Cell Bounds in Two-Way Contingency Tables Based on Conditional Frequencies

Byran Smucker and Aleksandra B. Slavković

Department of Statistics, Pennsylvania State University,
University Park, PA 16802, U.S.A.
bjs379@psu.edu, sesa@stat.psu.edu

Abstract. Statistical methods for disclosure limitation (or control) have seen coupling of tools from statistical methodologies and operations research. For the summary and release of data in the form of a contingency table some methods have focused on evaluation of bounds on cell entries in k-way tables given the sets of marginal totals, with less focus on evaluation of disclosure risk given other summaries such as conditional probabilities, that is, tables of rates derived from the observed contingency tables. Narrow intervals - especially for cells with low counts - could pose a privacy risk. In this paper we derive the closed-form solutions for the linear relaxation bounds on cell counts of a two-way contingency table given observed conditional probabilities. We also compute the corresponding sharp integer bounds via integer programming and show that there can be large differences in the width of these bounds, suggesting that using the linear relaxation is often an unacceptable shortcut to estimating the sharp bounds and the disclosure risk.

Keywords: Confidentiality; Contingency tables; Integer programming; Linear programming; Statistical disclosure control; Tabular data.

1 Introduction

Social or government agencies often collect data with intent to release a sufficient amount as public information that can be used for statistical inference, the results of which could affect policy decisions or further research. However, if too much information is released, confidentiality of individuals or organizations that has likely been guaranteed upon collection of the data could be compromised. Thus, there must be a trade-off between releasing the useful data and maintaining privacy.

Statistical disclosure limitation (SDL) deals with developments of methods and tools for evaluating trade-offs between disclosure risk and data usefulness. Many of the SDL methods developed in recent years lie at the interface of operations research and statistical methods; see a detailed review in [20]. There are many ways in which data confidentiality can be violated, as well as many ways to determine whether a violation has occurred. In this paper we are concerned with tabular data releases (e.g., marginal totals and conditional probabilities

J. Domingo-Ferrer and Y. Saygın (Eds.): PSD 2008, LNCS 5262, pp. 64–76, 2008.

with sample sizes) and the *feasibility interval* [25], that is, the bounds on a cell entry in a contingency table that can be induced by given released information. If these feasibility intervals are too narrow - or if the table is uniquely identified because the lower and upper bounds are the same - the risk of a disclosure could be high, particularly in cells with small counts.

We are particularly interested in the cell bounds that can be calculated when we are given observed conditional probabilities, that is, tables of rates derived from the observed table of counts. This is an important question because although many categorical data summaries are in the form of marginal tables, agencies often release rates or percentages representing proportions of individuals who fall in a certain category given some other characteristics (see [21], p. 7 for an example). Furthermore, the conditionals preserve association measures such as odds and odds-ratios relevant for data utility (e.g., see [21], [16]). We explore the question of what information about original cell counts can be extracted from knowing these conditional probabilities along with the sample size, and thus what is the effect on disclosure risk. This paper widens the statistical disclosure limitation literature by considering the effect of releasing a summary statistic - conditional probabilities - that heretofore has received little attention.

In this paper we calculate cell bounds given conditional probability information and sample size, using an integer/linear programming formulation. We improve upon the formulation proposed by Slavković and Fienberg [22] by requiring the marginals upon which we condition to be nonzero while allowing individual cells to be zero, and derive the closed-form solutions for the linear relaxation bounds on cell counts thus significantly reducing necessary computing time. This formulation actually produces somewhat wider bounds than those in [22], but is more realistic since it accommodates sampling zeros. In Section 2, we give some technical background on the optimization formulation and a brief review of the current results on calculation of cell bounds in contingency tables. In sections 3 and 4, we describe the linear and integer programming formulation for two-way tables and derive closed-form solutions for the linear relaxation, demonstrating them with two simple examples. We differentiate between two formulations depending if the calculation is done by a data owner or an intruder. We also compute the corresponding sharp integer bounds via integer programming and show that there can be large differences in the width of these bounds, suggesting that using the linear relaxation is often an unacceptable shortcut to estimating the sharp bounds and the disclosure risk for contingency tables given observed conditional frequencies.

2 Optimization Methods and Cell Bound Calculation for Contingency Tables

In this paper, we solve linear and integer programs in order to calculate cell bounds for two-way contingency tables given certain information. A linear

program consists of a linear objective function, optimized subject to linear constraints. It can be represented in standard form as:

$$\begin{aligned} &Minimize && \mathbf{cx} \\ &subject\ to && \mathbf{Ax} = \mathbf{b} \\ & && \mathbf{x} \geq \mathbf{0} \end{aligned} \tag{1}$$

where there are n decision variables and m constraints, $\mathbf{c}$ is a row vector of length n, $\mathbf{x}$ is a column vector of length n, $\mathbf{A}$ is a $m \times n$ matrix, and $\mathbf{b}$ is a column vector of length m. An integer program can be formulated as (1) with the additional constraints that all decision variables be integer. Note that decision variables in this context are variables within an optimization program whose values are to be optimized; a random variable in the larger statistical context are variables whose values are determined by some random process (or, more technically, random variables are functions from a given sample space to the real numbers).

In the context of this paper, we use integer programming (IP) to calculate exact integer upper and lower bounds on entries in contingency tables. Integer programs are solved using methods such as Branch-and-Bound and Branch-and-Cut algorithms (see [18]), as in CPLEX, the commercial software [15] used in this work.

Calculating cell bounds for the entries of contingency tables given marginal totals has a long history, and goes back to Bonferroni [1], Fréchet [12], and Hoeffding [13] in their work on bounds for cumulative distribution functions given univariate marginals ([9], [10]). Given an $I \times J$ table with total sample size (n_{++}) and marginal totals (n_{i+} and n_{+j}), the *Fréchet bounds* have the following form for the ij^{th} cell:

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

Work has been done on generalizations of these bounds, i.e. bounds for k-way contingency tables. Given marginal totals for k-way tables, Dobra and Fienberg [6] give explicit formulas for the bounds when the table can be represented as a decomposable graph, a construct in which the expected counts in the cells of the table can be written as functions of the marginals. They extended these results to the case in which the graph is reducible, though when the table cannot be represented as a graph (which is often the case), other methods such as linear programming must be employed. The same authors in [7] further extended this idea to general k-way tables by generalizing the "shuttle algorithm" originally developed by Buzzigoli and Gusti [2] for three-way tables. This algorithm exploits hierarchical relationships within the table, and sequentially updates the bounds for cells until they cannot be further improved. A number of similar problems and approaches have been addressed in the context of statistical disclosure control but these have generally either successfully focused on two-way tables (e.g. [17], [2]) or broken down in higher-dimensional contexts (see [3]), though Cox [4] demonstrates that so-called network tables can overcome some of these problems.

Significantly less work has been done on examining bounds induced by given observed conditional probabilities. Researchers have begun to examine the cell

bounds induced by conditional probabilities in conjunction with given marginals, as well as conditionals alone, using both mathematical programming (linear and integer) and tools from algebraic statistics such as Markov bases (e.g., see [21], [22], [11], and [5]). In this paper we only focus on the calculation of linear and integer bounds given conditionals alone, but offer an improved formulation with closed form bounds thus reducing the potential computational burden; one of the biggest criticisms of using Markov bases for these problems is in part computational inefficiency.

As mentioned above, a natural way to obtain sharp bounds given marginals and/or conditionals is via IP. Solutions to IP's can be difficult and computationally expensive, which may lead to the desire to use the linear relaxation bounds as an approximation to the sharp IP bounds. Given the marginals, the maximal gap between an IP and its linear relaxation has been studied and theoretically has been shown to be exponentially large ([24], [14]). These results imply that it could be misleading to assess disclosure risk by using the linear relaxation as an approximation to the sharp integer bounds. Onn [19] also showed that there could be arbitrary gaps in the bounds on cell entries given the margins, which could further increase the disclosure risk. In this paper, we show empirically that the same is true in the case of given conditional probabilities.

3 Bounds for Cells in Two-Way Tables Given Conditional Probabilities

In this section we consider $I \times J$ tables, using a simple 2×2 example to demonstrate the formulation of the integer and linear programming problems with result on cell bounds. Then, using this formulation we prove a theorem about the linear relaxation bounds for this situation. We assume a single, unweighted tabular data release.

3.1 Setting and Notation

Let X and Y be two random variables and $O = \{o_{ij}\}$ be the $I \times J$ table (matrix) of observed counts with sample size N. The joint probability distribution of these two random variables can be represented as $P = \{p_{ij}\}, i = 1, ..., I, j = 1, ..., J$, where $p_{ij} = P(X = i, Y = j)$ and $\sum_i \sum_j p_{ij} = 1$. Further, the marginal probability distributions for X and Y are $p_{i\cdot} = \sum_{j=1}^{J} p_{ij} = P(X = i)$ and $p_{\cdot j} = \sum_{i=1}^{I} p_{ij} = P(Y = j)$ respectively, and conditional probability distributions are $C = \{c_{ij}\}$ and $D = \{d_{ij}\}$ where $c_{ij} = \frac{p_{ij}}{p_{\cdot j}} = P(X = i|Y = j)$ and $d_{ij} = \frac{p_{ij}}{p_{i\cdot}} = P(Y = j|X = i)$ for $i = 1, ..., I, j = 1, ..., J$. Additionally, $\sum_i c_{ij} = 1$ and $\sum_j d_{ij} = 1$.

Note that these probability distributions involve true parameters, and under the assumption of multinomial sampling the observed counts are just estimators of those parameters. We are in particular interested in the estimated (observed)

conditional probabilities and will represent them as $\hat{C} = \{\hat{c}_{ij}\}$ and $\hat{D} = \{\hat{d}_{ij}\}$ with $\hat{c}_{ij} = \frac{o_{ij}}{o_{\cdot j}}$ and $\hat{d}_{ij} = \frac{o_{ij}}{o_{i\cdot}}$.

As we have stated, the observed counts for the ij^{th} cell are represented by o_{ij} while in the following integer and linear programs, the decision variables (those variables which can be varied subject to constraints) used to define cell bounds are represented by n_{ij}. One can think of the observed counts as fixed (as they are a realization from the joint probability distribution P), while the n_{ij}'s can vary with relation to the optimization programs. In what follows, we focus on the case of given row conditionals, $\hat{D}$ and sample size, but similar statements can be derived for the column conditionals, $\hat{C}$.

3.2 Formulation of Optimization Problem for a 2×2 Table

We use a simple fictitious example (see [22]) to demonstrate the optimization setup. Suppose we have a sample of 25 male students and 25 female students and we ask them whether they have ever illegally downloaded mp3's on the internet. Thus X=gender and Y=illegally downloaded? with $i = 1, 2$ (male, female), $j = 1, 2$ (yes, no). These data are summarized in Table 1.

Table 1. Counts for 2-way Table

	Download Yes	Download No
Male	15	10
Female	5	20

From Table 1, using $\hat{d}_{ij} = \frac{o_{ij}}{o_{i1}+o_{i2}}$, we can calculate the following 2×2 matrix of row conditional probabilities; that is, the percentage of students downloading activity given gender:

$$\hat{D} = \begin{bmatrix} 0.6 & 0.4 \\ 0.2 & 0.8 \end{bmatrix}$$

Similarly, we can calculate $P(Gender|Download) = \hat{c}_{ij} = \frac{o_{ij}}{o_{1j}+o_{2j}}$:

$$\hat{C} = \begin{bmatrix} \frac{3}{4} & \frac{1}{3} \\ \frac{1}{4} & \frac{2}{3} \end{bmatrix}$$

Since in this case we have conditional probabilities that are repeating decimals, rounding becomes an issue which can affect the calculated bounds. We explore this issue in Section 3.3.

In most statistical models for contingency table data, each population parameter, p_{ij}, is assumed to be greater than zero (i.e. no structural zeros). However, for a given sample we can certainly observe a sampling zero. Because of this, instead of placing a lower bound of 1 on each cell (as in [22]), we make the lower bound zero and instead require that each margin have a count of at least one. This is necessary to satisfy the definition of conditional probability.

With this in mind, to calculate linear relaxation lower bounds on the ij^{th} cell counts in the original table based on the row conditionals (matrix $\hat{D}$), the following linear program is constructed:

$$Min\ n_{ij} \tag{2}$$
$$s.t.\ n_{11} + n_{12} + n_{21} + n_{22} = 50 \tag{3}$$
$$-\hat{d}_{12}n_{11} + \hat{d}_{11}n_{12} = 0 \tag{4}$$
$$-\hat{d}_{22}n_{21} + \hat{d}_{21}n_{22} = 0 \tag{5}$$
$$n_{11} + n_{12} \geq 1 \tag{6}$$
$$n_{21} + n_{22} \geq 1 \tag{7}$$
$$n_{ij} \geq 0, \quad \forall i, j \tag{8}$$

In the above linear program, the $\hat{d}_{ij}$'s are assumed known and calculated from the observed data O. The corresponding integer program to calculate exact bounds is formulated by simply including integer constraints on all decision variables. This formulation corresponds to *Example 2* in [22], except for the constraints given by equations (6) and (7) above. In fact, the same linear/integer program can be written with the observed cell counts, o_{ij}'s instead of $\hat{d}_{ij}$'s, by replacing equations (4) and (5) by

$$-o_{12}n_{11} + o_{11}n_{12} = 0 \tag{9}$$
$$-o_{22}n_{21} + o_{21}n_{22} = 0. \tag{10}$$

This formulation with the original counts has not been considered before, but it is important for providing feasible IP solutions and a more precise assessment of the disclosure risk by the data owner. We discuss this further below and in Section 4.1.

To calculate lower bounds for each cell, we solve four optimization problems, each one having a different cell in the objective function. To calculate upper bounds on each cell, we solve the same four optimization problem but maximize the objective function instead of minimize. The results of the integer program are listed in Table 2. These bounds using the improved formulation are actually the same as calculated in [22], although this may not be the case in general.

Similarly, we can calculate the sharp integer bounds for the $\hat{C}$ conditionals (see Table 3). Because the $\hat{C}$ conditionals require rounding, the IP often gives infeasible solution, and these integer bounds can only be calculated if the original data are given. Thus the agency can calculate the exact IP bounds using o_{ij}'s while an intruder, given no other external information, can only calculate the LP-relaxation bounds (see Table 3). In the next section, we present the closed form solution for the LP-relaxation bounds, and their implications for disclosure.

3.3 Exact Formulas for Linear Relaxation Bounds Given Conditional Probabilities

For the linear relaxation as we have formulated it in (2)-(8), notice that the lower bounds for each cell in Table 2 are equal to the conditional probability for that

cell. We prove this along with the closed form solution for the upper bounds for $I \times J$ tables, and display the results of a simple calculation for the *mp3* data. In [22], the LP lower bounds were some integer or real-valued number greater than or equal to 1, but there were no closed form solutions. These new results can be extended to k-way contingency tables, as we show in [23], since the linear program associated with it has the same form.

Table 2. IP and LP Results for 2-way Table for $\hat{D}$ conditionals

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

Recall that I is the number of categories in the first variable, J is the number of categories in the second, and N is the total sample size. We prove *Theorem* 1 in the case of our 2×2 example. Any other size of contingency table would have a linear program with the same structure, and could be proved similarly.

Theorem 1. *Assume we have an $I \times J$ contingency table, and none of the rows in the contingency table sum to zero. Based on the conditional probabilities $P(Y = j|X = i)$ and the sample size N, we can construct a linear program of the form* (2)-(8)*. This linear program is minimized when $n_{ij} = \hat{d}_{ij}$, and maximized for the ij^{th} cell at $(N - (I - 1))\hat{d}_{ij}$.*

Proof. For the lower bound, note that the lower bound for n_{ij} cannot be zero (unless $\hat{d}_{ij} = 0$, for which the result holds), because if it were, the other cell which defines its conditional distribution would be forced to zero by (4) or (5). This cannot happen because of constraints (6) and (7). Constraints (4) and (5) are derived from the the conditional probability relationship $\hat{d}_{ij} = \frac{n_{ij}}{\sum_j n_{ij}}$. Since (6) and (7) hold, and n_{ij} is minimized when its marginal is as small as possible, n_{ij} will be minimized when its marginal is 1, which forces n_{ij} to be precisely equal to its conditional probability, $\hat{d}_{ij}$.

For the upper bound, we maximize the objective function defined in (2). Since we are maximizing n_{ij}, the marginal total for each of the rows (beside the i^{th} row) in the contingency table will be as small as possible, namely 1, as required by constraints (6) and (7). This is possible because each of the cells can have a value equal to their conditional probability. Thus, for all but the i^{th} row, the marginal total is 1. So now there are $N - (I - 1)$ counts to distribute among the J cells in row i. Because constraints (4) and (5) are derived from the given conditional probabilities (i.e. $\hat{d}_{ij} = \frac{n_{ij}}{\sum_j n_{ij}}$), n_{ij} can be no larger than the value which satisfies $\frac{n_{ij}}{N-(I-1)} = \hat{d}_{ij}$ which means n_{ij} is maximized at $n_{ij} = (N - (I - 1))\hat{d}_{ij}$. □

To demonstrate the calculation for the first cell, just note that the lower bound is simply the associated $\hat{d}$ conditional probability, 0.6, and the upper bound is $(50-(2-1))*0.6 = 29.4$. The LP-relaxation bounds given are slightly wider than

the IP bounds (Table 2); in this case, they seem to be a reasonable approximation. Although, we would argue that a more mathematically precise definition of "reasonable approximation" is needed.

Similarly, we can derive the result for given the observed column conditional, $\hat{C}$ and sample size. Now the bound would be: $\hat{c}_{ij} \leq n_{ij} \leq (N-(J-1))\hat{c}_{ij}$. For the $\hat{C}$ conditionals, we show in Table 3 the sharp integer bounds as well as the linear relaxations given conditional probabilities rounded to one and two decimal places. Notice the effect that rounding can have on the LP bounds.

Table 3. IP and LP results (rounded to one and two decimal places) for two-way table for $\hat{C}$ conditionals

	Download Yes	Download No
Male	[6,33], [0.8,39.2], [.75,36.75]	[2,14], [0.3,14.7], [0.33,16.17]
Female	[2,11], [0.2,9.8], [0.25, 12.25]	[4,28], [0.7,34.3], [0.67,32.83]

While the IP bounds calculated in this way give a more precise assessment of disclosure risk than their LP counterparts, it has been pointed out that the gaps exist even within the bounds; e.g., by using algebraic tools Slavković and Fienberg [22] showed that there are only four possible tables of counts satisfying these constraints. Agencies can use the rounding to release less precise values. This leads to some loss of utility but also to a gain in protection as the bounds become wider. The effect of such rounding on data utility and the number of possible tables is currently being explored by Lee and Slavković [16].

4 Example: Delinquent Children Data

In this section we consider a 4×4 table of counts originally used in [8] to demonstrate various statistical disclosure techniques for tabular data. Slavković and Fienberg [22] used this example to demonstrate the effect of released conditional frequencies in comparison to release of marginal totals, and utilized tools from computational algebra and Markov bases for the calculation of bounds and the number of tables. Table 4 shows the number of juvenile delinquents broken down by county and education level. Titles, row and column headings are fictitious.

Consider the case in which we are given the sample size, $N = 135$, as well as an estimate of $P(\textit{Education Level}|\textit{County})$, that is $\hat{D}$. We can calculate the linear relaxation bounds immediately using Theorem 1. Slavković and Fienberg [22] calculated sharp integer bounds using Markov bases, and showed, at that time a surprising result, that there is only one table of counts that satisfies these released conditionals. We show here that the data owner does not need to use the algebraic tools but can get the same bounds and thus the same result by solving the integer program described below by using the observed counts.

Table 4. 4×4 Table. Delinquent Children Data and Integer Programing Bounds.

	Low	Medium	High	Very High
Alpha	15	1	3	1
Beta	20	10	10	15
Gamma	3	10	10	2
Delta	12	14	7	2

4.1 Formulation of Optimization Problems for 4×4 Example

Similar to Section 3.2, an integer program can be constructed as follows:

$$
\begin{aligned}
&Min\ n_{ij} && (11)\\
&s.t.\ \sum_i \sum_j n_{ij} = N\\
&\quad \hat{d}_{ij} \sum_{k \neq j} n_{ik} + (\hat{d}_{ij} - 1) n_{ij} = 0,\ \forall i,\ j = 1, 2, 3 && (12)\\
&\quad \sum_j n_{ij} \geq 1\ \forall i\\
&\quad n_{ij} \geq 0\ \forall i, j\\
&\quad n_{ij}\ integer\ \forall i, j
\end{aligned}
$$

where $\hat{d}_{ij}$ are elements of $\hat{D}$, and equation (12) is derived from the following:

$$
\begin{aligned}
\hat{d}_{ij} \sum_k n_{ik} - n_{ij} &= \hat{d}_{ij} n_{ij} - n_{ij} + \hat{d}_{ij} \sum_{k \neq j} n_{ik}\\
&= \hat{d}_{ij} \sum_{k \neq j} n_{ik} + (\hat{d}_{ij} - 1) n_{ij} = 0
\end{aligned}
$$

If we know all but one of the conditional probabilities the last one is determined, and this eliminates four constraints of the original form: $\hat{d}_{ij} = \frac{n_{ij}}{\sum_k n_{ik}}\ \forall i, j$.

Because the decimal representations of the numbers in $\hat{D}$ must be rounded, this integer program is infeasible. However, if we consider the conditional probability in terms of the original data, we can construct an integer program that is feasible. Let $\hat{d}_{ij} = \frac{o_{ij}}{\sum_k o_{ik}} = \frac{n_{ij}}{\sum_k n_{ik}}$. Linearizing the second equality leads to:

$$
\begin{aligned}
0 = o_{ij} \sum_k n_{ik} - \sum_k o_{ik} n_{ij} &= o_{ij} n_{ij} - \sum_k o_{ik} n_{ij} + o_{ij} \sum_{k \neq j} n_{ik}\\
&= o_{ij} \sum_{k \neq j} n_{ik} + (o_{ij} - \sum_k o_{ik}) n_{ij}
\end{aligned}
$$

$\forall i, j = 1, 2, 3$, where the o_{ij}'s are the observed cell counts in Table 4. By replacing constraints in equation (12) by the one above, we can calculate the sharp integer

bounds on the cell counts. Note that the simplification of coefficients in the matrix assumes knowledge of the marginal distribution, and thus under the assumptions of this section would not be available to an intruder. Again, these bounds could only be calculated by the agency releasing the data.

Notice that in this example, the sharp integer bounds uniquely identify the original table (that is, the lower bound is equal to the upper bound). Slavković and Fienberg [22] showed the same result (for IP) but by using tools from algebraic geometry and Markov bases. This is also an extreme case of entry uniqueness problem which is related to the entry uniqueness given the margins (see [19]). Table 5 shows the linear relaxation results calculated using Theorem 1, rounding to one, two, and three decimal places, respectively. Note that the bounds given in [22] are uniformly narrower than the bounds presented here. However, this is the result of an unrealistic formulation which forces each cell to have a count of at least 1. When this constraint is relaxed, the wider bounds in Table 5 result.

It is evident, with this example as well, that rounding can have a significant effect on the bounds providing a "false" sense of disclosure risk since the bounds are much wider. However, notice that the cell with small counts do have short LP-relaxation bounds given the rounding at two or three decimal places, e.g. $o_{12} = [0.05, 6.6]$. The data owner would most likely decide in this case that these bounds are too tight and not release the conditional frequencies with the sample size, even without running the above described IP and knowing that there is only one possible table.

Table 5. Linear Relaxation Results for 4×4 Table (rounding to 1, 2, and 3 decimal places)

	Education Level			
County	Low	Medium	High	Very High
Alpha	[0.7,92.4], [0.75,99], [0.75,99]	[0.1,13.2], [0.05,6.6], [0.05,6.6]	[0.1,13.2], [0.15,19.8], [0.15,19.8]	[0.1,13.2], [0.05,6.6], [0.05,6.6]
Beta	[0.3,39.6], [0.37,48.84], [0.363,47.916]	[0.2,26.4], [0.18,23.76], [0.182,24.024]	[0.2,26.4], [0.18,23.76], [0.182,24.024]	[0.3,39.6], [0.27,35.64], [0.273,36.036]
Gamma	[0.1,13.2], [0.12,15.84], [0.12,15.84]	[0.4,52.8], [0.4,52.8], [0.4,52.8]	[0.4,52.8], [0.4,52.8], [0.4,52.8]	[0.1,13.2], [0.08,10.56], [0.08,10.56]
Delta	[0.3,39.6], [0.34,44.88], [0.343,45.276]	[0.4,52.8], [0.4,52.8], [0.4,52.8]	[0.2,26.4], [0.2,26.4], [0.2,26.4]	[0.1,13.2], [0.06,7.92], [0.057,7.524]

5 Conclusions

To date statistical disclosure limitation methodologies for tables of counts have been heavily focused on the release of unaltered marginal totals from such tables,

and in part on inferences that are possible by an intruder from such releases. Many statistical agencies also release other forms of summary data from tables, such as tables of observed conditional frequencies. These are predominantly released as two-way and three-way tables, with conditioning on a single variable.

In this paper, we improved on the LP/IP formulation initially proposed in [22] by not restricting the counts in individual cells to be greater than one. While a zero marginal would result in a division by zero when calculating a conditional probability, there need not be any such restriction upon individual cells. The result is wider - though more realistic - bounds as well as closed-form solutions for the linear relaxation bounds thus reducing typically necessary optimization computing time. The proposed bounds hold even if there are observed zero cell counts. These zeros, however, may reveal extra information about their complementary cells and this requires some further careful investigation in particular for k-way tables. These new results can be extended to k-way contingency tables, as we show in [23], since the linear program associated with it has the same form.

Our improved formulation also circumvents the feasibility problem with calculation of sharp IP bounds given observed conditionals and sample size by calculating them using the observed counts directly, a fact relevant for data owners. The simple examples also show that IP may produce significantly narrower bounds than the linear relaxation of the same optimization problem. These large discrepancies can be seen especially in large and sparse tables k-way tables which we further explore in [23]. Because of these discrepancies and potential gaps within the IP bounds similar to the gaps described by [19] in the case of margins, the LP bounds often do not seem to be good approximation to the IP bounds. Thus these LP bounds may not be often reasonable for detecting whether there is a "true" potential disclosure, except perhaps as a crude approximation in the event of time-prohibitive sharp IP calculations. More precise mathematical definition of "reasonable" approximation is needed.

Note that in the 4×4 table (in addition to some other example we considered but not reported here), the sharp integer bounds given full conditional probabilities uniquely identify the counts in the original table. This occurred more often with smaller tables, but actually the most elementary example of all (the 2×2 table) did not yield a unique specification. At this point, we do not fully understand the underlying characteristics of a table that would produce a unique specification. There is some kind of tradeoff between the sample size and the number of cells, though in our examples the ratio between these two quantities certainly does not suggest anything obvious.

To further examine the relationship between the sample size and bounds given conditionals, and their effect on risk and utility, we are currently running simple simulations. For example, if we multiply each entry in Table 4 by 10, this has the effect of changing the sample size from 135 to 1350, increasing the width of IP and LP-relaxation bounds, and increasing the number of possible tables while maintaining the same conditional probabilities. Having the same conditionals is important for the utility aspect of SDL as they preserve certain associations

within cell counts in the table. Also, in most of the datasets we analyzed (excluding the small example in Section 3.2) the IP based on the released conditionals proved infeasible because of rounding issues. Therefore, it is likely, without external information, that in practice releasing conditional probabilities would not allow intruders to calculate sharp integer bounds, but would give sufficient information for statistical inference.

Acknowledgments

The research reported here was supported in part by NSF Grant SES-0532407 to the Department of Statistics, Pennsylvania State University, and NSF Grant DMS-0439734 to the Institute for Mathematics and Its Application at the University of Minnesota.

References

1. Bonferroni, C.E.: Teoria statistica delle classi e calcolo delle probabilitá. Publicazioni del R. Instituto Superiore di Scienze Economiche e Commerciali di Firenze, 8 (1936)
2. Buzzigoli, L., Gusti, A.: An algorithm to calculate the upper and lower bounds of the elements of an array given its marginals. In: Statistical Data Protection (SDP 1998) Proceedings, pp. 131–147. Eurostat, Luxembourg (1998)
3. Cox, L.: Bounds on entries in 3-dimensional contingency tables. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. LNCS, vol. 2316, pp. 21–33. Springer, Heidelberg (2002)
4. Cox, L.: Contingency tables of network type: Models, markov basis and applications. Statistica Sinica 17, 1371–1393 (2007)
5. Dobra, A., Fienberg, S., Rinaldo, A., Slavković, A., Zhou, Y.: Algebraic statistics and contingency table problems: Log-linear models, likelihood estimation and disclosure limitation. In: Putinar, M., Sullivant, S. (eds.) IMA Volumes in Mathematics and its Applications: Emerging Applications of Algebraic Geometry, vol. 149, pp. 63–88. Springer, Heidelberg (2008)
6. Dobra, A., Fienberg, S.E.: Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. Statistical Journal of the United Nations Economic Commission for Europe 18(4), 363–371 (2001)
7. Dobra, A., Fienberg, S.E.: Bounds for cell entries in contingency tables induced by fixed marginal totals. Statistical Journal of the United Nations ECE 18, 363–371 (2003)
8. Federal Committe on Statistical Methodology, Statistical Policy Working Paper 22 (Version Two). Report on Statistical Disclosure Limitation Methodology (2005)
9. Fienberg, S.E.: Fréchet and Bonferroni bounds for multi-way tables of counts with applications to disclosure limitation. In: Statistical Data Protection: Proceedings of the Conference, pp. 115–129. Eurostat, Luxembourg (1999)
10. Fienberg, S.E.: Contingency tables and log-linear models: Basic results and new developments. Journal of the American Statistical Association 95(450), 643–647 (2000)

11. Fienberg, S.E., Slavkovic, A.B.: Preserving the confidentiality of categorical statistical data bases when releasing information for association rules. Data Mining and Knowledge Discovery 11, 155–180 (2005)
12. Fréchet, M.: Les Probabilitiés Associées a un Système dÉvénments Compatibles et Dépendants, Vol. Premiere Partie. Hermann & Cie, Paris (1940)
13. Hoeffding, W.: Scale-invariant correlation theory. Schriften des Mathematischen Instituts und des Instituts fur Angewandte Mathematik der Universit at Berlin 5(3), 181–233 (1940)
14. Hosten, S., Sturmfels, B.: Computing the integer programming gap (2003), http://www.citebase.org/abstract?id=oai:arXiv.org:math/0301266
15. ILOG CPLEX, ILOG CPLEX 10.1 User's Manual. ILOG (2006)
16. Lee, J., Slavković, A.: Synthetic tabular data preserving the observed conditional probabilities. In: PSD 2008 (submitted, 2008)
17. Lu, H., Li, Y., Wu, X.: Disclosure analysis for two-way contingency tables. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 57–67. Springer, Heidelberg (2006)
18. Nemhauser, G.L., Wolsey, L.A.: Integer and Combinatorial Optimization. Wiley-Interscience (1988)
19. Onn, S.: Entry uniqueness in margined tables. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 94–101. Springer, Heidelberg (2006)
20. Salazar-Gonzalez, J.-J.: Statistical confidentiality: Optimization techniques to protect tables. Computers and Operations Research 35, 1638–1651 (2008)
21. Slavković, A.B.: Statistical Disclosure Limitation Beyond the Margins: Characterization of Joint Distributions for Contingency Tables. PhD thesis, Carnegie Mellon University (2004)
22. Slavković, A.B., Fienberg, S.E.: Bounds for cell entries in two-way tables given conditional relative frequencies. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 30–43. Springer, Heidelberg (2004)
23. Smucker, B., Slavković, A.: Cell bounds in K-way tables given conditional frequencies. Journal of Official Statistics (to be submitted, 2008)
24. Sullivant, S.: Small contingency tables with large gaps. Siam J. Discrete Math. 18(4), 787–793 (2005)
25. Willenborg, L., de Waal, T.: Statistical Disclosure Control in Practice. Lecture Notes in Statistics III. Springer, New York (1996)