Calculating Bounds Based on Conditional Frequencies

Byran Smucker, Aleksandra Slavković and Xiaotian Zhu

The Pennsylvania State University

September 24, 2008

Setting: Statistical Disclosure Limitation (SDL)

- Social and government agencies collect data and would like to release it to the public and/or researchers
- Must strike a balance between utility and privacy
- This work concerns a particular measure of disclosure risk, the feasibility interval
  - Bounds on a cell count in a contingency table that can be induced by released information
  - If these intervals are too narrow, particularly for cells with low counts, the disclosure risk may be unacceptably high

Generic Example of 2-way Contingency Table

	Download Yes	Download No	Total
Male	15	10	25
Female	5	20	25
Total	20	30	50

Table: Basic 2-way Table

We can calculate

- individual probabilities (e.g. $P(Yes \cap Male) = \frac{15}{50}$)
- marginal probabilities (e.g. $P(Female) = \frac{25}{50}$)
- conditional probabilities (e.g. $P(Yes|Male) = \frac{15}{25}$)

Past Work

Classic result is *Fréchet bounds*: Given an $I \times J$ table with total sample size (n_{++}) and marginal totals (n_{i+} and n_{+j}), the cell bounds for the ij^{th} cell are:

$$max\{0, n_{i+} + n_{+j} - n_{++}\} \leq n_{ij} \leq min\{n_{i+}, n_{+j}\}$$

- Dobra and Feinberg (2001, 2003), Cox (2007) and others have made extensions to k-way tables, though this is a difficult problem
- Slavković (2004), Slavković and Feinberg (2004) first looked at bounds induced by conditional probabilities in the context of SDL

What We Did

- Calculated both exact and linear relaxation bounds on cells given observed conditional probabilities using formulation which improves that of Slavkovic (2004), in the sense of handling sampling zeros
- Also derived closed-form solutions for the linear relaxation bounds
- However, we show that the linear relaxation bounds can be far wider than the corresponding exact bounds
- Give an improvement which tightens the linear relaxation bounds which is easy to calculate

2-way Contingency Tables: Notation

- 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 $P = \{p_{ij}\}, i = 1, ..., I, j = 1, ..., J$ be the joint probability distribution of these two random variables, where $p_{ij} = P(X = i, Y = j)$ and $\sum_i \sum_j p_{ij} = 1$
- Let $D = \{d_{ij}\}$ be the conditional probabilities, where $d_{ij} = \frac{p_{ij}}{p_{i\cdot}} = P(Y = j|X = i)$ for $i = 1, ..., I, j = 1, ..., J$

Note that these probability distributions involve true parameters. We will represent the observed conditionals as $\hat{D} = \{\hat{d}_{ij}\}$ with $\hat{d}_{ij} = \frac{o_{ij}}{o_{i\cdot}}$.

Formulation of Optimization Problem for 2×2 Table

	Download Yes	Download No	Total
Male	15 (0.6)	10 (0.4)	25
Female	5 (0.2)	20 (0.8)	25
Total	20	30	50

$$
\begin{aligned}
&\textit{Min } n_{ij} \\
&s.t.\ n_{11} + n_{12} + n_{21} + n_{22} = 50 \\
&\quad -\hat{d}_{12} n_{11} + \hat{d}_{11} n_{12} = 0 \\
&\quad -\hat{d}_{22} n_{21} + \hat{d}_{21} n_{22} = 0 \\
&\quad n_{11} + n_{12} \geq 1 \\
&\quad n_{21} + n_{22} \geq 1 \\
&\quad n_{ij} \geq 0, \quad \forall i, j
\end{aligned}
$$

Results for 2×2 Example

LP/IP Results:

	Download Yes	Download No
Male	15 [3,27],[0.6,29.4]	10 [2,18],[0.4,19.6]
Female	5 [1,9],[0.2,9.8]	20 [4,36],[0.8,39.2]

Table: IP and LP Results for 2-way Table

- Closed-form bounds: $\hat{d}_{ij} \leq n_{ij} \leq (N - (I - 1))\hat{d}_{ij}$
- We have no closed-form solution for integer bounds, though we noticed some interesting patterns

k-way Example: 4-way Contingency Table

- Due to Koch, $N = 193$
- Response: Recovery
- Factors: Treatment, Status, Center

			Recovery		
Center	Status	Treatment	Poor	Modest	Excellent
1	1	1	3 (0.107)	20 (0.714)	5 (0.179)
		2	11 (0.333)	14 (0.424)	8 (0.243)
	2	1	3 (0.103)	14 (0.483)	12 (0.414)
		2	6 (0.25)	13 (0.542)	5 (0.208)
2	1	1	12 (0.5)	12 (0.5)	0 (0)
		2	11 (0.524)	10 (0.476)	0 (0)
	2	1	3 (0.188)	9 (0.562)	4 (0.25)
		2	6 (0.333)	9 (0.5)	3 (0.167)

k-way Contingency Tables: Results for Full Conditionals

For full conditionals in a k-way table, there is a direct extension to result for 2-way tables:

$$\hat{d}_{IJ} \leq n_{IJ} \leq (N - (R - 1))\hat{d}_{IJ}$$

where

- R is the number of nonzero marginals in the 2-way table constructed from the k-way table
- I and J represent collections of indices representing response variables (for J) and random variables upon which we are conditioning (for I)

Results for k-way Tables and Partial Conditionals

- For a 4-way table, could have $X_I = \{X_1, X_2\}$, $X_J = \{X_4\}$, and $X_K = \{X_3\}$ so that $d_{IJ\cdot} = P(X_J|X_I)$
- Then the partial conditional probability is:

$$\hat{d}_{IJ\cdot} = \frac{\sum_K o_{IJK}}{\sum_J \sum_K o_{IJK}} \tag{1}$$

Result for linear relaxations: $0 \leq n_{IJK} \leq (N - (R-1))\hat{d}_{IJ\cdot}$ where R is the number of rows in the 2-way table constructed from the partial conditionals.

Example: *4*-way Contingency Table

- Due to Koch, $N = 193$
- Response: Recovery
- Factors: Treatment, Status, Center

			Recovery		
Center	Status	Treatment	Poor	Modest	Excellent
1	1	1	3 (0.107)	20 (0.714)	5 (0.179)
		2	11 (0.333)	14 (0.424)	8 (0.243)
	2	1	3 (0.103)	14 (0.483)	12 (0.414)
		2	6 (0.25)	13 (0.542)	5 (0.208)
2	1	1	12 (0.5)	12 (0.5)	0 (0)
		2	11 (0.524)	10 (0.476)	0 (0)
	2	1	3 (0.188)	9 (0.562)	4 (0.25)
		2	6 (0.333)	9 (0.5)	3 (0.167)

LP/IP Results for Example

- Let $X_1 = \textit{Center}$ $(i_1 = 1, 2)$, $X_2 = \textit{Status}$ $(i_2 = 1, 2)$, $X_3 = \textit{Treatment}$ $(i_3 = 1, 2)$, and $X_4 = \textit{Recovery}$ $(i_4 = 1, 2, 3)$
- For full conditionals, let $I = \{i_1, i_2, i_3\}$ and $J = \{i_4\}$
- Since there are no zero marginals, $R = 2^3 = 8$

C	S	T	Poor	Modest	Excellent
1	1	1	[3,6], [0.11,19.93]	[20,40], [0.71,132.86]	[5,10], [0.18,33.21]
		2	11, [0.33,62]	14, [0.42,78.91]	8, [0.24,45.09]

- Lower bound for first cell: $\hat{d}_{1111} = 0.107$
- Upper bound for first cell: $(N - (R - 1)) \cdot \hat{d}_{1111} = 19.9$

Cell Bound Improvement

- Given exact (unrounded) conditional probabilities, tightens LP relaxations (similar results to Slavkovic, 2004)
- Computations much simpler than for IP bounds and don't require LP
- Idea of algorithm may serve as basis of more efficient algorithm for sharp bounds as well as one which can take into account error induced by rounding (future work)

$$\frac{\hat{d}_{IJ}}{\ell_I} \leq n_{IJ} \leq (N - \sum_{T \neq I} 1/\ell_T) \cdot \hat{d}_{IJ} \tag{2}$$

where ℓ_I is the smallest positive conditional probability in row I. Since we know that those n_{IJ}'s are integers, we can round the lower bounds up and round the upper bounds down.

Logic of Improvement: Lower Bounds

- Finding lower bounds
  - For a given row, find the smallest positive conditional probability. This must have a count of at least 1.
  - Divide the conditional probabilities for all other cells in the row by the smallest, which will give you a real number. Round this number up to the nearest integer and this is the minimum for that cell
    - e.g. if smallest is .1 and another cell has .25, you know that the smallest has a count of at least 1 and that the other cell has at least .25/.1=2.5 times more than the smallest. Round up to 3 for the lower bound for this cell.
  - Repeat this for each row.

Logic of Improvement: Upper Bounds

- Finding upper bounds
  - For a given row, need to find the minimum count that is taken up by other rows.
  - Use the minimum counts calculated earlier, add them up for all other rows, then subtract this from the total sample size. This gives the maximum count which can be distributed in this row.
  - Multiply this maximum count by the conditional probability for each cell and this gives the upper bound.

Discussion

- Calculated sharp bounds given observed conditionals
  - For smaller tables, IP seemed to produce intervals narrow enough to raise privacy concerns
  - For larger tables (e.g. 6-way and 8-way tables we computed), IP seemed to produce wider intervals
  - For 8-way table computation prohibitively long
- Because of rounding, released conditionals do not allow calculation of integer bounds in most realistic scenarios
  - Must have exact conditionals to calculate the sharp bounds
  - Thus, intruders could not calculate these bounds
- Closed-form solutions for linear relaxations
  - Noted that there can be large differences between the linear relaxations and the sharp bounds
  - Thus, unfortunately linear relaxation bounds are probably not an adequate shortcut to the sharp bounds

Conclusions and Future Work

- Conclusions
  - Provides additional tools for publishing agencies to determine if tables of conditional rates are safe to release.
  - It appears that for large tables, conditional probabilities are generally safe to release, and this could be checked by agencies beforehand.
- Future Work
  - Can we devise a more efficient algorithm to calculate sharp bounds?
  - Since we cannot calculate the sharp bounds unless we are given exact conditionals (because of rounding), can we "sharpen" the bounds given the rounded conditionals?
  - How much uncertainty does the rounding introduce?

Acknowledgements

- Thanks to advisor Dr. Slavković and Xiaotian Zhu
- Thanks to the NSF, grant SES-0532407 to Penn State Statistics
- For more information, visit http://www.stat.psu.edu/ sesa/privacy.html