table of observed counts with sample size N
- Let $P = \{p_{ij}\}$, $i = 1, ..., I$, $j = 1, ..., J$, where $p_{ij} = P(X = i, Y = j)$ and $\sum\sum p_{ij} = 1$
- Let $D = \{d_{ij}\}$, $i = 1, ..., I$, $j = 1, ..., J$, where $d_{ij} = P(Y = j \mid X = i) = p_{ij} / p_{i+}$.
- Note that these probability distributions involve true parameters. Let observed conditionals be: $\hat{d}_{ij} = o_{ij} / o_{i+}$

Consider an $I \times J$ table

Min **cn**

s.t. **An = b**

Gn ≥ h

n ≥ 0

c is a row vector of lenght d

n is a column vector of lenght d

A is a $m \times d$ matrix

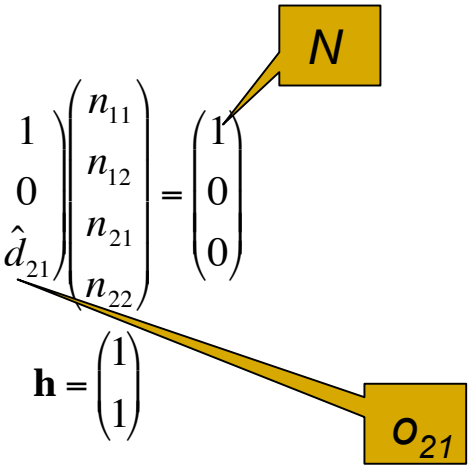
b is a column vector of lenght m

Consider a 2x2 table

$$\mathbf{c} = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ -\hat{d}_{12} & \hat{d}_{11} & 0 & 0 \\ 0 & 0 & -\hat{d}_{22} & \hat{d}_{21} \end{pmatrix} \begin{pmatrix} n_{11} \\ n_{12} \\ n_{21} \\ n_{22} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

$$\mathbf{G} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \quad \mathbf{h} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$



Potentially different **A**: agency vs. outsider

What we know: Conditionals

- Optimization: Linear & Integer programming

- Slavkovic (2004), Slavkovic and Feinberg (2004) first looked at bounds induced by conditional probabilities in the context of SDL.

- Smucker & Slavkovic (2007)
  - Distinction between the agency and an outsider
  - New sharp bounds on cells
  - Derived closed-form solutions for the linear relaxation bounds
    - Result for *IxJ* tables: $\hat{d}_{ij} \leq n_{ij} \leq (N-(I-1))\hat{d}_{ij}$
    - Extension for a full conditional in a *k*-way table
    - Extension for a small conditional in *k*-way table
  - Show empirically that the linear relaxation bounds can be far wider than the corresponding exact bounds
  - LP bounds depend on rounded conditionals!

What we know: Conditionals

- Conditional Inference: Sampling with Markov bases

- Perform a random walk on the space of all the tables with a given set of conditionals
  - Moves (Markov bases) are integer valued vectors in the kernel(A) that, added to the current table, will produce a table with same conditionals.

- Using Markov Bases, it is possible to build a Gibbs sampler 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 χ^2).
  - MCMC not as nice as for the margins
  - Sample from $P(\mathbf{n} \mid \{\hat{d}_{ij}\}, \{p_{+j}\}, N)$
  - But this requires prior distribution either on one cell or one marginal probability
  - Lee & Slavkovic (in progress)

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).

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

- Possible margins release for well-fitted models:
  [CST][CRT][CSR] [CST][CSR][TR] [CST][CSR]

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][CSR] **65,419,200 tables** and 2. [CST][CSR][RT] **108,490 tables**

Center	Status	Treatment	Response		
			Poor	Moderate	Excellent
1	1	Active	3 [0,14]	20 [1,28]	5 [0,13]
		Placebo	11 [0,14]	14 [6,33]	8 [0,13]
	2	Active	3 [0,9]	14 [3,27]	12 [1,17]
		Placebo	6 [0,9]	13 [0,24]	5 [0,16]
2	1	Active	12 [2,21]	12 [3,22]	0 [0,0]
		Placebo	11 [2,21]	10 [0,19]	0 [0,0]
	2	Active	3 [0,9]	9 [0,16]	4 [0,7]
		Placebo	6 [0,9]	9 [2,18]	3 [0,7]

- Is it safe to release?

Software: Latte, 4ti2, lpsolve

- For the **[CST][CSR]** release there are 12 elements in the Markov Basis.

Center	Status	Treatment	Response		
			Poor	Moderate	Excellent
1	1	Active	2	21	5
		Placebo	13	13	7
	2	Active	3	14	12
		Placebo	5	13	6
2	1	Active	12	12	0
		Placebo	11	10	0
	2	Active	3	9	4
		Placebo	6	9	3

-1	+1	0
+1	-1	0
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
-1	0	+1
0	0	0
0	0	0
0	0	0
0	0	0

Software: 4ti2, CoCoA

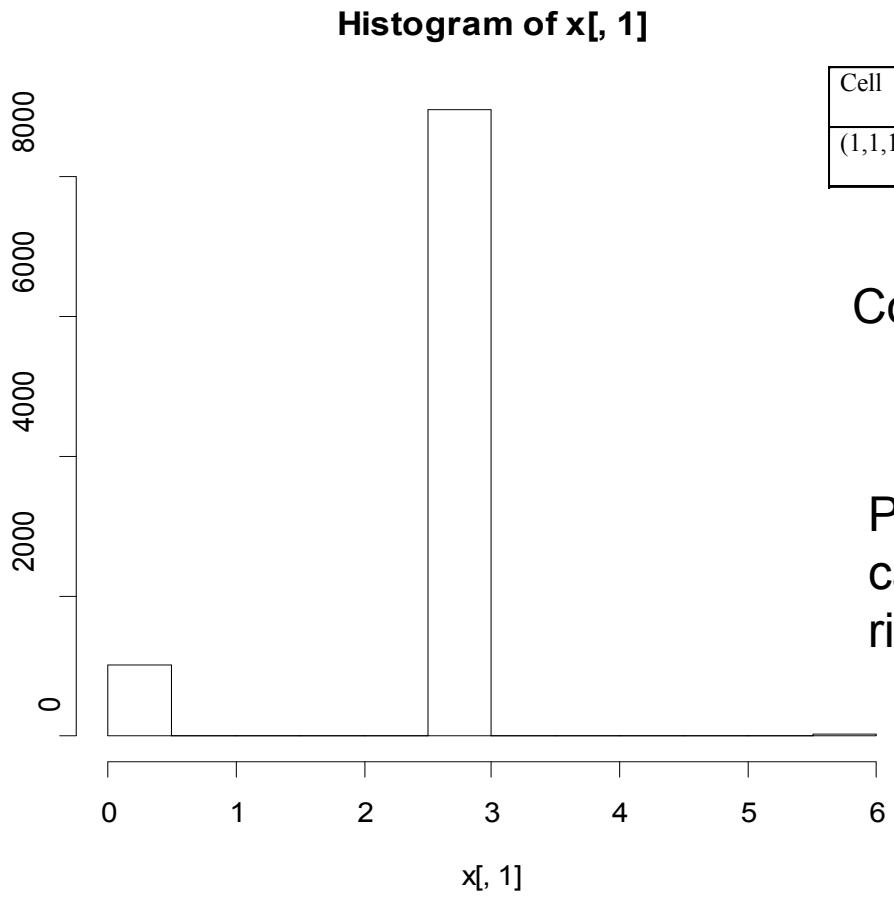
Conditional inference given the conditionals: counting & optimizing

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

Center	Status	Response / Treatment	Poor	Moderate	Excellent
1	1	Active	3 [3, 6] [1, 17.03]	20 [20, 40] [6.67,113.55]	5 [5, 10] [1.67, 28.39]
1	1	Placebo	11 [1.38, 51.26]	14 [1.75, 65.23]	8 [1, 37.28]
1	2	Active	3 [1, 16.48]	14 [4.67, 76.91]	12 [4, 65.92]
1	2	Placebo	6 [6, 12] [1.2, 38.61]	13 [13, 26] [2.60, 83.66]	5 [5, 10] [1 32.18]
2	1	Active	12 [1, 18] [1.10, 79.44]	12 [1, 18] [1, 72.26]	0
2	1	Placebo	11 [1.10, 79.48]	10 [1, 72.26]	0
2	2	Active	3 [3, 9] [1, 29.06]	9 [9, 27] [3, 87.17]	4 [4, 12] [1, 38.74]
2	2	Placebo	6 [2, 12] [2, 51.89]	9 [3, 18] [3, 77.83]	3 [1, 6] [1, 25.94]

- There are 7,703,002 tables
- These are LP relaxation bounds (plus constraint for cells to be greater than 1), but IP are much sharper
- Is it safe to release this conditional?

Bounds from the posterior distribution given [R|CST]



Cell	True Value	Lower Bound	Upper Bound	Mean	Median	Std. Dev.
(1,1,1,1)	3	0	6	2.7009	3	0.9176728

Compared to LP: [1, 17.03]

Presence of integer gaps which can strongly influence the disclosure risk and utility.

What we know: Marginals and Conditionals

- Small conditionals give the same sharp bounds as the corresponding margin, but the LP bounds are wider and the space of tables is larger

- Conditionals preserve odds and odds-ratios thus carry a lot of utility for inference, e.g., for log-linear models

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

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

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

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

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

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

- Certain combinations of marginals and conditionals will uniquely identify the table
  - E.g. trivial: P(R|CST) and P(CST)

Uniqueness: Complete specification of the joint

- *Theorem*: $\mathcal{T}=\{p_{A|B}, p_A\}$, A, B $\in$K uniquely specify p_{AB} if the k-way array $p_{A|B}$ has a full rank, and $d_A \geq d_B$.
  - Given *P(x|y) and P(x),* unique solution exists for *IxJ, if matrix with values P(x|y) has a full rank and I≥ J*
  - [R|T] and [R] give [RT]
- There are further simplifications
  - E.g., closed form unique solution for I x 2

Slavkovic (2005), and Dobra et al. (2008)

Bounds given [R|CS] vs. [CSR]

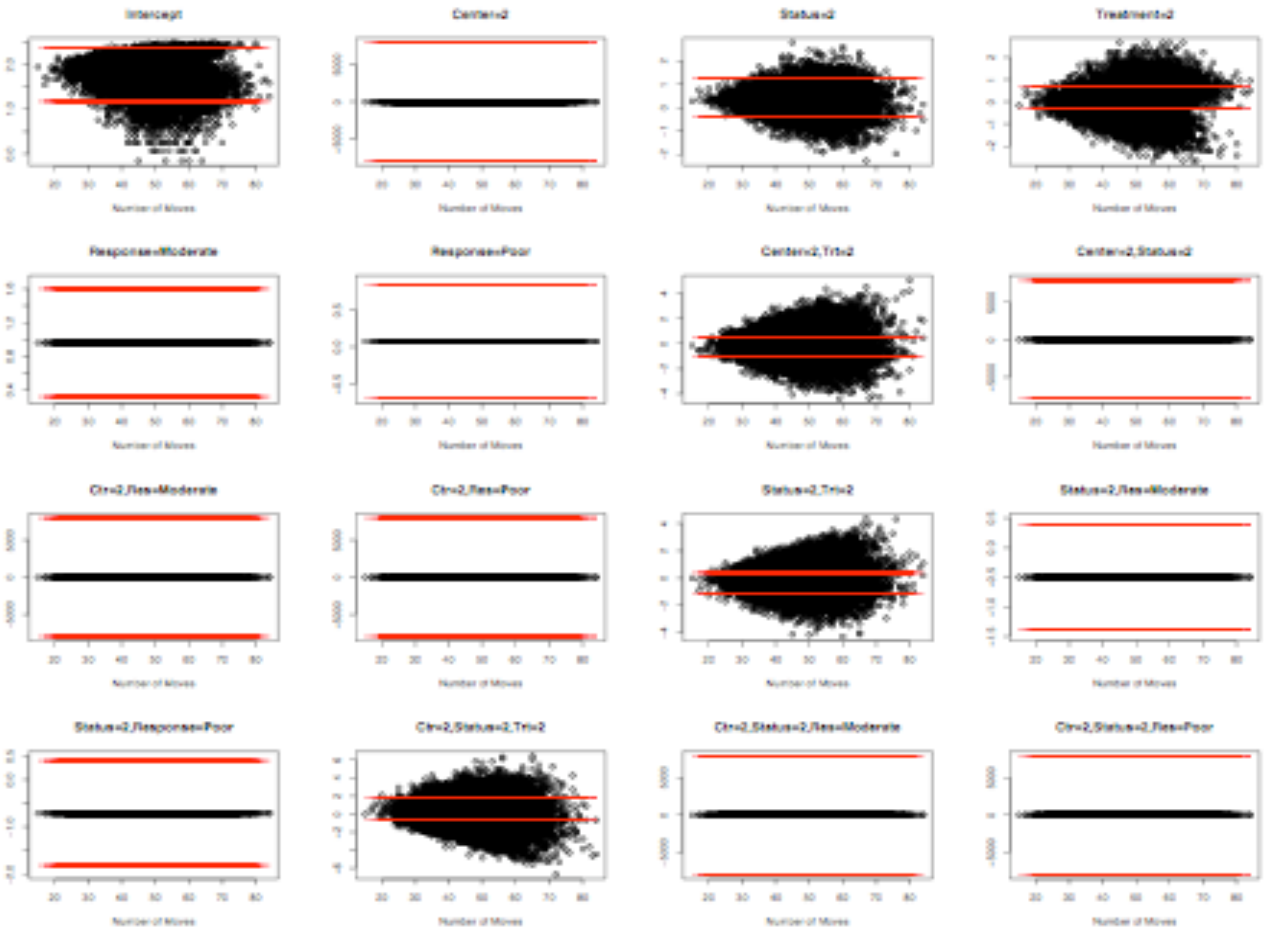
- Number (tables|conditional) ≥ Number (tables|corresponding margin)
  - 31,081,579,235,840 vs. 31,081,397,760,000

- Markov basis for [R|CS] includes elements from [CRS] plus

$$\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\
-1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-14 & -34 & -13 & 0 & 0 & 0 & -9 & -27 & -17 & 0 & 0 & 0 & -46 & -44 & 0 & 0 & 0 & 0 & 54 & 108 & 42 & 0 & 0 & 0 \\
-14 & -34 & -13 & 0 & 0 & 0 & 18 & 54 & 34 & 0 & 0 & 0 & -23 & -22 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-42 & -102 & -39 & 0 & 0 & 0 & -9 & -27 & -17 & 0 & 0 & 0 & 138 & 132 & 0 & 0 & 0 & 0 & -9 & -18 & -7 & 0 & 0 & 0 \\
-56 & -136 & -52 & 0 & 0 & 0 & 9 & 27 & 17 & 0 & 0 & 0 & 115 & 110 & 0 & 0 & 0 & 0 & -9 & -18 & -7 & 0 & 0 & 0 \\
-70 & -170 & -65 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 69 & 66 & 0 & 0 & 0 & 0 & 45 & 90 & 35 & 0 & 0 & 0
\end{pmatrix}$$

- LP relaxation bounds wider than for the margin
- But, sharp bounds are the same!
  - Last cell [3], [0, 39.74], [0,7]

Log-linear model [CSR][CST]: parameter estimates



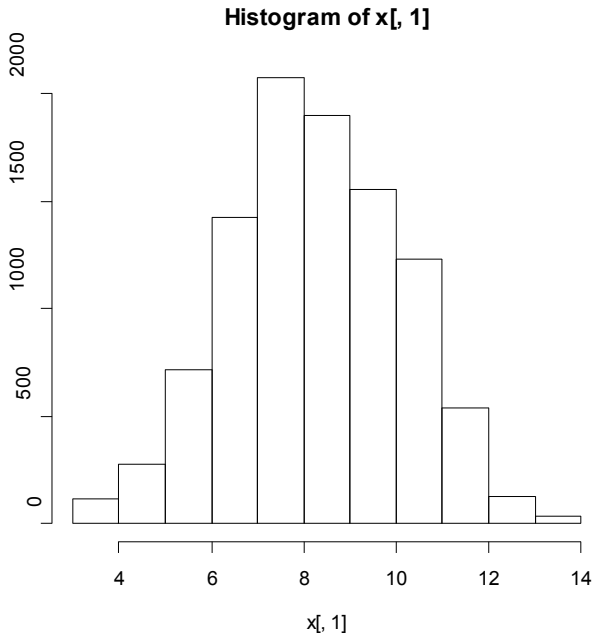
Simulate data, that is 4-way tables based on given [R|CS] and N.

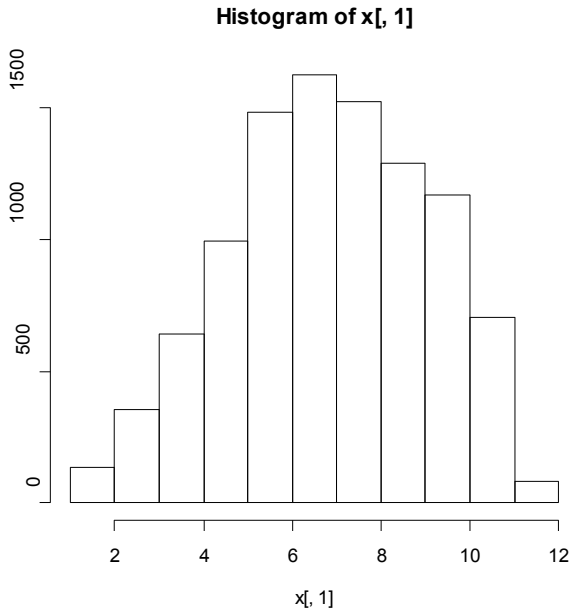
Fit a log-linear model [CSR][CST].

Red horizontal lines are 95% CIs from the original data.

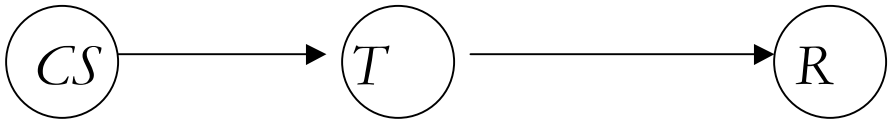
Posterior distributions given (CRS,T) vs (R|CS,T)

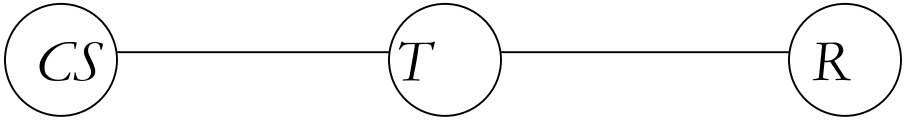
Cell	True Value	Lower Bound	Upper Bound	Mean	Median	Std. Dev.
(1,1,1,1)	3	3 1	14 12	8.7557 7.3366	9 7	1.8847 2.2334





Bounds on Multi-way Tables Using DAGs

- *Query: P(R|T), P(T|CS)*
  - *G:* 

  - $G^u = G^m$: 

- *Theorem*: When *G* satisfies *Wermuth* condition, the bounds imposed by a set of conditionals and marginals reduce to the bounds imposed by a set of marginals associated with G^u
- *Bounds:* $max\{0, p_{RT} + p_{TCS} - p_T\} \leq p_{RCST} \leq min\{p_{RT}, p_{TCS}\}$

DAG=Directed acyclic graph

Slavkovic (2004)

Non-existence of MLE

- Maximum Likelihood Estimates (MLE's) of the cell mean vector play a fundamental role in assessment of fit and model selection.
  - However, if the table is big and sparse, it may not exist
  - Is quite common and occurs even in non-sparse tables. Models like: **[CST][CSR][CTR]**, **[CST][CSR][TR]**, **[CST[CSR]** don't have a MLE. All existing software would ignore it and **report the wrong number of degrees of freedom**.

- When releasing margins associated to a non-existent MLE:
  - Some parameters cannot be estimated (with the MLE)
  - Increased risk of disclosure: some cells can be known **for certain** to be zero. This, for example, can affect the way bounds of cell entries are obtained.

- Algorithms to identify such zeros are currently being developed (Rinaldo 2005).

Practical Implications

- 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
  - In most realistic scenarios, because of lack of numerical precision, the integer program cannot be solved using the released conditional probabilities, though the agency could check these bounds using the actual data
    - sometimes unique specification results
  - Reveal zero counts

- Number of simplifications for quick assessment of risk, that is bounds

- Algebraic geometry useful for exploring the space of tables for smaller problems
  - Not computationally feasible for large tables
  - Works for margins & rates
  - Size of the move may determine uniqueness
  - Number of tables as a measure for disclosure evaluation
    - Space of tables too small & may reveal margins
  - Computing sharp bounds
  - Implication for distributions
  - Related to synthetic data methods and swapping

Open & ongoing questions

- Exploring further combinations of marginals and conditionals
- Exploring applications and extensions to magnitude tables

- Bounds & space of tables given odds, and odds-ratios
  - Requires non-linear programming

- When do combinations of margins and conditionals reduce to margins?
  - Wermuth condition!

- Definitions of utility and risk, and disclosure
  - Exploring utility and risk measures

- The usual hard problems remain hard
  - Modeling the joint distribution of multivariate categorical data (especially in presence of sparse data)

Acknowledgements

- NSF grant SES-0532407

- Department of Statistics, Penn State University
  - Juyoun Lee, Byran Smucker

- References:
  - http://www.stat.psu.edu/~sesa/research.html
  - Statistical Policy Working Paper 22 (2005) by Federal Committee on Statistical Methodology: http://www.fcsm.gov