BOUNDING ENTRIES IN 3-DIMENSIONAL CONTINGENCY TABLES

Lawrence H. Cox, Associate Director
National Center for Health Statistics, U.S. Centers for Disease Control and Prevention
6525 Belcrest Road, Room 915
Hyattsville, MD 20782 USA
LCOX@CDC.GOV

Abstract

Problems in statistical science including estimation and statistical data security have led to recent interest in the problem of determining exact integer bounds on entries in multi-dimensional contingency tables subject to fixed integer marginal totals and in related problems. We investigate the case of the 3-dimensional integer planar transportation problem (3-DIPTP) which has been the subject of recent investigations in the statistics and operations research literature. We provide a new, easily derived bound for the objective function 3-DIPTP and demonstrate its utility for analyzing convergence behavior. Our analysis discloses that previously proposed methods are not exact but based on necessary but not sufficient conditions for solving 3-DIPTP, and also are insensitive to whether a feasible table even exists. To expand the basis for understanding these estimation and data security problems, we compare previously proposed algorithms analytically, demonstrate the existence of fractional extremal points, discuss implications for statistical data base query systems, and present new directions involving MCMC computation.

Keywords: transportation; combinatorial optimization; data security; exact bounds

1. INTRODUCTION

A problem of recurring interest in operations research since the 1950s (e.g., Schell 1955), and in particular during the 1960s (e.g., Moravek and Vlach 1967) and 1970s (Smith 1973), is that of establishing sufficient conditions for the existence of a feasible solution to the *3-dimensional planar transportation problem (3-DPTP)*, viz., a solution to the linear program:

$$\sum_{j=1}^{d_1} n_{ijk} = n_{+jk}, \quad \sum_{i=1}^{d_2} n_{ijk} = n_{i+k}, \quad \sum_{k=1}^{d_3} n_{ijk} = n_{ij+}, \quad n_{ijk} \geq 0 \tag{1}$$

where n_{+jk}, n_{i+k}, $n_{ij+} \geq 0$ are constants, referred to as the *2-dimensional marginal totals* (in m-dimensions, the

(m-1)-dimensional marginal totals). Vlach (1986) provides an excellent summary of attempts on the problem. These methods typically involve iteration over a suitably selected system of linear equality and/or inequality constraints. When totals are integer, the number of iterations is necessarily finite. Unfortunately, each of these methods has been shown (e.g., by Moravek and Vlach 1967 and Vlach 1986) to yield necessary but not sufficient conditions for feasibility. Cox (2000) provides a sufficient condition for more general and multi-dimensional transportation problems based on an iterative nonlinear statistical procedure known as *iterative proportional fitting* (Deming and Stephan 1940). The purpose here is not to revisit solving 3-dimensional transportation problems but to examine the role of feasibility in the pursuit of exact integral lower and upper bounds on internal entries in a 3-DPTP subject to integer constraints (viz., a *3-DIPTP*) and to describe further research directions. These issues are of recurring interest and importance as they are at the intersection of statistical science and operations research.

Our notation suggests that internal and marginal entries are integer. Integrality is not required for 3-DPTP (1), nor by the feasibility procedure of Cox (2000). However, henceforth integrality of all entries is assumed as we focus on *contingency tables*--tables of nonnegative integer frequency counts and totals--and on the 3-DIPTP.

Obvious necessary condition for feasibility of 3-DIPTP are given by the *consistency conditions*:

$$\sum_{k=1}^{d_3} n_{ij+} = \sum_{j=1}^{d_2} n_{ij+}, \quad \sum_{i=1}^{d_1} n_{ij+} = \sum_{k=1}^{d_3} n_{+jk}, \quad \sum_{i=1}^{d_1} n_{i+k} = \sum_{j=1}^{d_2} n_{+jk}, \qquad n_{ijk} \geq 0 \tag{2}$$

Denote the respective values by n_{i++}, n_{+j+}, n_{++k}. These are the *1-dimensional marginal totals* (in m-dimensions, the *(m-2)-dimensional marginal totals*), and $n_{+++} = \sum_{i=1}^{d_1} n_{i++} = \sum_{j=1}^{d_2} n_{+j+} = \sum_{k=1}^{d_3} n_{++k}$ is the *grand total*. It is customary in the operations research literature to represent the 2-dimensional totals as:

$$\boldsymbol{A_{jk}} = \sum_{i=1}^{d_1} n_{ijk}, \; \boldsymbol{B_{ik}} = \sum_{j=1}^{d_2} n_{ijk}, \; \boldsymbol{C_{ij}} = \sum_{k=1}^{d_3} n_{ijk}, \qquad n_{ijk} \geq 0 \tag{3}$$

The *feasibility problem* is that of the existence of one or more integer solutions to (1) subject to consistency (2) and integrality conditions on the 2-dimensional marginal totals. The *bounding problem* is to determine integer lower and upper linear bounds on each internal entry n_{ijk} over all contingency tables satisfying (1)-(2). *Exact bounding* amounts to determining the interval $[\ \min\{n_{ijk}\},\ \max\{n_{ijk}\}\ ]$, computed over all integer

feasible solutions $\boldsymbol{n}^{(} = \{n^{(}_{ijk}\}$ of (1)-(2).

The bounding problem has received recent attention in statistics, particularly for data security, as follows. To prevent unauthorized disclosure of confidential subject-level data, it is often necessary to thwart narrow estimation of small counts (see U.S. Department of Commerce 1994). In lieu of releasing the internal entries of a 3-dimensional contingency table, a national statistical office (NSO) may instead release only the 2-dimensional marginal totals. This amounts to releasing data at one higher level of aggregation to avoid disclosure at the lower level. An important question for the NSO is then: how closely can a third party estimate the suppressed internal entries using the published marginal totals? During large-scale data production such as for a national census or survey, the NSO needs to answer this question thousands of times. NSOs are moving away from periodic release of static sets of tabulations towards providing data users with on-line, query-based access to potentially dynamic statistical data bases, thus magnifying confidentiality concerns. These factors stimulate research on the (exact) bounding problem aimed at the development of efficient bounding algorithms for repetitive, large scale use.

NSOs subject statistical data to a range of verification and "cleaning" processes. Much data is derived from probability samples, requiring estimation of population-level statistics. Data in a statistical data base can come from multiple sources, at various times, and may have been modified through a variety of statistical procedures such as rounding. Cox (2000) demonstrates that any of these factors can produce an *infeasible table*, viz., marginal totals that satisfy (1)-(2) but for which no feasible solution exists, and furthermore that infeasible tables are *ubiquitous*, viz., infeasible tables of every conceivable size exist in 3- and higher-dimensions, and *abundant*, viz., for every pair of one feasible and one infeasible table of equal size and dimension, there exists a countably infinite number of additional infeasible tables of the same size and dimension. This makes it possible and indeed likely that sets of consistent 2-dimensional (integer) marginal totals fail to define a feasible 3-dimensional (contingency) table. To be useful, then, bounding methods for 3- or multi-dimensional tables must be sensitive to infeasibility, viz., must fail to converge when presented with an infeasible problem (1)-(2). Otherwise, meaningless data will be released to the public and erroneous inferences made for policy purposes.

The advent of public access statistical data base query systems has stimulated recent research by statisticians on the bounding problem, some of which is based on mathematical programming models. Unfortunately, feasibility and its consequences have been ignored. One motivation of this paper is to highlight and explore this issue. We do so through examination of four papers representing separate approaches to bounding

problems. Three of the papers were presented at the International Conference on Statistical Data Protection (*SDP'98*), March 25-27, 1998, Lisbon, Portugal, sponsored by the Statistical Office of the European Communities (EUROSTAT). This conference is the most recent, in-depth technical forum for examination of statistical data protection methodology. Papers were refereed prior to acceptance and a proceedings volume was published in 1999. The first three papers selected for examination here comprised the session, "Bounds" of the conference, and presumptively represent current research thinking. The fourth paper appeared in *Management Science* and reports current research findings. These papers, plus an additional paper discussed in Section 5, are representative of current research on the bounding problem, but more can be said, as demonstrated here.

Section 2 provides preliminaries on bounding in 2-, 3- and m-dimensional tables. Section 3 deals with two of the three conference papers. These methods are related to results from probability theory developed during the 1930s and 1940s that are receiving recent attention in statistics (e.g., Rüschendorff et al. 1996). We compare the two algorithms, demonstrate that one is superior, and highlight their similarity to necessary conditions to 3-DIPTP introduced by Schell (1955). We demonstrate that neither algorithm is exact and that both are insensitive to infeasibility, thereby calling into question their usefulness by NSOs and others for practical application. Section 4 addresses the third conference paper and the *Management Science* paper, both based on networks. We show that conference paper fails to generalize as claimed to arbitrary 3-dimensional tables, and that the method of *Management Science* paper is insensitive to infeasibility and indeed over-complicated as the problem is easily solved by simple methods developed here. Section 5 deals with the presence of fractional extremal points and a method for avoiding them for a specialized class of problems different from those considered here, due to Dobra and Fienberg (2000). Section 6 contains discussion and presents a new avenue for research related to MCMC.

2. THE F-BOUNDS

Given a 2-dimensional table with consistent sets of column ($\{n_{\%j}\}$) and row ($\{n_{i\%}\}$) marginal totals, the *nominal upper bound* for n_{ij} equals $\min\ \{n_{\%j},\ n_{i\%}\}$. The *nominal lower bound* is zero.

It is well-known how to obtain exact bounds in 2-dimensions. The nominal upper bound is exact, by virtue of the *stepping stones algorithm*: set n_{ij} equal to its nominal upper bound, and subtract this value from the column, row and grand totals. Either the column total or the row total (or both) must become zero: set all entries in the corresponding column (or row, or both) equal to zero and drop this column (or row, or both) from the table. Arbitrarily pick an entry from the remaining table, set it equal to its nominal upper bound, and continue as above. In

a finite number of iterations, a completely specified, consistent 2-dimensional table exhibiting the nominal upper bound for n_{ij} will be reached.

Theorem 2.1: Exact lower bounds on internal entries n_{ij} in a 2-dimensional table are given by

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

Proof: As $n_{\%j} - n_{i\%} - n_{\%\%} \leq n_{ij} - \sum_{I \neq i,\ J \neq j} n_{IJ}$, then $n_{ij} \geq \max \{0,\ n_{\%j} - n_{i\%} - n_{\%\%}\}$. Exactness follows from observing that $\sum_{I \neq i,\ J \neq j} n_{IJ} \leq 0$ is feasible if $n_{\%j} - n_{i\%} - n_{\%\%} \geq 0$. *Q.E.D.*

Therefore, in 2-dimensions, exact bounds are given by:

$$\min \{n_{\%j},\ n_{i\%}\} \geq n_{ij} \geq \max \{0,\ n_{\%j} - n_{i\%} - n_{\%\%}\} \tag{4}$$

The bounds of Theorem 2.1 generalize to m-dimensions, as follows. Each internal entry in an m-dimensional table is contained in m (m-1)-dimensional marginal totals, each of which provides an upper bound on the entry. The minimum of these m totals is the nominal upper bound. Similarly, each entry is contained in precisely m(m-1)/2 2-dimensional tables, each of which yields a candidate lower bound. The maximum of these lower bounds and zero provides a lower bound on the entry. Unlike the 2-dimensional case, in m $\geq$ 3 dimensions these bounds are not necessarily exact (Cox 1999). To avoid confusion in following sections, we refer to these bounds as the *F-bounds* (as they are in fact due to Fréchet and others). By direct application of the logic of Theorem 2.1, we obtain:

Corollary 2.2: In 3-dimensions, the F-bounds are:

$$\min \{n_{\%jk}\ ,\ n_{i\%k}\ ,\ n_{ij\%}\} \geq n_{ijk} \geq \max \{0,\ n_{ij\%} + n_{i\%k} + n_{i\%\%}\ ,\ n_{ij\%} + n_{\%jk} + n_{\%j\%}\ ,\ n_{i\%k} + n_{\%jk} + n_{\%\%k}\} \tag{5}$$

While sharper bound than those of Corollary 2.2 can be derived by an iterative process, in the following section we demonstrate the utility of the F-Bounds for analysis.

3. THE BOUNDING METHODS OF FIENBERG (1999) AND BUZZIGOLI AND GIUSTI (1999)

Fienberg (1999) proposes an approach for bounding internal entries in a 3-dimensional contingency table given fixed 2-dimensional marginal totals, based on classical methods in probability theory due to Fréchet (1940, 1951) and Bonferroni (1936). Buzzigoli and Giusti (1999) offer a bounding method for 3-DPTP based on iterative improvement of upper bounds based on current lower bounds and of lower bounds based on current upper bounds. In this section, we analyze these methods separately and then compare them to each other and to the F-bounds of Corollary 2.2.

3.1 The Procedure of Fienberg (1999)

Fienberg (1999) does not specify a bounding algorithm precisely, but illustrates an approach via example. The example is a 3x3x2 table of sample counts from the 1990 Decennial Census Public Use Sample (Fienberg 1999, Table 1). For convenience, we present it in the following form. The internal entries are:

	INCOME					
	High	*Med*	*Low*	*High*	*Med*	*Low*
White	96	72	161	186	127	51
Black	10	7	6	11	7	3
Chines	1	1	2	0	1	0
	MALE			*FEMALE*		

Table 1: Fienberg (1999), Table 1

and the 2-dimensional marginal totals are: $\boldsymbol{A} = \begin{pmatrix} 107 & 197 \\ 80 & 135 \\ 169 & 54 \end{pmatrix}$, $\boldsymbol{B} = \begin{pmatrix} 329 & 364 \\ 23 & 21 \\ 4 & 1 \end{pmatrix}$, $\boldsymbol{C} = \begin{pmatrix} 282 & 199 & 212 \\ 21 & 14 & 9 \\ 1 & 2 & 2 \end{pmatrix}$.

In the remainder of this sub-section, we examine the properties of the bound procedure of Fienberg (1999).

Corresponding to each internal entry n_{ijk}, there exists a *collapsed* 2x2x2 table:

n_{ijk}	$\sum_{J \neq j} n_{iJk}$
$\sum_{I \neq i} n_{Ijk}$	$\sum_{I \neq i, J \neq j} n_{IJk}$

$\sum_{K \neq k} n_{ijK}$	$\sum_{J \neq j, K \neq k} n_{iJK}$
$\sum_{I \neq i, K \neq k} n_{IjK}$	$\sum_{I \neq i, J \neq j, K \neq k} n_{IJK}$

Table 2: Collapsing A 3-Dimensional Table Around Entry n_{ijk}

The entry in the lower right-hand corner, viz., entry (2, 2, 2), is referred to as the *complement* of n_{ijk}, denoted $\bar{n}_{ijk}$. Fix (i, j, k). Denote the 2-dimensional marginals of Table 2 to which $\bar{n}_{ijk}$ contributes by $\bar{n}_{\%22}$, $\bar{n}_{2\%2}$, $\bar{n}_{22\%}$. Observe that:

$$n_{\%\%\%} + n_{\bar{\imath}\%\%} + n_{\%\bar{\jmath}\%} + n_{\%\%k} - n_{i\bar{\jmath}\%} - n_{i\%k} - n_{\%\bar{\jmath}k} + \bar{n}_{ijk} = (n_{\%\%\%} + n_{\bar{\imath}\%\%} + n_{\%\bar{\jmath}\%} - n_{i\bar{\jmath}\%}) - (n_{i\%k} - n_{\%\bar{\jmath}k} + n_{\%\%k}) + (\bar{n}_{ijk}) = (\bar{n}_{22\%}) - (n_{ijk} + (\bar{n}_{22\%} + \bar{n}_{ijk})) + (\bar{n}_{ijk}) = n_{ijk} \quad (6)$$

From (6) results the *2-dimensional Fréchet lower bound* of Fienberg (1999):

$$n_{ijk} \geq \max\{0,\ n_{\%\%\%} - n_{i\%\%} - n_{\%j\%} - n_{\%\%k} + n_{ij\%} + n_{i\%k} + n_{\%jk} - \min\{\bar{n}_{\%22},\ \bar{n}_{2\%2},\ \bar{n}_{22\%}\}\} \tag{7}$$

The nominal (also called *Fréchet*) upper bound on n_{ijk} equals $\min\{n_{\%jk},\ n_{i\%k},\ n_{ij\%}\}$. The *2-dimensional Fréchet bounds* of Fienberg (1999) are thus:

$$\min\{n_{\%jk},\ n_{i\%k},\ n_{ij\%}\} \geq n_{ijk} \geq \max\{0,\ n_{\%\%\%} - n_{i\%\%} - n_{\%j\%} - n_{\%\%k} + n_{ij\%} + n_{i\%k} + n_{\%jk} - \min\{\bar{n}_{\%22},\ \bar{n}_{2\%2},\ \bar{n}_{22\%}\}\} \tag{8}$$

Simple algebra reveals that the lower F-bounds of Section 2 and the 2-dimensional lower Fréchet bounds of Fienberg (1999) are identical. The F-bounds are straightforward (as they do not require the computation of complements) and easier to implement (as they can be implemented at sight). Consequently, we replace (8) by (5).

From (6) also follows the *2-dimensional Bonferroni upper bound* of Fienberg (1999):

$$n_{\%\%\%} - n_{i\%\%} - n_{\%j\%} - n_{\%\%k} + n_{ij\%} + n_{i\%k} + n_{\%jk} \geq n_{ijk} \tag{9}$$

Fienberg (1999)'s Bonferroni upper bound is not redundant: if $\max\{\bar{n}_{ijk}\}$ is sufficiently small, it can be sharper than the nominal upper bound. This is illustrated by the n_{111} entry of the 2x2x2 table with marginals:

$$\boldsymbol{A} = \begin{pmatrix} 2 & 8 \\ 9 & 8 \end{pmatrix},\quad \boldsymbol{B} = \begin{pmatrix} 2 & 9 \\ 9 & 7 \end{pmatrix},\quad \boldsymbol{C} = \begin{pmatrix} 2 & 9 \\ 8 & 8 \end{pmatrix} \tag{10}$$

From $n_{i\%\%} + n_{\%j\%} + n_{\%\%k} - 2n_{\%\%\%} \leq n_{ijk} - \bar{n}_{ijk}$ results the *1-dimensional Fréchet lower bound* of Fienberg (1999):

$$n_{ijk} \geq n_{i\%\%} + n_{\%j\%} + n_{\%\%k} - 2n_{\%\%\%} \tag{11}$$

This bound is redundant with respect to the lower F-bounds. This is demonstrated as follows:

$$\begin{aligned} 1/3[(n_{ij\%}+n_{i\%k}-n_{i\%\%}) + (n_{ij\%}+n_{\%jk}-n_{\%j\%}) + (n_{i\%k}+n_{\%jk}-n_{\%\%k})] - (n_{i\%\%}+n_{\%j\%}+n_{\%\%k}-2n_{\%\%\%}) \\ = 2/3(\bar{n}_{\%22} + \bar{n}_{2\%2} + \bar{n}_{22\%}) \geq 0 \end{aligned} \tag{12}$$

Thus, at least one of the three candidate lower F-bounds must equal or exceed the 1-dimensional Fréchet lower bound, which renders the 1-dimensional Fréchet lower bound redundant.

Therefore, the *Fréchet-Bonferroni bounds* of Fienberg (1999) can be replaced by the F-bounds augmented

by the 2-dimensional Bonferroni upper bound, and expressed as:

$$\min \{n_{\%\%\%} \& n_{i\%\%} \& n_{\%j\%} \& n_{\%\%k} \% n_{ij\%} \% n_{i\%k} \% n_{\%jk} \ , \ n_{\%jk} \ , \ n_{i\%k} \ , \ n_{ij\%}\} \ \$ \ n_{ijk} \ \$$$
$$\max \{0 \ , \ n_{ij\%} \% n_{i\%k} \& n_{i\%\%} \ , \ n_{ij\%} \% n_{\%jk} \& n_{\%j\%} \ , \ n_{i\%k} \% n_{\%jk} \& n_{\%\%k}\} \quad (13)$$

We return to the example Fienberg (1999, Table 1). Here, the 2-dimensional Bonferroni upper bound (9) is not effective for any entry, and can be ignored. Thus, in this example, the Fienberg (1999) bounds (13) are identical to the F-bounds (5), and should yield identical results. They in fact do not, most likely due to numeric error somewhere in Fienberg (1999). This discrepancy is need to be kept in mind as we compare computational approaches below, that of Fienberg, using (13), and an alternative using (5).

Fienberg (1999) computes the Fréchet bounds, without using the Bonferroni bound (9), resulting in Fienberg (1999, Table 7). We applied the F-bounds (5), but in place of his Table 7, obtained sharper bounds. Both sets of bounds are presented in Table 3, as follows. If our bound agrees with Fienberg (1999, Table 7), we present the bound. If there is disagreement, we include the Fienberg (1999, Table 7) bound in parentheses alongside ours.

	INCOME					
	High	*Med*	*Low*	*High*	*Med*	*Low*
White	(80)85&107	(53)64&80	(142)158&169	175&197	(113)119&135	(32)43&54
Black	0&21	0&14	0&9	0&21	0&14	0&9
Chines	0&1	0&2	0&2	0&1	0&1	0&1
	MALE			*FEMALE*		

Table 3: F-Bounds and Fienberg (1999, Table 7) Fréchet Bounds for Table 1

Fienberg (1999) next applies the Bonferroni upper bound (9) to Table 7, and reports improvement in five cells, resulting in a Table 8, which is the table of exact bounds for the entire table. We were unable to reproduce these results; indeed, the Bonferroni bound provides improvement for none of the entries. Neither (5) nor (13) produced exact bounds for Table 1. We return to this example in a later sub-section.

3.2 The Shuttle Algorithm of Buzzigoli and Giusti (1999)

Buzzigoli and Giusti (1999) present an iterative algorithm called the *shuttle algorithm.* It is based on two *principles of subadditivity*:

- a sum of lower bounds on entries is a lower bound for the sum of the entries, and
  a sum of upper bounds on entries is an upper bound for the sum of the entries;
- the difference between the value (or an upper bound on the value) of an aggregate and a lower bound on the sum of all but one entry in the aggregate is an upper bound for that entry, and
  the difference between the value (or a lower bound on the value) of an aggregate and an upper bound on the sum of all but one entry in the aggregate is a lower bound for that entry.

The Buzzigoli-Giusti shuttle algorithm begins with the nominal lower and upper bounds for each entry. For each entry and each 2-dimensional marginal total containing the entry, the sum of current upper bounds of all other entries contained in the 2-dimensional marginal is subtracted from the marginal. This is a candidate lower bound for the entry. If the candidate improves the current lower bound, it replaces it. This is followed by an analogous procedure using sums of lower bounds, replacing current upper bounds with candidates whenever there is improvement. This two-step procedure is repeated until all bounds are stationary. The authors fail to note but it is evident that stationarity is reached in a finite number of iterations because the marginals are integer.

3.3 Comparative Analysis of the Fienberg (1999), Shuttle and F-Bounding Methods

It is worthwhile to compare the procedure of Fienberg (1999), the shuttle algorithm and the F-bounds. As previously observed, the Fienberg (1999) bounds can be reduced to the F-bounds plus the 2-dimensional Bonferroni upper bound, viz., (13). The shuttle algorithm produces bounds at least as sharp as the F-bounds, for two reasons. First, the iterative shuttle procedure enables improved lower bounds to improve the nominal and subsequent upper bounds. Second, lower F-bounds can be no sharper than those produced during the first set of steps of the shuttle algorithm. To see this, for concreteness consider the candidate lower F-bound $n_{i\%k} + n_{\%jk} - n_{\%\%k}$ for n_{ijk}. One of the three candidate shuttle lower bounds for n_{ijk} equals $n_{i\%k} - \sum_{J..j} n^{U}_{iJk}$, where n^{U}_{iJk} denotes the current upper bound for n_{iJk}. Thus, $n^{u}_{iJk} \le n_{\%Jk}$ and therefore

$\sum_{J..j} n^{U}_{\%Jk} \le \sum_{J..j} n_{\%Jk} - n_{\%\%k} + n_{\%jk}$. Consequently, the shuttle candidate lower bound is greater than or equal to

$n_{i\%k} - (n_{\%\%k} - n_{\%jk}) \le n_{i\%k} + n_{\%jk} - n_{\%\%k}$, so the shuttle candidate is at least as sharp as the F-candidate. If the shuttle algorithm is employed, all but the nominal lower Fienberg (1999) and F-bounds (namely, 0) are redundant.

Buzzigoli and Giusti (1999) illustrate the 3-dimensional shuttle algorithm for the case of a 2x2x2 table. It is not clear, for the general case of a $d_1 x d_2 x d_3$ table, if they intend to utilize the collapsing procedure of Table 2, but in what follows we assume that they do. Consider the 2-dimensional Bonferroni upper bounds (9). From (6), the Bonferroni upper bound for n_{ijk} equals $n_{ijk} + \bar{n}_{ijk}$. Consider the right-hand 2-dimensional table in Table 2. Apply the standard 2-dimensional lower F-bound to the entry in the upper left-hand corner:

$$\sum_{K...k} n_{ijK} \ge \sum_{K...k} n_{ijK} - \bar{n}_{ijk} \ge ((n_{i\%\%} - n_{i\%k}) + (n_{\%j\%} - n_{\%jk}) - (n_{\%\%\%} - n_{\%\%k})) \tag{14}$$

As previously observed, the shuttle algorithm will compute this bound during step 1and, if it is positive, replace the nominal lower bound with it, or with something sharper. During step 2, the shuttle algorithm will use this lower bound (or something sharper) to improve the upper bound on n_{ijk}, as follows:

$$n_{ijk} \le n_{ij\%} - \sum_{k...K} n_{ijK} \le n_{ij\%} - ((n_{i\%\%} - n_{i\%k}) + (n_{\%j\%} - n_{\%jk}) - (n_{\%\%\%} - n_{\%\%k})) \le n_{\%\%\%} - n_{i\%\%} - n_{\%j\%} - n_{\%\%k} + n_{ij\%} + n_{i\%k} + n_{\%jk} \tag{15}$$

Thus, the Fienberg (1999) 2-dimensional Bonferroni upper bound is redundant relative to the shuttle algorithm. Consequently, if the shuttle and collapsing methodologies are applied in combination, it suffices to begin with the nominal bounds and run the shuttle algorithm to convergence. Application of this approach (steps 1-2-1) to Table 1 yields Table 4 of exact bounds:

	INCOME					
	High	*Med*	*Low*	*High*	*Med*	*Low*
White	85&107	64&79	158&168	175&197	120&135	44&54
Black	0&21	0&14	0&9	0&21	0&14	0&9
Chinese	0&1	1&2	1&2	0&1	0&1	0&1
	MALE			*FEMALE*		

Table 4: Exact Bounds For Table 1 From the Shuttle Algorithm

3.4 Limitations of all Three Procedures

Although the shuttle algorithm produced exact bounds for this example, the shuttle algorithm, and consequently the Fienberg (1999) procedure and the F-bounds, are inexact, as follows. Cox (2000, Examples 7b,c) are

3-DIPTP exhibiting one or more non-integer continuous exact bounds. Because it is based on iterative improvement of integer upper bounds, in these situations the shuttle algorithm can come no closer than one unit larger than the exact integer bound, and therefore is incapable of achieving the exact integer bound.

In addition, the shuttle algorithm is not new: it was introduced by Schell (1955), who presented purported sufficient conditions on the 2-dimensional marginals to assure feasibility of 3-DPTP. Moravek and Vlach (1967) demonstrated that the *Schell conditions* are necessary, but not sufficient, for the existence of a solution to the 3-DPTP by means of the 5x8x2 3-dimensional table given in (16).

The counterexample (16) is applicable here. Each 1-dimensional Fréchet lower bounds is negative, thus not effective. Each Bonferroni upper bound is too large to be sharp. No lower F-bound is positive. Iteration of the shuttle produces no improvement. Therefore, each procedure yields nominal lower (0) and upper (1) bounds for each entry. Each procedure converges. Consequently, all three procedures produce seemingly correct bounds when in fact no table exists, viz., all three procedures are insensitive to infeasibility. A related but simpler counterexample, Cox (1999, Example 2), given by (17), appears in the next section.

$$\boldsymbol{A} = \begin{pmatrix} 1 & 4 \\ 1 & 4 \\ 1 & 4 \\ 4 & 1 \\ 4 & 1 \\ 4 & 1 \\ 4 & 1 \\ 3 & 2 \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 1 & 7 \\ 2 & 6 \\ 7 & 1 \\ 6 & 2 \\ 6 & 2 \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \tag{16}$$

4. THE ROEHRIG et al. (1999) AND CHOWDHURY et al. (1999) NETWORK MODELS FOR BOUNDS

Roehrig et al. (1999) (the third conference paper) and Chowdhury et al. (1999) (the *Management Science* paper) offer network models for computing exact bounds on internal entries in 3-dimensional tables. Network models are extremely convenient and efficient, and most important enjoy the *integrality property*, viz., integer constraints (viz., 2-dimensional marginal totals) assure integer optima. Network models provide a natural mechanism and language in which to express 2-dimensional tables, but most generalizations beyond 2-dimensions are apt to fail, as follows.

Roehrig et al. (1999) construct a network model for 2x2x2 tables and claim that it can be generalized to all 3-dimensional tables. This must be false. Cox (2000) showed that the class of 3-dimensional tables (of size $d_1 x d_2 x d_3$) representable as a network is the set of tables for which $d_i < 3$ for at least one index i. This is true because, if all $d_i \geq 3$, it is possible to construct a 3-dimensional table with integer marginals whose corresponding polytope has non-integer vertices (Cox 2000, Roehrig 1999), which contradicts the integrality property.

Chowdhury et al. (1999) address the following problem related to 3-DIPTP, also appearing in Roehrig et al. (1999). Suppose that the NSO releases only two of the three sets of 2-dimensional marginal totals (**A** and **B**), but not the third set (**C**) or the internal entries n_{ijk}. Is it possible to obtain exact lower and upper bounds for the remaining marginals (**C**)? The authors construct d_3 independent networks corresponding to the 2-dimensional tables defined by the appropriate (**A**, **B**) pairs and provide a procedure for obtaining exact bounds.

Unfortunately, this problem is quite simple and can be solved directly without recourse to networks or other mathematical formalities. In particular, the F-bounds of Section 2 suffice, as follows. Observe that, absent the **C**-constraints, the minimum (maximum) feasible value of a **C**-marginal total $\mathbf{C_{ij}}$ equals the sum of the minimum (maximum) values for the corresponding internal entries n_{ijk}. As the n_{ijk} are subject only to 2-dimensional constraints within their respective k-planes, then exact bounds for each n_{ijk} are precisely its 2-dimensional lower and upper F-bounds. These can be computed at sight and added along the k-dimension thus producing the corresponding **C**-bounds without recourse to networks or other formulations.

In addition, the Chowdhury et al. (1999) method is insensitive to infeasibility. This again is demonstrated by the Moravek and Vlach (1967) example (16): all Fréchet lower and nominal upper bounds computed in the two 2-dimensional tables defined by k = 1 and k = 2 contain the corresponding $\mathbf{C_{ij}}$ (equal to 1 in all cases), but there is no underlying table as the problem is infeasible. Insensitivity is also demonstrated by Cox (1999, Example 2), presented at the 1998 conference, viz.,

$$\boldsymbol{A} \text{ ' } \boldsymbol{B} \text{ ' } \boldsymbol{C} \text{ ' } \begin{pmatrix} 1 & 1 & 3 \\ 3 & 1 & 1 \\ 1 & 3 & 1 \end{pmatrix} \tag{17}$$

5. FRACTIONAL EXTREMAL POINTS

Cox (2000, Theorem 4.3) demonstrates that the only multi-dimensional integer planar transportation problems (*m-DIPTP*) for which integer extremal points are assured are those of size 2^{m-2}xbxc. In these situations,

linear programming methods can be relied upon to produce exact integer bounds on entries, and to do so in a computationally efficient manner even for problems of large dimension or size. The situation for other *integer problems of transportation type*, viz., m-dimensional contingency tables subject to k-dimensional marginal totals, k = 0, 1, ..., m-1, is quite different: until a direct connection can be demonstrated between exact continuous bounds on entries obtainable from linear programming and exact integer bounds on entries, linear programming will remain an unreliable tool for solving multi-dimensional problems of transportation type. Integer programming is not a viable option for large dimensions or size or repetitive use. Methods that exploit unique structure of certain subclasses of tables are then appealing, though possibly of limited applicability.

Dobra and Fienberg (2000) present one such method, based on notions from mathematical statistics and graph theory. Given an m-dimensional integer problem of transportation type and specified marginal totals, if these marginals form a *set of sufficient statistics* for a specialized log-linear model known as a *decomposable graphical model*, then the model is feasible and exact integer bounds can be obtained from straightforward formulae. These formulae yield, in essence, the F- and Bonferroni bounds. The reader is referred to Dobra and Fienberg (2000) for details, Bishop et al. (1975) for details on log-linear models, and Lauritzen (1996) for development of graphical models. The m-dimensional planar transportation problem considered here, m > 2, corresponds to the *no m-factor effect* log-linear model, which is not graphical, and consequently the Dobra-Fienberg method does not apply to problems considered here.

The choice here and perhaps elsewhere of the 3- or m-DIPTP as the initial focus of study for bounding problems is motivated by the following observations. If for reasons of confidentiality the NSO cannot release the full m-dimensional tabulations (viz., the m-dimensional marginal totals), then its next-best strategy is to release the (m-1)-dimensional marginal totals, corresponding to the m-DIPTP. If it is not possible to release all of these marginals, perhaps the release strategy of Chowdhury et al. (1999) should be investigated. Alternatively, release of the (m-2)-dimensional marginals might be considered. This strategy for release is based on the principle of releasing the most specific information possible without violating confidentiality. Dobra-Fienberg offers a different approach, namely, a class of marginal totals, perhaps of varying dimension, that can be released while assuring confidentiality through easy computation of exact bounds on suppressed internal entries.

A formula driven bounding method is valuable for large problems and for repetitive, large scale use. Consider the m-dimensional integer problem of transportation type specified by its 1-dimensional marginal totals. In the statistics literature, this is known as the *complete independence* log-linear model (Bishop et al. 1975). This

model, and in particular the 3-dimensional complete independence model, is graphical and decomposable. Thus, exact bounding can be achieved using Dobra-Fienberg.

Such problems can exhibit non-integer extremal points. For example, consider the 3x3x3 complete independence model with 1-dimensional marginal totals given by the vector:

$$(\sum_{i,j} n_{ijk}) \ ' \ (\sum_{i,k} n_{ijk}) \ ' \ (\sum_{j,k} n_{ijk}) \ ' \ (2 \ \ 1 \ \ 2) \tag{18}$$

Even though all continuous exact bounds on internal entries in (18) are integer, one extremal point at which n_{312} is maximized (at 1) contains four non-integer entries and another contains six. Bounding using linear programming would only demonstrate that $n_{312} \ ' \ 1$ is the continuous, not the integer, maximum if either of these extremal points were exhibited. A strict network formulation is not possible because networks exhibit only integer extremal points (although use of networks with side constraints is under investigation). Integer programming is undesirable for large or repetitive applications. Direct methods such as Dobra-Fienberg may be required. A drawback of Dobra-Fienberg is that it applies only in specialized cases.

6. DISCUSSION

In this paper we have examined the problem of determining exact integer bounds for entries in 3-dimensional integer planar transportation problems. This research was motivated by previous papers presenting heuristic approaches to similar problems that failed in some way to meet simple criteria for performance or reasonableness. We examined these and other approaches analytically and demonstrated one's superiority. We demonstrated that this method is imperfect and a reformulation of a method from the operations literature of the 1950s. We demonstrated that these methods are insensitive to infeasibility and can produce meaningless results otherwise undetected. We demonstrated that a method purported to generalize from 2x2x2 tables to all 3-dimensional tables could not possibly do so. We demonstrated that a problem posed and solved using networks in a *Management Science* paper can be solved by simple, direct means without recourse to mathematical programming. We illustrated the relationship between computing integer exact bounds, the presence of non-integer extremal points and the applicability of specialized mathematical programming formulations such as networks.

National statistical offices rely on automated methods for statistical operations including estimation, tabulation, quality assurance, imputation, rounding and disclosure limitation. Algorithms implementing these

methods must converge to meaningful quantities. In particular, these procedures should not report meaningless, misleading results such as seemingly correct bounds on entries in tables for which no feasible values exist. These risks are multiplied in statistical data base settings where data are often combined from different sources. Methods reported elsewhere and discussed here for computing bounds on suppressed internal entries in 3-dimensional contingency tables fail this requirement. This is because these methods are heuristic and based on necessary, but not sufficient, conditions for the existence of a solution to the 3-dimensional planar transportation problem. In addition, most of these methods fail to ensure exact bounds, and are incapable of identifying if and when they do in fact produce an exact bound. Consequently, nothing is gained by extending these methods to higher dimensions.

The first of these problems would be solved if sufficient conditions for the solution of the 3-DPTP were identified. Based on past attempts, this is a formidable problem, but certainly still worth investigating. Failing that, heuristic methods could be combined with linear programming or the method of Cox (2000) to determine first if a feasible table exists. The heuristic method then can be applied to produce (weak) bounds.

The second problem, assuring exact bounds, remains open and is complicated by the existence of non-integer optima, e.g., (18). One approach would be to explore the *integer rounding property* (Schrivjer 1986). Methods considered here as well as previous methods have worked from the *outside-in* to bound internal cells by exploiting subadditive relations between bounds to improve upon weak lower and upper bounds. Outside-in approaches are stymied by noninteger optima. The feasibility test of Cox (2000) works from the *inside-out*, viz., expands intermediate solutions outwards towards the boundary of the linear constraints. A potential statistical approach to the m-dimensional bounding problem would also work from the inside-out, as follows.

The NSO is aware of one feasible integer solution to the m-dimensional transportation problem, namely, the original counts that generated the m-dimensional table. Starting with this feasible integer solution, *Markov chain Monte Carlo* methods can be used to draw a large sample of integer feasible solutions. Methods based on objects from algebraic geometry known as *Gröbner bases* ensure that integer feasible solutions can be selected from a uniform distribution on the set of all integer feasible solutions (Diaconis and Sturmfels 1998). The sample is used as an empirical distribution of feasible integer solutions. Consequently, the NSO can estimate, say, a 95% credible region for feasible integer values of each internal entry n_{ijk}. The NSO may choose to use the credible intervals as credible estimates of the true exact integer bounds, or as constraints for a tree search or integer programming analysis. Note that in order to use this approach the attacker requires a feasible integer starting value. This approach is in the realm of sampling contingency tables with fixed marginal totals, a difficult and computationally

demanding area but one of current interest in the setting of Markov chain Monte Carlo analysis. Further developments on computing Gröbner bases would pay benefits in both MCMC and the current setting.

Acknowledgments

The information and options expressed in this article are the work of the author and should not be interpreted as representing the policies or practices of the U.S. National Center for Health Statistics, Centers for Disease Control and Prevention. Constructive comments of Fred Glover on an earlier draft are acknowledged with appreciation.

REFERENCES

Bishop, Y., S. Fienberg, P. Holland 1975. Discrete Multivariate Analysis: Theory and Practice, MIT Press, Cambridge.
Buzzigoli, L., A. Giusti, 1999, An algorithm to calculate the lower and upper bounds of the elements of an array given its marginals, In: Statistical Data Protection: Proceedings of the Conference, EUROSTAT, Luxembourg, pp. 131-147.
Chowdhury, S., G. Duncan, R. Krishnan, S. Roehrig, S. Mukherjee, 1999, Disclosure detection in multivariate categorical databases: auditing confidentiality protection through two new matrix operators,. Management Science 45 1710-1723.
Cox, L., 1999, Invited talk: some remarks on research directions in statistical data protection, In: Statistical Data Protection: Proceedings of the Conference, EUROSTAT, Luxembourg, pp. 163-176.
Cox, L., 2000, On properties of multi-dimensional statistical tables, Technical report, April 2000, 29 pp.
Diaconis, P., B. Sturmfels, 1998, Algebraic algorithms for sampling from conditional distributions, Annals of Statistics 26 363-397.
Dobra, A., S. Fienberg, 2000, Bounds for cell entries in contingency tables given marginal totals and decomposable graphs, Proceedings of the National Academy of Sciences 97 11185-11192.
Fienberg, S., 1999, Fréchet and Bonferroni bounds for multi-way tables of counts with applications to disclosure limitation, In: Statistical Data Protection: Proceedings of the Conference, EUROSTAT, Luxembourg, pp. 115-129.
Lauritzen, S. 1996. Graphical Models, Clarendon Press, Oxford.
Moravek, J., M. Vlach, 1967, On necessary conditions for the existence of the solution to the multi-index transportation problem, Operations Research 15 542-545.
Roehrig, S., 1999, Auditing disclosure in multi-way tables with cell suppression: simplex and shuttle solutions, manuscript, Heinz School of Public Policy & Management, Carnegie Mellon University, Pittsburgh, 21pp.
Roehrig, S., R. Padman, G. Duncan, R. Krishman, 1999, Disclosure detection in multiple linked categorical datafiles: a unified network approach, In: Statistical Data Protection: Proceedings of the Conference, EUROSTAT, Luxembourg, pp. 149-162.
Rüschendorf, L., B. Schweizer, M.D. Taylor (eds.)1996. Distributions with Fixed Marginals and Related Topics, IMS Lecture Notes–Monograph Series, Vol. 28, Institute of Mathematical Statistics, Hayward, CA.
Schell, E., 1955, Distribution of a product over several properties, In: Proceedings, 2nd Symposium on Linear Programming. Washington, DC, pp, 615-642.
Schrivjer, A. 1986. Theory of Linear and Integer Programming, John Wiley & Sons, New York.
Vlach, M., 1986, Conditions for the existence of solutions of the three-dimensional planar transportation problem, Discrete Applied Mathematics 13 61-78.

Original: July 13, 2000
Revised: December 3, 2001