Tabular Data: Releases of Conditionals and Marginals

Aleksandra Slavkovic-Gross
Department of Statistics
Carnegie Mellon University

Data Confidentiality Day, The BLS, November 20, 2003

Goal of Statistical Disclosure Limitation

- Preserving confidentiality

- Providing access to useful statistical data, not just few numbers
  - Inferences should be the same as if we had original complete data
    - Requires ability to reverse disclosure protection mechanism, not for individual identification, but for inferences about parameters in statistical models (e.g, likelihood function for disclosure procedure)
  - Sufficient variables to allow for proper multivariate analyses
  - Ability to assess goodness of fit of models
    - Need most summary information, residuals, etc

Goal of Statistical Disclosure Limitation

- 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

Wealth (W)		Weak			Strong	
Location (L)		Center	Outskirts		Center	Outskirts
Gender (G)	Male	**8**	**6**		**2**	**9**
	Female	**0**	**3**		**5**	**1**

  - Risk-Utility (R-U) confidentiality maps (Duncan et al. (2001))
  - Bayesian framework (Trottini (2001), Trottini & Fienberg (2002))

Survey of self-employed shop-owners
Source: Willenborg & deWaal, adapted example

NISS DG & Statistical Disclosure Methods

- Partial data releases for tabular data

- Release of *marginals*
  - Maintains existing statistical correlations
  - Determine safe releases via *bounds* and *distributions*
    - Linear/Integer programming
      - Roehrig et al. (1999), Dobra (2001)

    - Decomposable and graphical log-linear models
      - Dobra & Fienberg(2000, 2002)

    - Shuttle Algorithm
      - Dobra (2002)

    - Gröbner (& Markov) bases to enumerate or sample
      - Diaconis & Sturmfels (1998), Dobra & Fienberg (2000, 2002), Dobra et al. (2003)

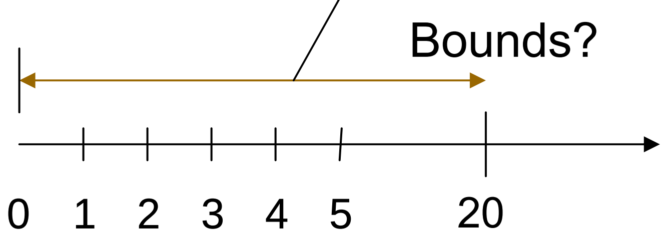
- Release of *conditionals* (A. Slavkovic)

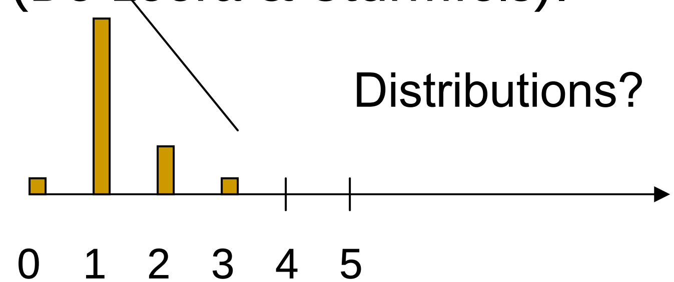
- Release of *regressions* (Jerry Reiter)

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).





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

NCHS: National Health Interview Survey, 2000

Table 21. Percent distributions (with standard errors) of any period without health insurance coverage during the past 12 months and percents (with standard errors) of persons who were without coverage for 6 months or less or 7–12 months, among currently insured persons under age 65 years, by selected characteristics: United States, 2000

Selected characteristic	Total	Any period without coverage: No	Any period without coverage: Yes[1]	Duration of period without coverage[2]: 6 months or less	Duration of period without coverage[2]: 7–12 months
	Percent distribution[3] (standard error)			Percent[4] (standard error)	
Total[5]	100.0	94.8 (0.15)	5.2 (0.15)	3.4 (0.12)	1.7 (0.08)
Sex					
Male	100.0	95.0 (0.17)	5.0 (0.17)	3.3 (0.14)	1.6 (0.09)
Female	100.0	94.7 (0.17)	5.3 (0.17)	3.4 (0.13)	1.8 (0.09)
Age					
Under 12 years	100.0	94.8 (0.27)	5.2 (0.27)	3.5 (0.22)	1.7 (0.16)
12–17 years	100.0	95.6 (0.32)	4.4 (0.32)	2.7 (0.25)	1.7 (0.20)
18–44 years	100.0	93.0 (0.21)	7.0 (0.21)	4.7 (0.17)	2.3 (0.11)
45–64 years	100.0	97.5 (0.13)	2.5 (0.13)	1.6 (0.11)	0.8 (0.08)
Race					
1 race[6]	100.0	94.9 (0.15)	5.1 (0.15)	3.4 (0.12)	1.7 (0.08)
White	100.0	95.0 (0.16)	5.0 (0.16)	3.3 (0.13)	1.6 (0.09)
Black or African American	100.0	94.3 (0.35)	5.7 (0.35)	3.6 (0.28)	2.0 (0.21)
American Indian or Alaska Native	100.0	93.8 (2.11)	*6.2 (2.11)	*4.5 (2.02)	*1.4 (0.59)
Asian	100.0	95.9 (0.70)	4.1 (0.70)	2.7 (0.61)	1.3 (0.37)
Native Hawaiian or Other Pacific Islander	100.0	100 (0.00)	*–	*–	*–
2 or more races[7]	100.0	92.2 (1.00)	7.8 (1.00)	5.6 (0.82)	2.3 (0.57)
Black or African American, white	100.0	92.8 (1.66)	7.2 (1.66)	5.0 (1.45)	*2.3 (0.95)
American Indian or Alaska Native, white	100.0	88.9 (2.61)	11.1 (2.61)	8.0 (2.33)	*3.1 (1.31)
Hispanic or Latino origin[8] and race					
Hispanic or Latino	100.0	93.3 (0.39)	6.7 (0.39)	3.7 (0.28)	2.8 (0.25)
Mexican or Mexican American	100.0	93.3 (0.50)	6.7 (0.50)	3.6 (0.34)	3.0 (0.34)
Not Hispanic or Latino	100.0	95.0 (0.15)	5.0 (0.15)	3.4 (0.13)	1.6 (0.08)
White, single race	100.0	95.1 (0.17)	4.9 (0.17)	3.3 (0.14)	1.5 (0.10)
Black or African American, single race	100.0	94.3 (0.35)	5.7 (0.35)	3.6 (0.29)	2.0 (0.21)
Education[9]					
Less than a high school diploma	100.0	93.7 (0.40)	6.3 (0.40)	3.4 (0.32)	2.7 (0.28)
High school diploma or GED[10]	100.0	95.3 (0.22)	4.7 (0.22)	2.9 (0.19)	1.7 (0.13)
Some college	100.0	95.0 (0.24)	5.0 (0.24)	3.4 (0.21)	1.5 (0.13)
Bachelor's degree or higher	100.0	96.8 (0.20)	3.2 (0.20)	2.5 (0.18)	0.7 (0.09)
Family income[11]					
Less than $20,000	100.0	89.6 (0.49)	10.4 (0.49)	5.8 (0.35)	4.4 (0.32)
$20,000 or more	100.0	95.6 (0.16)	4.4 (0.16)	3.1 (0.13)	1.3 (0.08)
$20,000–$34,999	100.0	90.3 (0.52)	9.7 (0.52)	6.3 (0.42)	3.3 (0.29)
$35,000–$54,999	100.0	93.9 (0.38)	6.1 (0.38)	4.5 (0.33)	1.5 (0.16)
$55,000–$74,999	100.0	96.1 (0.38)	3.9 (0.38)	3.0 (0.33)	0.8 (0.19)
$75,000 or more	100.0	97.9 (0.21)	2.1 (0.21)	1.5 (0.18)	0.5 (0.10)

Source: http://www.cdc.gov/nchs/data/series/sr_10/sr10_214.pdf

NCES: Parent Survey of NHES Program

Data access" http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2000079

Table 3.— Distribution of all students, homeschooled students, and nonhomeschooled students ages 5-17, with a grade equivalent of kindergarten to grade 12, by selected characteristics: 1999

Characteristic	Number of students	All students Percent	All students s.e.	Homeschoolers[1] Percent	Homeschoolers[1] s.e.	Non-homeschoolers Percent	Non-homeschoolers s.e.
Total	**50,188,000**	**100.0**		**100.0**		**100.0**	
Grade equivalent[2]							
K– 5	24,428,000	48.7	0.07	50.4	3.75	48.7	0.09
Kindergarten	3,790,000	7.6	0.04	10.8	2.31	7.5	0.05
Grades 1–3	12,692,000	25.3	0.04	23.5	3.61	25.3	0.07
Grades 4–5	7,946,000	15.8	0.02	16.0	2.34	15.8	0.05
Grades 6–8	11,788,000	23.5	0.04	21.9	2.83	23.5	0.06
Grades 9–12	13,954,000	27.8	0.10	27.7	3.21	27.8	0.11
Race/ethnicity							
White, non-Hispanic	32,474,000	64.7	0.32	75.3	3.36	64.5	0.33
Black, non-Hispanic	8,047,000	16.0	0.20	9.9	2.80	16.1	0.21
Hispanic	7,043,000	14.0	0.17	9.1	2.06	14.1	0.17
Other	2,623,000	5.2	0.23	5.8	2.01	5.2	0.23
Sex							
Female	24,673,000	49.2	0.47	51.0	3.27	49.1	0.47
Male	25,515,000	50.8	0.47	49.0	3.27	50.9	0.47
Number of children in the household							
One child	8,226,000	16.4	0.30	14.1	2.53	16.4	0.30
Two children	19,883,000	39.6	0.42	24.4	3.06	39.9	0.42
Three or more children	22,078,000	44.0	0.48	61.6	3.97	43.7	0.49

Source: "Homeschooling in the U.S.: 1999". July 2001

U.S. Census Bureau: Pennsylvania: 2000 Census

Table 1. **Place of Birth, Residence in 1995, and Language: 2000**

[Data based on a sample (except Tables 65-68). For information on confidentiality protection, coverage, sampling error, and nonsampling error, see Appendix G.
For location of definitions, see "How to Use This Census Report"]

State County County Subdivision Place	Total population	Native population—Percent born in state of residence	Foreign-born population			Population 5 years and over		Speak a language other than English at home			
								Population 5 to 17 years		Population 18 years and over	
			Number	Percent of total population	Percent naturalized citizens	Number	Percent living in different house in 1995	Number	Percent who speak English less than "very well"	Number	Percent who speak English less than "very well"
The State	**12 281 054**	**81.1**	**508 291**	**4.1**	**50.6**	**11 555 538**	**36.5**	**189 885**	**33.5**	**782 599**	**38.9**
Adams County	91 292	71.2	3 130	3.4	35.6	85 917	39.6	1 034	41.8	3 649	50.0
Abbottstown borough	884	76.5	20	2.3	55.0	813	50.9	12	-	36	36.1
Arendtsville borough	849	80.0	94	11.1	-	775	46.5	45	15.6	82	69.5
Bendersville borough	577	90.2	58	10.1	-	529	34.2	10	-	24	58.3
Berwick township	1 817	78.7	53	2.9	47.2	1 721	35.6	22	50.0	49	44.9
Biglerville borough	1 096	75.5	84	7.7	21.4	1 039	44.3	22	18.2	64	40.6
Bonneauville borough	1 378	72.0	49	3.6	6.1	1 257	39.4	9	77.8	42	69.0
Butler township	2 683	73.3	91	3.4	19.8	2 551	33.2	24	4.2	95	54.7
Carroll Valley borough	3 287	44.5	87	2.6	78.2	3 023	46.6	6	33.3	80	35.0
Conewago township	5 628	80.7	62	1.1	27.4	5 236	36.4	56	41.1	107	49.5
Midway CDP	2 362	82.6	30	1.3	33.3	2 197	32.8	13	-	56	60.7
Cumberland township	5 812	67.1	361	6.2	47.9	5 624	37.6	58	39.7	384	33.3
East Berlin borough	1 365	82.0	34	2.5	70.6	1 256	43.4	7	71.4	37	59.5
Fairfield borough	485	72.2	-	-	(X)	466	37.1	1	-	-	(X)
Franklin township	4 589	74.9	106	2.3	49.1	4 364	37.4	83	24.1	129	58.9

Source: http://www.census.gov/prod/cen2000/phc-2-40-pt1.pdf

BLS: Data from 2002 CPS supplement

Table 2. Volunteer rates by sex, race, Hispanic origin, and selected characteristics, September 2002

Selected characteristics	White			Black			Hispanic		
	Total	Men	Women	Total	Men	Women	Total	Men	Women
Age									
Total, 16 years and over	29.4	25.1	33.4	19.2	16.7	21.1	15.7	12.9	18.4
16 to 19 years	28.6	24.3	33.0	18.8	16.3	21.1	18.1	15.3	20.9
20 to 24 years	19.3	15.7	22.9	13.1	9.9	15.8	9.4	7.6	11.3
25 to 34 years	26.8	20.8	32.7	20.2	15.6	24.0	16.9	12.9	21.0
35 to 44 years	37.1	30.8	43.4	22.4	19.1	25.2	20.6	15.7	25.4
45 to 54 years	33.5	29.4	37.6	20.4	19.3	21.2	16.1	15.1	17.1
55 to 64 years	28.8	26.1	31.4	20.6	19.1	21.7	13.2	12.0	14.2
65 years and over	23.9	22.2	25.2	13.9	14.9	13.3	6.9	6.2	7.4
Employment status among persons aged 16 years and over									
Employed	31.4	27.1	36.6	21.9	18.9	24.6	17.0	14.0	21.1
Unemployed	26.5	21.3	32.6	21.5	18.2	24.6	17.9	12.3	25.5
Not in the labor force	25.6	20.1	28.9	14.1	12.1	15.5	12.6	9.0	14.4
School enrollment status among persons aged 16 to 24 years									
Enrolled in high school	32.3	26.0	39.5	18.2	17.0	19.4	19.6	15.6	23.9
Enrolled in college	28.3	25.2	31.1	23.9	19.7	26.4	19.6	19.4	19.7
Not enrolled in school	16.0	13.0	19.1	10.5	8.4	12.7	8.6	7.2	10.3
Educational attainment among persons aged 25 years and over									
Less than a high school diploma	10.5	9.0	11.8	9.2	8.6	9.6	8.4	5.8	11.0
High school graduate, no college[1]	22.8	18.2	26.8	14.1	12.7	15.4	16.3	13.4	19.2
Less than a bachelor's degree[2]	34.5	28.9	39.4	26.1	23.2	28.0	25.2	22.9	27.2
College graduate	46.0	40.9	51.4	36.6	33.4	39.1	31.9	27.2	36.4

[1] Includes high school diploma or equivalent.

[2] Includes the categories of some college, no degree; and associate's degree.

NOTE: Data on volunteers relate to persons who performed unpaid volunteer activities for an organization at any point from September 1, 2001, through the survey week in September 2002. Details for the above race and Hispanic-origin groups will not sum to totals because data for the "other races" group are not presented and Hispanics are included in both the white and black population groups.

Source: Boraas, S. "Volunteerism in the U.S." Monthly Labor Review, August 2003

Why conditionals? – Causal Inference

Wealth (W)		Weak		Strong	
Location (L)		Center	Out	Center	Out
Gender (G)	Male	**8 [24,0] [20,0] [8,5]**	**6 [12,0] [10,0] [9,6]**	**2 [6,0] [5,0] [5,2]**	**9 [18,0] [15,0] [9,6]**
	Female	**0 [0,0] [0,0] [3,0]**	**3 [24,0] [6,0] [3,0]**	**5 [34,0] [9,1] [5,2]**	**1 [8,0] [2,0] [4,1]**

Survey of self-employed shop-owners
Source: Willenborg & deWaal, adapted example

f(w|l,g)

f(w|l,g), f(g)

f(w,l)
f(w,g)
f(l,g)

- Assess causal distribution: $P(W=w|L:=l)=\sum_g P(w|l,g)P(g)$
- Assess treatment effect: $P(W=1|L=1, G=1)-P(W=1|L=0, G=0)$
- Release f(w,l,g) vs. f(w|l,g), f(g)

New research question

- Determine safe releases in terms of arbitrary sets of marginal and conditionals
  - Assume data reported without error
  - Assume compatible margins and conditionals
  - Assume unweighted counts

  - Investigate conditions under which sets of marginals and conditionals give:
    - unique specifications
    - upper/lower bounds on cell entries
    - distributions over the cell entries

  - Determine/compute the bounds and distributions

Uniqueness: Complete specification of the joint

- Full disclosure
- Ability to completely reconstruct the original table

- Uniqueness Theorem for *k*-way tables
  - Gelman & Speed 1993, Arnold et al. 1999
  - Example: Two-way table
    - *P(x), P(y|x)*
    - *P(x|y), P(y|x)*
  - Arnold et al. 1999
    - Sometimes: *P(y), P(y|x)*
    - Define the missing marginal
    - Vardi & Lee (1993) algorithm

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

Uniqueness: Complete specification of the joint

- *Prop*: Given *P(x|y) and P(x),* unique solution exists for *IxJ, if matrix with values P(x|y) has full rank and I≥ J*

- *Prop*: Unique solution for *I x 2*

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

- 2 x 2 table, release *P(x|y), P(x)*

X|Y	1	1
	a_{11}	a_{12}
	a_{21}	a_{22}

	Y		
X	p_{11}	p_{12}	p_{1+}
	p_{21}	p_{22}	p_{2+}

$$p_{11} = a_{11} \frac{p_{1+} - a_{12}}{a_{11} - a_{12}}$$

I x J Tables Summary

Queries	Assume	Unique	Assume	Bounds
P(x|y), P(y)		√		
P(x|y), P(y|x)		√		
P(x), P(y)	$X \perp\!\!\!\perp Y$	√	X not $\perp\!\!\!\perp Y$	$\max\{0, x_i+y_j-n\} \leq x_{ij} \leq \min\{x_i , y_j\}$
P(x|y), P(x)	$I \geq J$	√	$I < J$	
P(x)				$0 \leq x_{ij} \leq x_i$
P(x|y)				$0 \leq p_{ij} \leq p_{i\|j}$

Understanding 2-way tables is important for solving k-way tables

LP/IP: Bounds given $P(x|y)=\{a_{ij}\}$ or $P(y|x)=\{b_{ij}\}$

- Conditionals maintain odds-ratio which makes this problem different from marginals:

$$\alpha = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{b_{11}b_{22}}{b_{12}b_{21}} = \frac{a_{11}a_{22}}{a_{12}a_{21}}$$

- *N* unknown
  - LP bounds: $0 \le p_{ij} \le a_{ij}$ or $0 \le p_{ij} \le b_{ij}$
  - Not sharp for integer tables
  - Closed form solutions for 2 x J tables

- *N* known
  - IP gives sharp bounds when feasible
  - May not be computationally feasible for *k-way* tables
  - LP relaxation gives fractional bounds

LP = Linear Programming
IP = Integer Programming

LP: Bounds given $P(y|x)$

Download (Y) / Gender (X)	Yes	No
Male	*15 (0.3)*	*10 (0.1)*
Female	*5 (0.2)*	*20 (0.4)*

- Release *P(y|x)*

Y|X		
1	0.6	0.4
1	0.2	0.8

Bounds, unknown N

X,Y		
	[0,0.6]	[0, 0.4]
	[0,0.2]	[0,0.8]

Bounds, known N

X,Y		
	[0,30]	[0, 20]
	[0,10]	[0,40]

- Problems:
  - α=6
  - None of the conditional values are zero
    $\Rightarrow$ Cell in the original table CANNOT be zero.
    $\Rightarrow$ These are NOT the sharp bounds for integer tables

Fictitious example: survey of 50 students about illegally downloading MP3s

IP: Bounds given *P(gender | download)*, known *N*=50

- Explicitly forcing the lower bound to be 1

  Max n_{11}

  Subject to $n_{11}+n_{12}+n_{21}+n_{22}=50$

  $0.4\, n_{11} - 0.6\, n_{12} = 0,$

  $0.8\, n_{21} - 0.2\, n_{22} = 0,$

  $n_{ij} \geq 1$, $i=1,2$, $j=1,2$

X,Y		
	15 [3, 27]	*10 [2, 18]*
	5 [1, 9]	*20 [4, 36]*

- LP Relaxation

X,Y		
	15 [3, 27]	*10 [2, 18]*
	5 [1, 9]	*20 [4, 36]*

IP=Integer Programming

IP: Bounds given *P(download|gender)*, known *N*=50

- Release:

X|Y	1	1
	0.75	*0.33*
	0.25	*0.67*

- IP: no feasible solution

- LP Relaxation

X,Y		
	15 [3, 35.25]	*10 [1, 15.33]*
	5 [1, 11.74]	*20 [2, 30.67]*

- These are not tight bounds!

IP=Integer Programming

Delinquent Children by County & Education Level

- Release observed conditional frequencies

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

- IP: no feasible solution
- LP relaxation bounds:

County	Low	Medium	High	Very High
Alpha	15 [15, 74.6]	1 [1, 4.97]	3 [3, 14.9]	1 [1, 4.97]
Beta	20 [1.99, 30.8]	10 [1, 15.5]	10 [1, 15.5]	15 [1.5, 23.2]
Gamma	3 [1.5, 11.0]	10 [5, 36.8]	10 [5, 36.8]	2 [1, 7.36]
Delta	12 [6.02, 33.27]	14 [7.02, 38.8]	7 [3.51, 19.4]	2 [1, 5.53]

Release margins [0, 20]

- Is it safe to release this conditional?

LP: Bounds given P(x|y), P(x), if I<J

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

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

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

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

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

- Generalizes to 2 x J tables

LP=Linear Programming

Multi-way Tables: 3x3x2

				Gender	
				Male	Female
Race	White	Income Level	≤ \$10K	96	186
			\$10K and ≤ \$25K	72	127
			> \$25K	161	51
	Black	Income Level	≤ \$10K	10	11
			> \$10K and ≤ \$25K	7	7
			> \$25K	6	3
	Chinese	Income Level	≤ \$10K	1	0
			>10K and ≤ \$25K	1	1
			> \$25K	2	0

- Release full conditional *Income | Gender, Race*
- There are 2,083,240,054,713 tables
- Not safe to release!

Source: 1990 Census Public Use Microdata File (Fienberg, Markov, Steel (1998))

Bounds on multi-way tables using DAGs

- *Query:*
  $P(x_2{=}income \mid x_3{=}race)$, $P(x_3{=}race \mid x_1{=}gender)$, $P(x_1{=}gender)$

  - *G*: $X_1 \rightarrow X_3 \rightarrow X_2$
  - *$G^u{=}G^m$:* $X_1 - X_3 - X_2$

- *Prop*: 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:* $\max\{0, p_{13} + p_{23} - p_3\} \leq p_{123} \leq \min\{p_{13}, p_{23}\}$

DAG=Directed acyclic graph

Bounds on three-way tables using DAGs

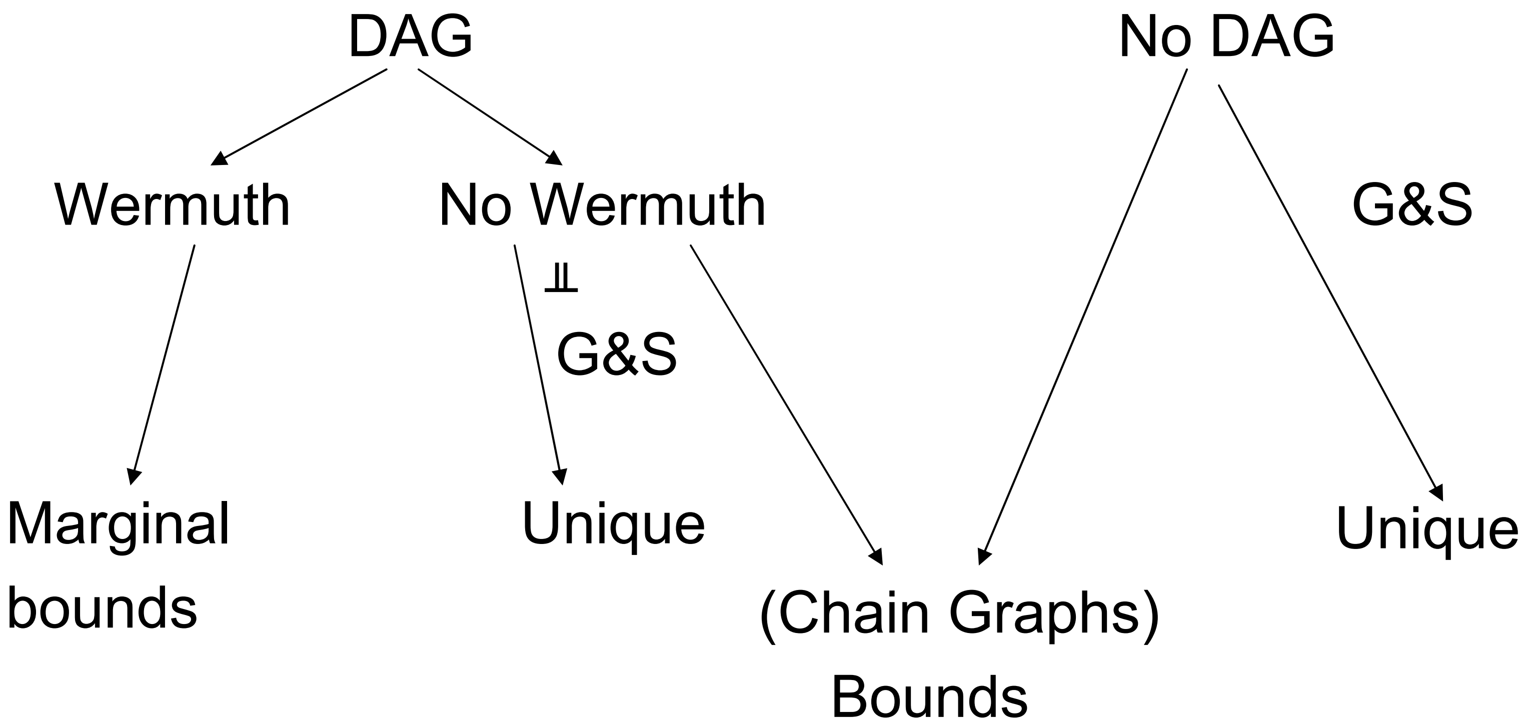
- Query:
  P$(x_3\textit{=income}|\ x_2\textit{=race}, x_1\textit{=gender}\ ),\ P(x_2\textit{=race}),\ P(x_1\textit{=gender})$

- *Wermuth fails,* $G^u \neq G^m$*:*
  - G: $X_1 \longrightarrow X_3 \longleftarrow X_2$

- If DAG implies $X_1 \perp\!\!\!\perp X_2$
  - Special case of Gelman & Speed Uniqueness Theorem
    - $P(x_3 \mid x_2, x_1),\ P(x_2, x_1)$
    - $P(x_3, x_2 \mid x_1),\ P(x_1)$
    - $P(x_3, x_1 \mid x_2),\ P(x_2)$

- What if X_i, X_j, X_k are dependent?
  - *2x2x2 table:* $f(x_i, x_j | x_k),\ f(x_i, x_j)$ gives unique specification

Framework: Bounds on Multi-way Tables



In combination with optimization methods

Computational Commutative Algebra in Statistics

- Methods from computational commutative algebra to explore the space of *all* possible tables given the constraints
  - Polynomial rings & ideals give a way of representing tables of counts
  - Markov bases are equivalent to a set of generators of an ideal (Diaconis & Sturmfels)

- Bounds & distributions given margins
  - Gröbner (Markov) bases to enumerate or sample via MCMC
  - Decomposable and reducible graphical models
    - Diaconis & Sturmfels (1998), Dobra & Fienberg (2001, 2002), Dobra & Sullivant (2002)

Markov Bases for contingency tables with fixed marginals

- *Moves* are tables with *integer* entries.
- A move leaves unchanged fixed marginals.
- *Markov bases* <u>connect</u> all tables having a set of fixed marginals.
- *Markov bases* <u>do not depend</u> on the actual value of the margins but just their functional form.

Delinquent Children by County & Education Level

- Markov bases for fixed margins
  - 36 primitive moves (e.g. $n_{41}n_{22}-n_{21}n_{42}$)

County	Low		Medium		High		Very High		Total
Alpha	15	0	1	0	3	0	1	0	20
Beta	20	-1	10	1	10	0	15	0	55
Gamma	3	0	10	0	10	0	2	0	25
Delta	12	1	14	-1	7	0	2	0	35
Total	50		35		30		20		135

18,272,363,056 possible tables given fixed row sums and column sums

Can calculate bounds and distributions

Software: 4ti2

Markov bases: Delinquent Children by County & Education Level

- Markov bases for fixed conditional *P(Education|County)*
- 4 basic moves, but more complex

-45	-3	-9	-3
0	0	0	0
3	10	10	2
12	14	7	2

$$n_{31}^{3}n_{32}^{10}n_{33}^{10}n_{34}^{2}n_{41}^{12}n_{42}^{14}n_{43}^{7}n_{44}^{2} - n_{11}^{45}n_{12}^{3}n_{13}^{9}n_{14}^{3}$$

-15	-1	-3	-1
20	10	10	15
0	0	0	0
-12	-14	-7	-2

-15	-1	-3	-1
0	0	0	0
-6	-20	-20	-4
24	28	14	4

-30	-2	-6	-2
0	0	0	0
9	30	30	6
-12	-14	-7	-2

Software: 4ti2

Enumeration: Delinquent Children by County & Education Level

- Only 1 possible integer table given the fixed conditional !
  - Higher disclosure risk than in the case of LP/IP bounds

- Compare to 18,272,363,056 possible tables given fixed row sums and column sums

County	Low	Medium	High	Very High
Alpha	15 [15, 74.6]	1 [1, 4.97]	3 [3, 14.9]	1 [1, 4.97]
Beta	20 [1.99, 30.8]	10 [1, 15.5]	10 [1, 15.5]	15 [1.5, 23.2]
Gamma	3 [1.5, 11.0]	10 [5, 36.8]	10 [5, 36.8]	2 [1, 7.36]
Delta	12 [6.02, 33.27]	14 [7.02, 38.8]	7 [3.51, 19.4]	2 [1, 5.53]

Markov bases: Fixed *P(download|gender)*

- 4 possible tables

X,Y	8	42
20	*6*	*14*
30	*2*	*28*

X,Y	20	30
35	*15*	*10*
30	*5*	*20*

X,Y	32	18
30	*24*	*6*
20	*8*	*12*

X,Y	44	6
35	*33*	*2*
15	*11*	*4*

+9	-4
+3	-8

- Issue: What's the initial table?

- Tighter bounds!

X,Y		
	15 [3, 35.25] [6, 33]	*10 [1, 15.33] [2, 14]*
	5 [1, 11.74] [2, 11]	*20 [2, 30.67] [4, 28]*

Markov moves: Fixed $P(y|x)$ & Rounding Issues

- 2x2 table [4, 10, 2, 20]=[0.3,0.2,0.1,0.4] , N=36, α=4
- 3 decimals: fixed P(y|x) = [0.286,0.714,0.091,0.909]
- 2 decimals: fixed P(y|x) = [0.27,0.71,0.09,0.91]
- 1 decimals: fixed P(y|x) = [0.3,0.7,0.1,0.9]

X,Y		
	+286	+714
	-91	-909

X,Y		
	+29	+ 71
	- 9	- 91

X,Y		
	+3	+ 7
	- 1	- 9

X,Y		
	7	17
	1	11

α=4.53

X,Y		
	1	3
	3	29

α=3.22

- Issue 1:
  - Need integer values from floating point approximations
  - Margins may be revealed!
- Issue 2:
  - Do we have a unique solution?
  - Do we accept approximation?

0.285714, 0.71426, 0. 09090, 0.90909

Markov bases for fixed conditionals & confidentiality

- The moves must maintain odds ratio, α, but do not need sample size N
- *Space of tables* is determined by knowing the bases <u>and</u> sample size N
- Depend on the value of the conditional distribution
  - Rounding to different decimal place gives different moves
  - Margins may be revealed as the denominators, and often give the unique solution for two-way tables!
  - Size of the move determines uniqueness
- Not feasible to release conditionals in two-way tables
  - Space of tables too small
  - Margins may be revealed

Example: 3x3x2 1990 Census Data

- Release all three 2-way margins:
  - *Race x Income*, *Race x Gender*, *Gender x Income*
- Release, instead, three conditionals:
  - *Income|Race*, *Race|Gender*, *Income|Gender*
- There are 441 tables and the bounds are the same

Bounds Given All Two-Way Margins or Corresponding Conditionals

				Gender		Gender	
				Male	Female	Male	Female
Race	White	Income Level	≤ \$10K	96	186	[85,107]	[175,197]
			\$10K and ≤ \$25K	72	127	[64,79]	[120,135]
			> \$25K	161	51	[158,168]	[44,54]
	Black	Income Level	≤ \$10K	10	11	[0,21]	[0,21]
			> \$10K and ≤ \$25K	7	7	[0,14]	[0,14]
			> \$25K	6	3	[0,9]	[0,9]
	Chinese	Income Level	≤ \$10K	1	0	[0,1]	[0,1]
			>10K and ≤ \$25K	1	1	[1,2]	[0,1]
			> \$25K	2	0	[1,2]	[0,1]

Czech Autoworkers example

- Risk factors for heart disease
- 2^6 table
- population data
  - "0" cell
  - population unique, "1"
  - 2 cells with "2"

- Suppose we release margins: [ADE][ABCE][BF]
  - Decomposable graph.

				B	no		yes	
F	E	D	C	A	no	yes	no	yes
neg	< 3	< 140	no		44	40	112	67
			yes		129	145	12	23
		≥ 140	no		35	12	80	33
			yes		109	67	7	9
	≥ 3	< 140	no		23	32	70	66
			yes		50	80	7	13
		≥ 140	no		24	25	73	57
			yes		51	63	7	16
pos	< 3	< 140	no		5	7	21	9
			yes		9	17	1	4
		≥ 140	no		4	3	11	8
			yes		14	17	5	2
	≥ 3	< 140	no		7	3	14	14
			yes		9	16	2	3
		≥ 140	no		4	0	13	11
			yes		5	14	4	4

Bounds Given Margins

				B	no		yes	
F	E	D	C	A	no	yes	no	yes
neg	< 3	< 140	no		[0,88]	[0,62]	[0,224]	[0,117]
			yes		[0,261]	[0,246]	[0,25]	[0,38]
		≥ 140	no		[0,88]	[0,62]	[0,224]	[0,117]
			yes		[0,261]	[0,151]	[0,25]	[0,38]
	≥ 3	< 140	no		[0,58]	[0,60]	[0,170]	[0,148]
			yes		[0,115]	[0,173]	[0,20]	[0,36]
		≥ 140	no		[0,58]	[0,60]	[0,170]	[0,148]
			yes		[0,115]	[0,173]	[0,20]	[0,36]
pos	< 3	< 140	no		[0,88]	[0,62]	[0,126]	[0,117]
			yes		[0,134]	[0,134]	[0,25]	[0,38]
		≥ 140	no		[0,88]	[0,62]	[0,126]	[0,117]
			yes		[0,134]	[0,134]	[0,25]	[0,38]
	≥ 3	< 140	no		[0,58]	[0,60]	[0,126]	[0,126]
			yes		[0,115]	[0,134]	[0,20]	[0,36]
		≥ 140	no		[0,58]	[0,60]	[0,126]	[0,126]
			yes		[0,115]	[0,134]	[0,20]	[0,36]

“Safe” to release these margins; low risk of disclosure.

Bounds given fixed conditionals [E|A,D], [B|F], [AD|BC]

- LP relaxation bounds wider than for margins

			B	no		yes	
F	E	D	C / A	no	yes	no	yes
neg	< 3	< 140	no	44	40	112	67
			yes	129	145	12	23
		≥ 140	no	35	12	80	33
			yes	109	67	7	9
	≥ 3	< 140	no	23	32	70	66
			yes	50	80	7	13
		≥ 140	no	24	25	73	57
			yes	51	63	7	16
pos	< 3	< 140	no	5	7	21	9
			yes	9	17	$\boxed{1}$	4
		≥ 140	no	4	3	11	8
			yes	14	17	5	$\boxed{2}$
	≥ 3	< 140	no	7	3	14	14
			yes	9	16	$\boxed{2}$	3
		≥ 140	no	4	0	13	11
			yes	5	14	4	4

[0, 20]
[1, 195.38]

[0, 38]
[1, 164.41]

[0, 25]
[1, 134.49]

Number(tables|conditional) ≥ Number (tables|corresponding margin)

"Safe" to release these conditionals

Summary: Tabular data releases

- Agencies already release conditionals in 2-way and 3-way tables
- Conditionals reveal zero counts
- 2-way tables
  - Do not release conditionals!
- K-way table
  - Releasing full conditionals could be too risky
  - Small conditionals may release less information (less disclosure) than corresponding marginals
- Graphical models useful for calculating bounds
- Algebraic geometry useful for exploring the space of tables
- Beginning to understand implications on rounding on disclosure limitation

Ongoing Research & Open Questions

- Investigate combining compatible pieces of information (e.g. odds ratios, margins, conditionals, regressions, etc…)
    - Implications on data usability, reconstructing data and disclosure risk

- Distribution functions over the space of tables
    - Log-linear models for marginals
    - No analogy theory model for conditionals

- Understand influence of weighted counts on disclosure

- Understand influence of rounding on disclosure

Public Access & Unique identification

- Publicly available data
  - American Fact Finder website (Source: U.S. Census Bureau: Block data)

RACE	All ages Number	18 years and over Number
Total population	**83**	**70**
White	70	63
Black or African American	1	1
American Indian and Alaska Native	0	0
Asian	9	6
Native Hawaiian and Other Pacific Islander	0	0
Two or more races	3	0

Data measured with error & reported after data swapping

- Uniqueness
  - Sweeney(2000): **Date of birth, gender, 5-digit ZIP**
  - Likely unique identification of **87%** U.S. population

Delinquent Children by County & Education Level

- Bounds:
  - Linear Programming/Integer Programming
  - Two-Way Fréchet Bounds

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

- Enumeration via Markov Basis (e.g. moves) & algebraic geometry

- Exact probability distribution for log-linear model given its MSS marginals:

$$\sigma(\mathbf{n}) = \frac{\prod_{i \in I} \frac{1}{n(i)!}}{\sum_{m \in S(c)} \left(\prod_{i \in I} \frac{1}{m(i)!} \right)}$$

  - MCMC using Markov basis (Diaconis-Sturmfels (1998), Fienberg, Makov, Meyer, Steele (2002)) & computational algebra

Some General Principles for Developing DL Methods

- All data are informative for intruder, including non-release or suppression.

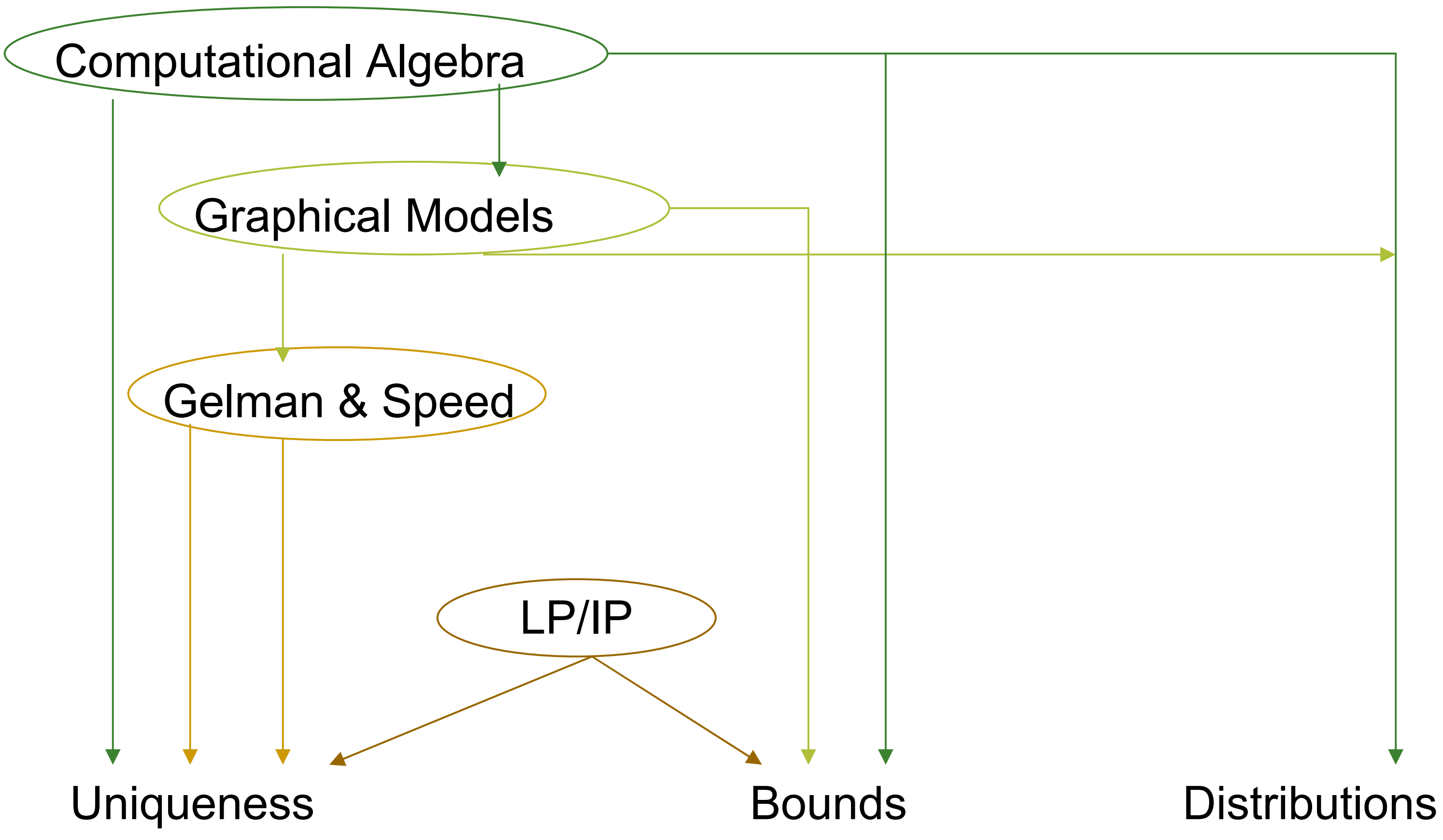
- Need to define and understand potential statistical uses of data in advance:
  - Leads to useful reportable summaries (e.g., MSSs).

- Methods should allow for reversibility for inference purposes:
  - Missing data should be "ignorable" for inferences.
  - Assessing goodness of fit is important.

Expected Contributions

- Disclosure Limitation (DL):
  - Extension of marginal query space by conditionals
  - Enhancement of data usability
- Statistics:
  - Integration of diverse results & methods from
    - Disclosure limitation
    - Conditional specification of the joint distribution
    - Graphical models
    - Algebraic geometry
  - New results on bounds and distributions on contingency tables
- New theoretical links between DL, Statistical Theory and Computational Algebraic Geometry

Framework


Computational Algebra
Graphical Models
Gelman & Speed
LP/IP
Uniqueness
Bounds
Distributions

Can we do MCMC now for fixed conditionals?

- We can find the bounds
- We can enumerate
- We have irreducible Markov chain

- What is the family of distributions that has marginals AND conditionals as MSS?
  - What's the stationary distribution?
  - What is the distribution of tables for fixed conditionals?
    - $\Pr(\boldsymbol{N}=\boldsymbol{n} | \mathcal{C}) = \Pr(\boldsymbol{N}=\boldsymbol{n} | \boldsymbol{N} \in \mathcal{T})$

Perturbation for Protection

- Perturbation preserving marginals involves a parallel set of results to those for bounds:
  - Markov basis elements for decomposable case requires only "simple" moves. (Dobra, 2002)
  - Efficient generation of Markov basis for reducible case. (Dobra and Sullivent, 2002)
  - Simplifications for 2^k tables ("binomials")??? (Aoki and Tachimura, 2003)
  - Rooted in ideas from likelihood theory for log-linear models and computational algebra of toric ideals.

Definitions & Notation

- Observed counts n_{ij} with sample size N
- Joint probability distribution $P=(p_{ij})$, $p_{ij}=P(X=x_i,\ Y=y_j)$
- Marginal probability distribution

$$p_{i+} = P(X = x_i) = \sum_{j=1}^{J} p_{ij} = \frac{n_{i+}}{N} = \frac{\sum_{j=1}^{J} n_{ij}}{\sum_{i,j} n_{ij}}$$

$$p_{+j} = P(Y = y_j) = \sum_{i=1}^{I} p_{ij}$$

- Conditional probability distribution

$$a_{ij} = P(X = x_i \mid Y = y_j) = \frac{p_{ij}}{p_{+j}} = \frac{n_{ij}}{n_{+j}}, \quad i = 1,\dots, I, \ \ j = 1,\dots, J$$

$$b_{ij} = P(Y = y_j \mid X = x_i) = \frac{p_{ij}}{p_{i+}} = \frac{n_{ij}}{n_{i+}}, \quad i = 1,\dots, I, \ \ j = 1,\dots, J$$

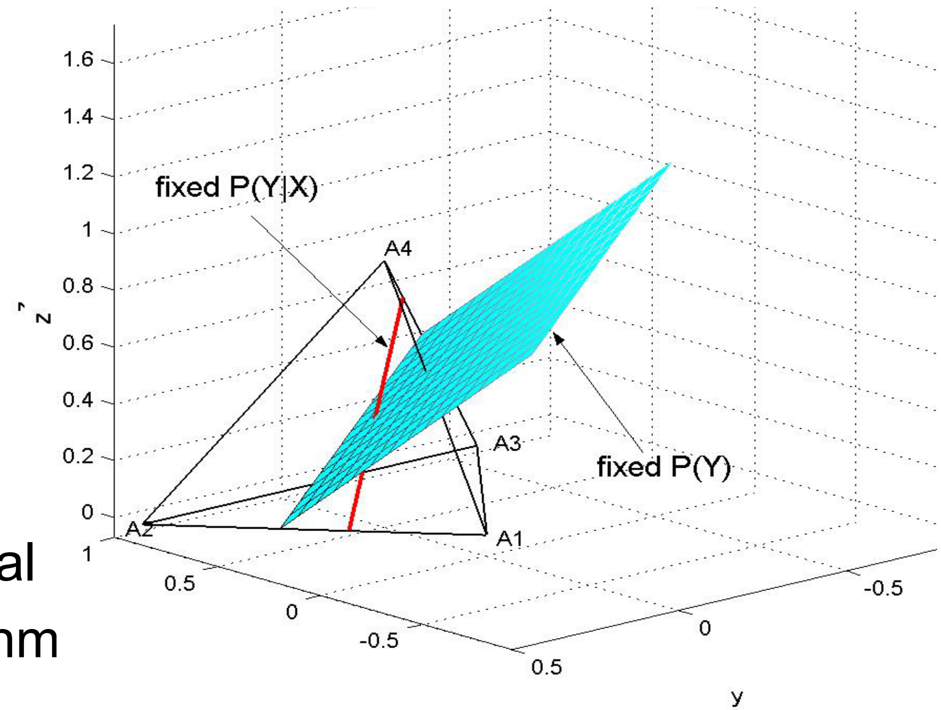
General structure 3x3 example

- *Max(min)* x_{ij} subject to Ax=b and $x \in \mathbb{N}^n$, $n = I \; x \; J$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1-b_{11} & -b_{11} & -b_{11} & 0 & 0 & 0 & 0 & 0 & 0 \\ -b_{12} & 1-b_{12} & -b_{12} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1-b_{21} & -b_{21} & -b_{21} & 0 & 0 & 0 \\ 0 & 0 & 0 & -b_{22} & 1-b_{22} & -b_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1-b_{31} & -b_{31} & -b_{31} \\ 0 & 0 & 0 & 0 & 0 & 0 & -b_{32} & 1-b_{32} & -b_{32} \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{21} \\ x_{22} \\ x_{23} \\ x_{31} \\ x_{32} \\ x_{33} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Uniqueness: Complete specification of the joint

- Uniqueness Theorem for *k*-way tables
  - Two-way table (Gelman & Speed 1993, Arnold et al. 1999)
    - *P(x), P(y|x)*
    - *P(x|y), P(y|x)*
  - Arnold et al. 1999
    - Sometimes: *P(y), P(y|x)*
    - Define the missing marginal
    - Vardi & Lee (1993) algorithm



Markov basis

- A set of generators of a toric ideal
  - Ring::=Q[n_{11}, n_{12}, n_{21}, n_{22}]=Q[χ]
  - $I_A = \langle \chi^{u+} - \chi^{u-} \mid \forall u \in Z^n_{\geq 0}, Au=0 \rangle$
- Example
  - 2x2 table [15,10,5,20]=[0.3,0.2,0.1,0.4]
  - Fixed f(y|x) = [0.6, 0.4, 0.2, 0.8] = [3/5, 2/5, 1/5, 4/5]
  - $A = \begin{vmatrix} 1 & 1 & 1 & 1 \\ 0.4 & -0.6 & 0 & 0 \\ 0 & 0 & 0.8 & -0.2 \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 & 1 \\ 2 & -3 & 0 & 0 \\ 0 & 0 & 4 & -1 \end{vmatrix}$
  - $n_{11}^{3} n_{12}^{2} - n_{21}^{1} n_{22}^{4}$

X, Y		
	+3	+2
	- 1	- 4

Software: 4ti2

Markov basis: reduction

- Step 1: given conditional frequencies as rationals $\begin{pmatrix} \frac{3}{5} & \frac{2}{5} \\ \frac{1}{5} & \frac{4}{5} \end{pmatrix}$

- Step 2: vector of denominators or sum of numerators $\boldsymbol{r}=(5\ \ 5)$

- Step 3: Markov basis for ***r*** $n_{1+} - n_{2+}$ $m=[1\ \ -1]$

- Step 4: exponents for Markov basis for fixed conditionals
  - $m_{ij}*(b_{j1}\ \ b_{j2} \ldots b_{jJ})$ 1*(3 2) and -1*(1 4)

One move: $\boldsymbol{n_{11}^3} n_{12}^2 - \mathrm{n}_{21}^1 n_{22}^4$

Theorem to be proven