Conditional Distributions, Log-linear Models, and Disclosure Limitation Methods*

Stephen E. Fienberg
Department of Statistics and Center for Automated Learning and Discovery
Carnegie Mellon University
Pittsburgh PA 15213-3890

August 31, 2001

Abstract

Much of the recent methodological literature on statistical disclosure limitation has dealt with methods for altering the interior of tables, especially in the form of cross-classification of counts, given certain marginal totals or subtables. These methods are closely related to those that use the exact distribution of a contingency table under a log-linear model given its sufficient statistics. Diaconis and Sturmfels have articulated the role of Gröbner bases in the calculation of such distributions. This talk will give an overview of disclosure limitation problems and methods to address them based on exact distributions and it will also discuss some interesting features of Gröbner bases that arise in these problems.

*Handout to accompany presentation at GROSTAT V, September, 2001.

An Example of Bounds for Table Entries

[These tables are taken from Dobra and Fienberg [16]. We include some additional tables based on a 10% random sample of the data.]

				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

Table 1: Prognostic factors in coronary heart disease. Source: Edwards and Havranek [7].

F	E	D	C	B / A	no / no	no / yes	yes / no	yes / 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]

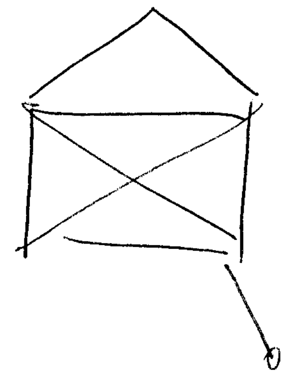

Table 2: Bounds for cell counts in the coronary heart disease table given margins corresponding to [BF][ADE][ABCE].

→ "Safe" to release.

F	E	D	C	B / A	no / no	no / yes	yes / no	yes / yes
neg	< 3	< 140	no		7	6	14	4
			yes		6	19	2	3
		≥ 140	no		6	3	5	1
			yes		8	6	1	1
	≥ 3	< 140	no		0	6	8	15
			yes		4	9	0	1
		≥ 140	no		2	2	6	4
			yes		3	3	0	3
pos	< 3	< 140	no		0	1	2	0
			yes		1	1	0	0
		≥ 140	no		0	0	1	0
			yes		1	1	0	0
	≥ 3	< 140	no		2	0	2	2
			yes		1	0	2	1
		≥ 140	no		0	0	3	2
			yes		1	2	0	0

Table 3: 10% sample selected from the population with coronary heart disease.

F	E	D	C	B / A	no / no	no / yes	yes / no	yes / yes
neg	< 3	< 140	no		[0,13]	[0,10]	[0,22]	[0,5]
			yes		[0,16]	[4,27]	[0,3]	[0,4]
		≥ 140	no		[0,13]	[0,10]	[0,22]	[0,5]
			yes		[0,16]	[0,12]	[0,3]	[0,4]
	≥ 3	< 140	no		[0,4]	[0,8]	[0,19]	[0,23]
			yes		[0,9]	[0,14]	[0,2]	[0,5]
		≥ 140	no		[0,4]	[0,8]	[0,15]	[0,16]
			yes		[0,9]	[0,14]	[0,2]	[0,5]
pos	< 3	< 140	no		[0,11]	[0,10]	[0,15]	[0,5]
			yes		[0,11]	[0,11]	[0,3]	[0,4]
		≥ 140	no		[0,11]	[0,10]	[0,15]	[0,5]
			yes		[0,11]	[0,11]	[0,3]	[0,4]
	≥ 3	< 140	no		[0,4]	[0,8]	[0,15]	[0,15]
			yes		[0,9]	[0,11]	[0,2]	[0,5]
		≥ 140	no		[0,4]	[0,8]	[0,15]	[0,15]
			yes		[0,9]	[0,11]	[0,2]	[0,5]

Table 4: Bounds for cell counts in the 10% sample table given margins corresponding to [BF][ADE][ABCE].

The Diaconis-Sturmfels Algorithm

[This material is extracted from Fienberg, Makov, Meyer, and Steele [24].]

Let **n** is the observed table, μ is the table of expected values under the model, **c** is the constraint vector representing the conditioning involving marginal totals, and $S(\mathbf{c})$ is the set of all nonnegative tables satisfying the marginal constraints. Let $\{f_1, f_2, \dots, f_L\}$ be a generating set for the tables in $S(\mathbf{c})$.

> *Lemma*: Let σ be a positive function on $S(\mathbf{c})$. Generate a Markov chain on $S(\mathbf{c})$ by choosing I uniformly in $\{1, 2, \dots, L\}$ and $\epsilon = \pm 1$ with probability 1/2 independently of I. If the chain is currently at $\mathbf{m}$ it moves to $\mathbf{m}' = \mathbf{m} + \epsilon f_I$ (provided that $\mathbf{m}' \in S(\mathbf{c})$ with probability $\min(1, \sigma(\mathbf{m}')/\sigma(\mathbf{m}))$. In all other cases the chain stays at $\mathbf{m}$. This is a connected, reversible Markov chain on $S(\mathbf{c})$ with a stationary distribution proportional to $\sigma(\mathbf{m})$.

By decoupling the "positive" and "negative" versions of the move to f_i for $i = 1, 2, \dots, L$, Diaconis and Sturmfels get transition probabilities that can be calculated for any model, even for nondecomposable loglinear models, as long as the margins we condition on are those that correspond to the minimal sufficient statistics. The argument is as follows.

From Haberman [11], we know that the underlying hypergeometric distribution for the exact distribution of the table under a loglinear model given a set of marginal constraints is

$$\sigma(\mathbf{n}) = \frac{(\prod_{i \in I} \frac{1}{n(i)!}) exp[\mathbf{n}, \mu]}{\sum_{\mathbf{m} \in S(\mathbf{c})} (\prod_{i \in I} \frac{1}{m(i)!}) exp[\mathbf{m}, \mu]} \tag{1}$$

where $\mathbf{n}$ is the observed table, μ is the table of expected values under the model, $\mathbf{c}$ is the constraint vector representing the conditioning involving marginal totals, and $S(\mathbf{c})$ is the set of all nonnegative tables satisfying the marginal constraints. When we condition on the margins that correspond to the minimal sufficient statistics under the model, the probabilities in equation (1) simplify because all of the exponential components are the same, yielding:

$$\sigma(\mathbf{n}) = \frac{\prod_{i\in I} \frac{1}{n(i)!}}{\sum_{\mathbf{m}\in S(\mathbf{c})}(\prod_{i\in I} \frac{1}{m(i)!})}. \quad (2)$$

The denominator in equation (2) is the same for each table with the specified margins and so the ratio of two such probabilities is only a function of the corresponding numerators.

There is a total of $9 + 6 = 15$ possible moves for the $3 \times 3 \times 2$ table, and these can occur with a change of sign as well. There are 9 basic or simple moves of the form:

1	−1	0
−1	1	0
0	0	0

−1	1	0
1	−1	0
0	0	0

formed by choosing a pair of rows, and a pair of columns in all possible ways. These take the form of embedding a Darroch-like local move in the corresponding $2 \times 2 \times 2$ subtable and set the other entries equal to 0. In addition, there are also $3! = 6$ possible "compound" moves of the form

1	−1	0
0	1	−1
−1	0	1

−1	1	0
0	−1	1
1	0	−1

The compound moves can be thought of as combinations of pairs of simple moves of the first type which allow one to reach extremal tables by first making a move outside the space of positive tables with fixed margins and then coming back via a second move. For the compound move given above we have

$$\frac{\sigma(\mathbf{m}')}{\sigma(\mathbf{m})} = \frac{m_{121}m_{231}m_{311}m_{112}m_{222}m_{332}}{(m_{111}+1)(m_{221}+1)(m_{331}+1)(m_{122}+1)(m_{231}+1)(m_{312}+1)}. \quad (3)$$

The 15 moves constitute a minimal generating set for the table and they correspond to a universal Gröbner basis. For each move there is a corresponding ratio of probabilities of the form $\sigma(\mathbf{m}')/\sigma(\mathbf{m})$.

In Table 6 we present the maximum likelihood estimates for the expected counts corresponding to the entries in Table 5 under the no 2nd-order interaction model with multinomial sampling We computed these in S-plus. The likelihood ratio chi-squared value for the fit of this model was 2.89 on 4 d.f. This is indicative of a moderately good model fit, although it is actually somewhat difficult to assess the fit given the sparseness of the row in the first layer which has a total count of 1 in it.

Gender = Male

Race	Income Level: ≤ $10,000	> $10000 and ≤ $25000	> $25000	Total
White	96	72	161	329
Black	10	7	6	23
Chinese	1	1	2	4
Total	107	80	169	356

Gender = Female

Race	Income Level: ≤ $10,000	> $10000 and ≤ $25000	> $25000	Total
White	186	127	51	364
Black	11	7	3	21
Chinese	0	1	0	1
Total	197	135	54	386

Table 5: Three-way cross-classification of Gender, Race, and Income for a selected U.S. census tract. (*Source*: 1990 Census Public Use Microdata Files)

Gender = Male

Race	Income Level: ≤ $10,000	> $10000 and ≤ $25000	> $25000	Total
White	97.09	72.15	159.76	329
Black	9.21	6.41	7.38	23
Chinese	0.70	1.44	1.86	4
Total	107	80	169	356

Gender = Female

Race	Income Level: ≤ $10,000	> $10000 and ≤ $25000	> $25000	Total
White	184.91	126.85	52.24	364
Black	11.79	7.58	1.62	21
Chinese	0.30	0.56	0.14	1
Total	197	135	54	386

Table 6: Maximum likelihood estimates for data in Table 5 under the no 2nd-order interaction model.

References

[1] Birch, M. W. (1963). Maximum Likelihood in Three-Way Contingency Tables. *Journal of the Royal Statistical Society, Series B*, **25**, 220–233.

[2] Bishop, Y. M. M. (1971). Effects of Collapsing Multidimensional Contingency Tables. *Biometrics*, **27**, 545–562.

[3] Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analyses: Theory and Practice.* MIT Press, Cambridge, MA.

[4] Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980). Markov Fields and Log-linear Interaction Models for Contingency Tables, *Annals of Statistics*, **8**, 522–539.

[5] Deming, W. E. and Stephan, F. F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known. *Annals of Mathematical Statistics*, **11**, 427-444.

[6] Diaconis, P. and Sturmfels, B. (1998). Algebraic Algorithms for Sampling From Conditional Distributions. *Annals of Statistics*, **26**, 363–397.

[7] Edwards, D.E. and Havranek, T. (1985). A Fast Procedure for Model Search in Multidimensional Contingency Tables. *Biometrika*, 72, 339–351.

[8] Fienberg, S. E. (1980). *The Analysis of Cross-classified Categorical Data (2nd edition).* MIT Press, Cambridge, MA.

[9] Good, I. J. (1963). Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables, *Annals of Mathematical Statistics.* **34**, 911–934.

[10] Haberman, S. J. (1973). Log-linear Models for Frequency Data: Sufficient Statistics and Likelihood Equations. *Annals of Statistics*, **1**, 617–632.

[11] Haberman, S. J. (1974). *Analysis of Frequency Data.* University of Chicago Press, Chicago IL.

[12] Lauritzen, S. (1996). *Graphical Models.* Oxford University Press, New York.

[13] Dobra, A. (2000). Computing Bounds for Entries in Contingency Tables Given a Set of Fixed Marginals. Technical Report, Department of Statistics, Carnegie Mellon University.

[14] Dobra, A. (2001). Markov Bases for Decomposable and Reducible Graphical Models. Submitted for publication.

[15] Dobra, A. and Fienberg, S. E. (2000). Bounds for Cell Entries in Contingency Tables Given Marginal Totals and Decomposable Graphs. *Proceedings of the National Academy of Sciences*, **97**, 11185–11192.

[16] Dobra, A. and Fienberg, S. E. (2001). Bounds for Cell Entries in Contingency Tables Induced by Fixed Marginal Totals. Paper prepared for 2nd Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, 14-16 March 2001, Skopje, Macedonia.

[17] Duncan, G. T. and Fienberg, S. E. (1999). Obtaining Information While Preserving Privacy: A Markov Perturbation Method for tabular Data. In *Statistical Data Protection, Proceedings of the Conference, Lisbon*, Eurostat, Luxembourg, 351–362.

[18] Duncan, G. T., Jabine, T. B., and Wolf, V. A. de (Eds.). (1993). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics.* National Academy Press, Washington, DC.

[19] Duncan, G. T., and Pearson, R. B. (1991). Enhancing Access to Microdata While Protecting Confidentiality: Prospects for the Future (with discussion). *Statistical Science*, **6**, 219–239.

[20] Fienberg, S. E. (1994). Conflicts Between the Needs for Access to Statistical Information and Demands for Confidentiality. *Journal of Official Statistics*, **10**, 115–132.

[21] Fienberg, S. E. (1999). Fréchet and Bonferroni bounds for Multi-way Tables of Counts with Applications to Disclosure Limitation. In *Statistical Data Protection, Proceedings of the Conference. Lisbon*, Eurostat, Luxembourg, 115–129.

[22] Fienberg, S. E. and Makov, U. E. (1998). Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data. *Journal of Official Statistics*, **14**, 385–397.

[23] Fienberg, S. E. and Makov, U. E. (2001). Uniqueness and Disclosure Risk: Urn Models and Simulation. *Research in Official Statistics*, **3**, in press.

[24] Fienberg, S. E., Makov, U. E., Meyer, M. M., and Steele, R.J. (2001). Computing the exact Distribution for a Multi-way Contingency Table Conditional on its Marginal Totals. In A.K.E. Saleh, ed., *Data Analysis from Statistical Foundations: Papers in Honor of D.A.S. Fraser*, Nova Science Publishing (2001), in press.

[25] Fienberg, S. E., Makov, U. E., and Sanil, A. P. (1997). A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data. *Journal of Official Statistics*, **13**, 75–89.

[26] Fienberg, S. E. and Makov, U. E., and Steele, R. J. (1998). Disclosure Limitation Using Perturbation and Related Methods for Categorical Data (with discussion). *Journal of Official Statistics*, **14**, 485–511.

[27] Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. R. J., and Wolf, P. P. de. (1998). Post Randomization for Statistical Disclosure Control: Theory and Implementation. *Journal of Official Statistics*, **14**, 463–478.

[28] Hundepool, A., Willenborg, L., Van Gemerden, L., Wessels, A., Fischetti, M., Salazar, J.-J., and Caprara, A. (1998b).*τ-Argus User's Manual.* Department of Statistical Methods, Statistics Netherlands.

[29] Hundepool, A., Willenborg, L., Wessels, A., Van Gemerden, L., Tiourine, S., and Hurkens, C. (1998). *μ-Argus User's Manual.* Department of Statistics, Statistics Netherlands.

[30] Roehrig, S. F., Padman, S., Duncan, G., and Krishnan, R. (1999). Disclosure Detection in Multiple Linked Categorical Datafiles: A Unified Network Approach. In *Statistical Data Protection, Proceedings of the Conference, Lisbon*, Eurostat, Luxembourg, 149–162.

[31] Raghunathan, T. E. and Rubin, D. B. (2001). Multiple Imputation for Statistical Disclosure Limitation. Unpublished manuscript.

[32] Rubin, D. B. (1993). Satisfying Confidentiality Constraints Through the Use of Synthetic Multiply Imputed Microdata. *Journal of Official Statistics*, **9**, 461–468.

[33] Samuels, S. M. (1998). A Bayesian, Species-sampling-inspired Approach to the Uniqueness Problem in Microdata Disclosure Risk Assessment. *Journal of Official Statistics*, **14**, 373–383.

[34] Skinner, C. J. and Holmes, D. J. (1998). Estimating the Re-identification Risk Per Record in Microdata. *Journal of Official Statistics*, **14**, 361–372.

[35] Trottini, M. (2001). A Decision-Theoretic Approach to Data Disclosure Problems. *Research in Official Statistics*, **3**, in press.

[36] Willenborg, L. and De Waal, T. (1996). *Statistical Disclosure Control in Practice.* Lecture Notes in Statistics Vol. 111, Springer Verlag, New York.

[37] Willenborg, L. and De Waal, T. (2001). *Elements of Statistical Disclosure Control.* Lecture Notes in Statistics Vol. 155, Springer Verlag, New York.