

Statistical Disclosure Limitation: Releasing Useful Data for Statistical Analysis*

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

April 28, 2003

An Example of Bounds for Table Entries

F	E	D	C	B			
				A	no		yes
				no	yes	no	yes
neg	< 3	< 140	no	44	40	112	67
			yes	129	145	12	23
	≥ 3	< 140	no	35	12	80	33
			yes	109	67	7	9
		≥ 140	no	23	32	70	66
			yes	50	80	7	13
pos	< 3	< 140	no	5	7	21	9
			yes	9	17	1	4
	≥ 3	< 140	no	4	3	11	8
			yes	14	17	5	2
		≥ 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 (1985).

*Handout to accompany presentation at Bureau of Transportation Statistics, Washington DC, April 28, 2003.

F	E	D	C	B		no		yes	
				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]	
	≥ 3	< 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]	
pos	< 3	< 140	no		[0,88]	[0,62]	[0,126]	[0,117]	
			yes		[0,134]	[0,134]	[0,25]	[0,38]	
	≥ 3	< 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]	
		≥ 140	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]. Source: Fienberg and Dobra (2001).

F	E	D	C	B		no		yes	
				A	no	yes	no	yes	
neg	< 3	< 140	no		[44,45]	[39,40]	[111,112]	[67,68]	
			yes		[128,129]	[145,146]	[12,13]	[22,23]	
	≥ 3	< 140	no		[34,35]	[12,13]	[80,81]	[32,33]	
			yes		[109,110]	[66,67]	[6,7]	[9,10]	
	≥ 3	≥ 140	no		[22,23]	[32,33]	[70,71]	[65,66]	
			yes		[50,51]	[79,80]	[6,7]	[13,14]	
		≥ 140	no		[24,25]	[24,25]	[72,73]	[57,58]	
		≥ 140	yes		[50,51]	[63,64]	[7,8]	[15,16]	
pos	< 3	< 140	no		[4,5]	[7,8]	[21,22]	[8,9]	
			yes		[9,10]	[16,17]	[0,1]	[4,5]	
	≥ 3	< 140	no		[4,5]	[2,3]	[10,11]	[8,9]	
			yes		[13,14]	[17,18]	[5,6]	[1,2]	
	≥ 3	≥ 140	no		[7,8]	[2,3]	[13,14]	[14,15]	
			yes		[8,9]	[16,17]	[2,3]	[2,3]	
		≥ 140	no		[3,4]	[0,1]	[13,14]	[10,11]	
		≥ 140	yes		[5,6]	[13,14]	[3,4]	[4,5]	

Table 3: Bounds for cell counts in Table 1 given all 5-way marginals.

Log-linear Models and Related Methods

References

- [1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.
- [2] Birch, M. W. (1963). Maximum Likelihood in Three-Way Contingency Tables. *Journal of the Royal Statistical Society, Series B*, **25**, 220–233.
- [3] Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: 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] Edwards, D. (2000). *Introduction to Graphical Modelling (2nd edition)*. Springer-Verlag, New York.
- [6] Edwards, D.E. and Havranek, T. (1985). A Fast Procedure for Model Search in Multidimensional Contingency Tables. *Biometrika*, **72**, 339–351.
- [7] Fienberg, S. E. (1980). *The Analysis of Cross-classified Categorical Data (2nd edition)*. MIT Press, Cambridge, MA.
- [8] Good, I. J. (1963). Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables, *Annals of Mathematical Statistics*. **34**, 911–934.
- [9] Haberman, S. J. (1973). Log-linear Models for Frequency Data: Sufficient Statistics and Likelihood Equations. *Annals of Statistics*, **1**, 617–632.
- [10] Diaconis, P. and Sturmfels, B. (1998). Algebraic Algorithms for Sampling From Conditional Distributions. *Annals of Statistics*, **26**, 363–397.
- [11] Haberman, S.J. (1974). *The Analysis of Frequency Data*. University of Chicago Press, Chicago.
- [12] Lauritzen, S. (1996). *Graphical Models*. Oxford University Press, New York.
- [13] Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

Computational Algebraic Geometry

References

- [1] Diaconis, P. and Sturmfels, B. (1998). Algebraic Algorithms for Sampling From Conditional Distributions. *Annals of Statistics*, **26**, 363–397.
- [2] 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), 145–165.
- [3] Garcia, L.D., Stillman, M., and Sturmfels, B. (2003). Algebraic Geometry of Bayesian Networks. Unpublished manuscript. <http://arXiv.org/abs/math/0301255>.

- [4] Dan Geiger, D., Meek, C., and Sturmfels, B. (2002). On the Toric Algebra of Graphical Models. February 2002, MSR-TR-2002-47. <http://math.berkeley.edu/~bernd/papers.html>.
- [5] Hosten, S. and Sturmfels, B. (2003). Computing the integer programming gap. Unpublished manuscript.
- [6] Pistone, G., Riccomagno, E., Wynn, H.P. (2001). *Algebraic Statistics: Computational Commutative Algebra in Statistics*. Chapman & Hall, New York.
- [7] Sturmfels, B. (1995). *Gröbner Bases and Convex Polytopes*. American Mathematical Society, Providence, RI.
- [8] Sturmfels, B. (2002). *Solving Systems of Polynomial Equations*. American Mathematical Society, Providence, RI.

Statistical Disclosure Limitation General Sources References

- [1] Domingo-Ferrer, J., ed. (2002). *Inference Control in Statistical Databases*. Lecture notes in Computer Science Vol. 2316, Springer-Verlag, Berlin.
- [2] Doyle, P., Lane, J. Theeuwes, J., and Zayatz, L., eds. (2001). *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Elsevier, Amsterdam.
- [3] *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems* (2002). Volume 10. Special Issue on Statistical Disclosure and Related Methods.
- [4] Duncan, G.T., Jabine, T.B., and de Wolf, V.A. (eds.) (1993). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. Panel on Confidentiality and Data Access, Committee on National Statistics, National Academy Press, Washington, DC.
- [5] *Journal of Official Statistics* (1993), Volume 9, and (1998), Volume 14. Special Issues on Statistical Disclosure Limitation.
- [6] 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.
- [7] Federal Committee on Statistical Methodology (1978). *Report on Statistical Disclosure and Disclosure-Avoidance Techniques*. Statistical Policy Working Paper 2. Subcommittee on Disclosure-Avoidance Techniques. U.S. Department of Commerce, Washington, DC.
- [8] Federal Committee on Statistical Methodology (1994). *Report on Statistical Disclosure Limitation Methodology*. Statistical Policy Working Paper 22. Subcommittee on Disclosure Limitation Methodology. Office of Management and Budget, Executive Office of the President, Washington, DC.
- [9] Fienberg, S. E. (1994). Conflicts Between the Needs for Access to Statistical Information and Demands for Confidentiality. *Journal of Official Statistics*, 10, 115–132.
- [10] Willenborg, L. and de Waal, T. (1996). *Statistical Disclosure Control in Practice*. Lecture Notes in Statistics, Vol. 111. Springer-Verlag, New York.

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

Categorical Data Bounds and Related Disclosure Methods References

- [1] Cox, L. (1980). Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association*, **75**, 377–385.
- [2] Cox, L. (1995). Network models for complementary cell suppression. *Journal of the American Statistical Association*, **90**, 1453–1462.
- [3] Cox, L. and Dandekar, R. (2002). Synthetic Tabular Data—An Alternative to Complementary Cell Suppression. Unpublished Manuscript.
- [4] Dobra, A. (2000). Computing Bounds for Entries in Contingency Tables Given a Set of Fixed Marginals. Technical Report, Department of Statistics, Carnegie Mellon University.
- [5] Dobra, A. (2001). Markov Bases for Decomposable Models. Submitted for publication.
- [6] Dobra, A. (2002). Statistical Tools for Disclosure Limitation in Multi-way Contingency Tables. Unpublished Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University.
- [7] Dobra, A., Erosheva, E., and Fienberg, S.E. (2003). Disclosure Limitation Methods Based on Bounds for Large Contingency Tables With Application to Disability Data,” In *Proceedings of Conference on the New Frontiers of Statistical Data Mining*, (H. Bozdogan, ed.), CRC Press, to appear.
- [8] 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**, 11885–11892.
- [9] Dobra, A. and Fienberg, S. E. (2001). Bounds for Cell Entries in Contingency Tables Induced by Fixed Marginal Totals. *Statistical Journal of the United Nations ECE*, **18**, 363–371.
- [10] Dobra, A. and Fienberg, S. E. (2003). Bounding entries in multi-way contingency tables given a set of marginal totals. *Foundations of Statistical Inference: Proceedings of the Shores Conference 2000* (Y. Haitovsky, H.R. Lerche, and Y. Ritov, eds.) Springer-verlag, New York, 3–16.
- [11] Dobra, A., Fienberg, S. E., and Trottni, M. (2003). Assessing the Risk of Disclosure of Confidential Categorical Data. In J. Bernardo, et al. eds., *Bayesian Statistics 7*, Oxford University Press, 125–144.
- [12] Dobra, A., Karr, A., Sanil, A., and Fienberg, S. E. (2002). Software Systems for Tabular Data Releases, *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, **10**, 529–544.
- [13] Dobra, A., Karr, A., and Sanil, A. (2002). Preserving Confidentiality of High-dimensional Tabulated Data: Statistical and Computational Issues. National Institute of Statistical Sciences, Technical Report 130.

- [14] Fienberg, S.E. (1994). *A radical proposal for the provision of micro-data samples and the preservation of confidentiality*. Technical Report No. 611, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.
- [15] 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.
- [16] Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R. and Roehrig, S.F. (2001). Disclosure Limitation Methods and Information Loss for Tabular Data. In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, (P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, eds.), Elsevier, Amsterdam, 135–166.
- [17] Duncan, G.T., Keller-McNulty and S.A.ynne Stokes, L. (2001). Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Submitted for publication.
- [18] Fienberg, S.E. (1994). *A radical proposal for the provision of micro-data samples and the preservation of confidentiality* . Technical Report No. 611, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.
- [19] 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), 145–165.
- [20] 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.
- [21] 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.
- [22] Raghunathan, T. E., Reiter, J., and Rubin, D. B. (2003). Multiple Imputation for Statistical Disclosure Limitation. *Journal of Official Statistics*, in press.
- [23] Rubin, D. B. (1993). Satisfying Confidentiality Constraints Through the Use of Synthetic Multiply Imputed Microdata. *Journal of Official Statistics*, **9**, 461–468.
- [24] Trottini, M. (2001). A Decision-Theoretic Approach to Data Disclosure Problems. *Research in Official Statistics*, **4**, 7–22.
- [25] Trottini, M. and Fienberg, S. E. (2002). Modelling user uncertainty for disclosure risk and data utility. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, **10**, 511–528.
- [26] Trottini, M., Fienberg, S.E., Makov, U.E., and Meyer, M.M. (2003). Additive noise and multiplicative bias as disclosure limitation techniques for continuous microdata: A simulation study. *Journal of Computational Methods for Science and Engineering*, **3**, 297–309.