Diaconis

BASICS

PROBABILITY THEORY:

$\mathcal{X}$ FINITE SET

$p(x) \geq 0 \quad \sum p(x) = 1$ PROBABILITY ON $\mathcal{X}$

PBLM; GIVEN $A \subseteq \mathcal{X}$

· COMPUTE APPROXIMATE UNDERSTAND

$$P(A) = \sum_{x \in A} p(x)$$

EX $\mathcal{X} = \{$ 4X4 ARRAYS WITH ROW/COL SUMS

·	·	·	·	220
·	·	·	·	215
·	·	·	·	93
·	·	·	·	64
108	286	71	127	592

$p(x) = 1/|\mathcal{X}|$ UNIFORM DISTRIBUTION

$A = \{x : T(x) \geq 138\}$, $T(x) = \sum_{i,j=1}^{4} \frac{(x_{ij} - r_i c_j / n)^2}{r_i c_j / n}$

ANS $P(A) \doteq .1831$

STATISTICS $\mathcal{X}$ FINITE SET

$p_\theta(x)$ $\theta \in \Theta$, FAMILY OF PROBABILITIES

OBSERVE x PICKED FROM SOME p_θ

PBLM GUESS θ AND SAY HOW SURE YOU ARE

EXPONENTIAL FAMILY $\Theta = \mathbb{R}^d$

$$p_\theta(x) = z^{-1}(\theta)\, e^{\theta \cdot T(x)}$$

$T(x) = (T_1(x), T_2(x) \cdots T_d(x))$ SUFFICIENT STATISTIC

$z(\theta) = \sum_x e^{\theta \cdot T(x)}$ NORMALIZING CONSTANT

NOTE $p_\theta(x \mid T(x) = t) = \dfrac{p_\theta(x)}{p_\theta(T(y) = t)} = \dfrac{1}{|y : T(y) = t|}$

DOESN'T DEPEND ON θ

3 BASIC TASKS:

TESTING, ESTIMATION, CONFIDENCE INTER.

<u>EX</u> MEN'S RESPONSE TO 'WOMEN SHOULD MIND THEIR HOMES AND LEAVE MEN TO RUN THE COUNTRY'

i	0	1	2	3	4	5	6	7	8	9	10
$n_1(i)$	4	2	4	6	5	13	25	27	75	29	32
$n(i)$	6	2	4	9	10	20	34	42	44	58	77
RATIO	.67	1	1	.67	.5	.65	.74	.64	.60	.50	.42

	11	12	13	14	15	16	17	18	19	20
	36	115	31	28	9	15	3	1	2	3
	95	360	101	107	32	125	32	29	15	23
	.38	.32	.31	.26	.28	.12	.09	.03	.13	.13

$\mathcal{X}$ = All BINARY VECTORS LENGTH 1309

$\Theta = \mathbb{R}^3$, $P_\theta(x)$ BUILT FROM $p(\text{YES} \mid i) = \dfrac{e^{\alpha + \beta i + \gamma i^2}}{1 + e^{\text{SAME}}}$

$$P_\theta(x) = z^{-1} \prod_{i=0}^{20} \binom{n(i)}{n_1(i)} e^{\alpha T_0 + \beta T_1 + \gamma T_2}$$

$$T_0 = \sum_1^{20} n_1(i), \quad T_1 = \sum_1^{20} i\, n_1(i) \quad T_2 = \sum_1^{20} i^2 n_1(i)$$

<u>TEST</u> iFey $\gamma = 0$ eg $p(Y|i)$ DOESN'T DEPEND ON i QUAD.

<u>ESTIMATE</u> β, γ

<u>SPECIFY</u> $\beta \pm \epsilon_2$ $\gamma \pm \epsilon_3$ SO

95% SURE TRUE β, γ COVERED

ONE APPROACH: GENERATE OTHER DATA SETS WITH SAME VALUES OF

$\sum i\, n_1(i)$, $\left(+ \text{ MAYBE } \sum i^2 n_1(i)\right)$

(All $n(i)$ FIXED)

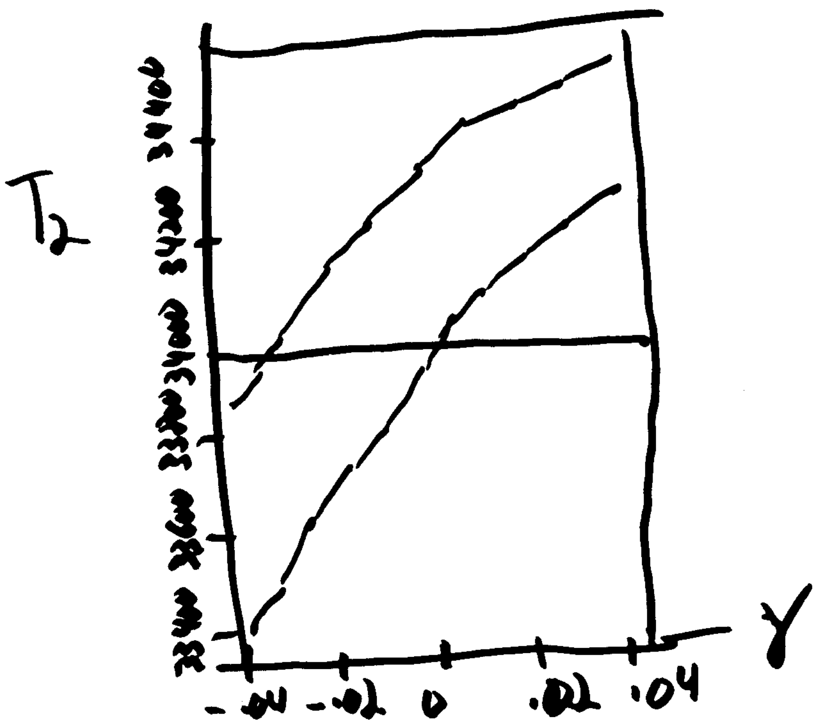
<u>EX</u> 95% C.I. FOR γ is $[-.032, .001]$

DONE BY USING MARKOV BASIS OF DIACONIS-GRAHAM-STURMFELS

+ METROPOLIS WITH γ FROM $-.04$ TO $.04$

FROM ASYA TAKENS' THESIS

OBSERVED $T_2 = 33965$



THREE 'DOABLE' OPEN PROBLEMS

1) TRY PERMUTATIONS: SEC 6 D+S 98

<u>EX</u> 2,262 GERMAN CITIZENS RANK ORDER

1) MAINTAIN ORDER (2) GIVE PEOPLE MORE SAY (3) FIGHT RISING PRICES (4) PROTECT FREEDOM OF SPEECH

1234	137	2134	48	3124	330	4123	21	
1243	29	2143	23	3142	294	4132	30	
1324	309	2314	61	3214	117	4213	29	
1342	255	2341	55	3241	69	4231	52	
1423	52	2413	33	3412	70	4312	35	
1432	93	2431	39	3421	34	4321	27	2262
	875		279		914		194	

	ITEM			
	875	279	914	194
POSITION	746	433	742	341
	345	773	419	725
	296	777	187	1002

$$\mathcal{X} = \{ f : S_4 \to \mathbb{N} : \sum_{\sigma} f(\sigma) = 2262 \}$$

	$\mathcal{X}$	V_0	$\oplus\ V_1$	$\oplus\ V_2$	$\oplus\ V_3$	$\oplus\ V_4$
DIM	24	1	9	4	9	1
LENGTH	657	462	381	268	48	4

<u>PBLM</u> FIX FIRST ORDER SUMMARY, GEN RANDOMLY

2) CONTINGENCY TABLES W. QUADRATIC STATISTICS

A. AGRESTI 'A SURVEY OF EXACT INFERENCE FOR CONTINGENCY TABLES' STAT. SCI 7 (1992)

$$\mathcal{X} = \left\{ T_{ij} \quad \begin{smallmatrix} 1 \le i \le I \\ 1 \le j \le J \end{smallmatrix} \quad \sum_j T_{ij} = r_i \quad \sum_i T_{ij} = c_j \right\}$$

WITH $\sum_{i,j} (i-2)(j-2) T_{ij} = t$ FIXED

3) POLYHEDRAL CONES AND TESTING

A. COHEN, J. KEMPERMAN H. SACKROWITZ
ANN STATIST 22 (1994)

OBSERVE $X_{ij} = \{ {}^1_0, \quad p(X_{ij} = 1 \mid x_{ij}) = z^{-1}(\theta_i)\, e^{x_{ij}\theta_i}$

$1 \le i \le I, \quad 1 \le j \le J$

TEST $\theta_i \equiv \theta$ VS $A \cdot \underset{\sim}{\theta} \ge 0$ $\quad$ A FIXED

eg, $\theta_1 \ge \theta_2 \ge \cdots \ge \theta_I$ OR $\theta_i \ge \theta_I$ All i

OR $\theta_1 \le \cdots \le \theta_a \ge \theta_{a+1} \ge \cdots \ge \theta_I$

NICE INTERACTION BETWEEN POLYHEDRAL COMBINATORICS + CONDITIONAL INFERENCE

COMPETING TECHNIQUES

a) CLASSICAL ASYMPTOTICS

b) EDGEWORTH CORRECTIONS

R. STRAWDERMAN M. WELLS JASA. 93 (1998)

c) STAT XACT

d) SEQUENTIAL IMPORTANCE SAMPLING

(Y. CHEN, P. DIACONIS, S. HOLMES, J. LIU)

SEQUENTIAL MONTE CARLO METHODS FOR STATISTICAL ANALYSIS OF TABLES, STANFORD TECH. REPORT

e) OTHERS

HOW DO THESE RELATE?

WARNING: CONDITIONAL INFERENCE is TRICKY

LEVEL	DELAY	RESPONSE L	RESPONSE D
1/8	NONE	0	6
	1½	0	5
1/4	N	3	3
	1½	0	6
1/2	N	6	0
	1½	2	4
1	N	2	4
	1½	5	1
2	N	2	0
	1½	5	0

2×2×5 TABLE

TEST NO-3 WAY INTERACTION

OR CONSTANT ODDS RATIO

$$\frac{p(11k)\,p(22k)}{p(21k)\,p(12k)} \equiv \Delta$$

SUFF STATISTICS
All LINE SUMS

D + S GIVE A BASIS

<u>EX</u> ASSUME CONSTANT Δ, TEST $\Delta = 1$, TEST BASED ON $U = \sum_{k=1}^{5} T_{11k}$ $(=16)$

USING EDGEWORTH $P(U \geq 16)$ "=" .... (6 DIGITS GIVEN)

<u>BUT</u> ONLY 4 TABLES WITH $U \geq 16$