Fréchet and Bonferroni Bounds for Multi-way Tables of Counts With Applications to Disclosure Limitation

Stephen E. Fienberg
Department of Statistics, Carnegie Mellon University
Pittsburgh, PA 15213, USA.

Abstract: Upper and lower bounds on cell counts in cross-classifications of positive counts play important roles in a number of the disclosure limitation procedures, e.g., cell suppression and data swapping. Some features of the Fréchet bounds are well-known, intuitive, and are regularly used by those working on disclosure limitation methods, especially those for two-dimensional tables. The multivariate versions of these bounds and other related bounds such as those calculated using the Bonferroni approach are more complex, however, but they have potentially great import for current disclosure limitation methodology. The purpose of this paper is to describe the key results on this topic.

Keywords: Contingency tables; Copulas; Loglinear models; Marginal bounds.

1. Introduction

Upper and lower bounds on cell counts in cross-classifications of positive counts play important roles in a number of the disclosure limitation procedures, e.g., see the discussion in Cox [6] [7], Dalenius and Reiss [8], Duncan and Fienberg [11], Fienberg [12], Fienberg et al. [14], and Fischetti and Salazar [15]. The purpose of this paper is to introduce the key results on Fréchet and Bonferroni bounds and to link them to problems in disclosure limitation. For related details on bounds, see Joe [22], Kwerel [23], Rüschendorf [28] [29], and Warmuth [32].

For most problems of interest in the disclosure limitation context, we begin by knowing that there exists a table with the known marginals, i.e., it is the one for whose entries we want to "protect." Thus many of the deep probabilistic results about the existence of multi-dimensional distributions with fixed marginals are not of direct concern to us. While the terminology we will use in this paper is that for tables of counts, i.e., "unweighted" contingency tables, virtually all of the results are applicable to general positive or "weighted" tables of counts.

The class of bounds we focus largely upon is usually attributed to Fréchet [16] (thus the name Fréchet bound), but Rüschendorf et al. [30] suggest co-attribution to Hoeffding [21]. The simplicity of the bounds has led many others to rediscover them repeatedly in the context of data security and contingency tables (e.g., see Gusfield [19]). Fréchet's original presentation (and many subsequent ones) was in terms of cumulative distribution functions (c.d.f.) for a random vector $(D_1, D_2, \ldots, D_n)$ in R^k:

$$F_{1,2,\ldots,k}(x_1, x_2, \ldots, x_k) = \Pr(D_1 \le x_1, D_2 \le x_2, \ldots, D_n \le x_k), \tag{1}$$

which are essentially equivalent to contingency tables when the underlying variables are categorical. For example, suppose we have a two-dimensional table of counts, $\{n_{ij}\}$ adding up to the total $n_{++} = n$. If we normalize each entry by dividing by n and then create a table of partial sums, by cumulating the proportions from the first row and first column to the present ones, we have a set of values of the form (1). Thus, for the purposes of converting the Fréchet bound results for distribution functions to those for tables of counts, the values $\{x_i\}$ in (1) represent"cut-points" between categories for the i-th categorical variable.

We present the bound results here in their contingency table form (c.f., Fréchet's presentation in [17]) for tables of counts in large part to emphasize close linkages to the theory of loglinear models for the analysis of contingency table data. Further, we illustrate the results using an example of data from the 1990 U.S. decennial census public use sample for a local area,in the form of a $3 \times 3 \times 2$ table of counts given in Table 1, as well as a collapsed $2 \times 2 \times 2$ table version of it given in Table 2. Table 1 has some noteworthy features. First, it includes three counts of "1", or sample uniques. Second, there are counts of "1" in two of the three two-way marginal totals. Thus, if we think in terms of constraining the interior cells of the table given the margins, we can expect to get tight bounds for some of the cell entries.

Gender = Male

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

Gender = Female

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

Table 1: Three-way cross-classification of Gender, Race, and Income for a selected U.S. census tract. (*Source*: 1990 Census Public Use Microdata Files)

2. Some Preliminary Considerations for $2 \times 2 \times 2$ Tables

We begin by illustrating the basic bound notions using the collapsed $2 \times 2 \times 2$ table in Table 2. If we consider layers 1 and 2 separately, then we have a pair of 2×2 tables. Thus the simple Fréchet bounds, given by

$$\min\{n_{i+}, n_{+j}\} \geq n_{ij} \geq \max\{n_{i+} + n_{+j} - n, 0\}, \tag{2}$$

Male

Race	Income Level $\leq \$10,000$	$> \$10000$	Total
White	96	233	329
Black/Chinese	11	16	27
Total	107	249	356

Female

Race	Income Level $\leq \$10,000$	$> \$10000$	Total
White	186	178	364
Black/Chinese	11	11	22
Total	197	189	386

Table 2: Collapsed $2 \times 2 \times 2$ version of cell counts in Table 1.

for $i, j = 1, 2$, are directly applicable. These bounds, given in Table 3, in effect fix the entries in two of the three 2-way margins of the full $2 \times 2 \times 2$ table.

Male

Race	Income Level $\leq \$10,000$	$> \$10000$	Total*
White	107,80	249,222	329
Black/Chinese	27,0	27,0	27
Total*	107	249	356

Female

Race	Income Level $\leq \$10,000$	$> \$10000$	Total*
White	197,175	189,167	364
Black/Chinese	22,0	22,0	22
Total*	197	189	386

Table 3: Fréchet bounds fixing the 1-way margins for each layer of Table 2.

Next we consider fixing all three 2-way margins. This problem has a simple generic form. In effect, we are given 7 values: the sums for each of the (1,1) cells of the three 2-way margins, the sums for the 1st entry in each of the three 1-way margins and the grand total. All of the other marginal values can be computed from these. Thus we need only one more quantity to determine the entries of the full table. Let x be the true but unknown value of the count in the $(1, 1, 1)$ cell.

Male

Race	Income Level $\leq \$10,000$	$> \$10000$	Total*
White	107, 85	244, 222	329
Black/Chinese	22, 0	27, 5	27
Total*	107	249	356

Female

Race	Income Level $\leq \$10,000$	$> \$10000$	Total*
White	197, 175	189, 167	364
Black/Chinese	22, 0	22, 0	22
Total*	197	189	386

Table 4: Upper and lower bounds for entries in Table 2 given all three 2-way margins.

We thus have:

$$\begin{aligned} n_{111} &= x \\ n_{121} &= n_{1+1} - x \\ n_{112} &= n_{11+} - x \\ n_{211} &= n_{+11} - x \\ n_{122} &= n_{1++} - n_{1+1} - n_{11+} + x \\ n_{212} &= n_{+1+} - n_{11+} - n_{+11} + x \\ n_{122} &= n_{1++} - n_{+11} - n_{1+1} + x \\ n_{122} &= n - n_{1++} - n_{+1+} - n_{++1} + n_{11+} + n_{1+1} + n_{+11} - x. \end{aligned} \tag{3}$$

Now if we add the non-negativity constraint for cell counts in a contingency table, i.e., $n_{ijk} \geq 0$ for $i, j, k = 1, 2$, we get 4 upper bounds and 4 lower bounds. Three of the 4 upper bounds components involve the 2-way marginal totals corresponding to the (1,1,1) cell of the table and the fourth can be written as the sum of the diagonally opposite or complementary cells, i.e.,

$$n - n_{1++} - n_{+1+} - n_{++1} + n_{11+} + n_{1+1} + n_{+11} = n_{111} + n_{222}. \tag{4}$$

The result is the following bounds on x:

$$\begin{aligned} &\min\{n_{11+}, n_{1+1}, n_{+11}, n_{111} + n_{222}\} \geq x \\ &\quad \geq \max\{n_{1++} - n_{1+1} - n_{11+}, n_{+1+} - n_{11+} - n_{+11}, n_{1++} - n_{+11} - n_{1+1}, 0\}. \end{aligned} \tag{5}$$

For the counts in Table 2, the upper and lower bounds of the form (5) yield the values in Table 4.

Despite the existence of explicit upper and lower bounds in the case of the $2 \times 2 \times 2$ contingency table with fixed 2-way margins various authors have suggested the need to resort to linear programming and other indirect methods to find the tightest possible bounds (e.g., see de Vries [31] and Chowdhury et al. [5]). In the following sections, we address the formal structure of the types of bounds explored here and we give formulas that allow the computation of sharp bounds in many problems of interest, thus obviating the need for linear programming or even network models. In the second last section we consider the relationship between these bounds and a simple alternation scheme proposed by Buzzigoli and Giusti [4], which they conjectures requires only a finite number of steps, as well as a network algorithm proposed by Roehrig et al. [26].

3. 1-Dimensional Fréchet Bounds for k-Way Tables

Consider a k-way contingency table with entries $\{n_{i_1 i_2 \ldots i_k}\}$ where $i_j = 1, 2, \ldots, I_j$ for $j = 1, 2, \ldots, k$. Then if we know the 1-way margins, the entries in the tables are bounded above and below by:

$$\begin{aligned} &\min\{n_{i_1+\cdots+}, n_{+i_2+\cdots+}, \ldots, n_{+\cdots+i_k}\} \\ &\quad \geq \quad n_{i_1 i_2 \ldots i_k} \geq \max\{n_{i_1+\cdots+} + n_{+i_2+\cdots+} + \cdots + n_{+\cdots+i_k} - n(k-1), 0\}. \end{aligned} \tag{6}$$

This result, which also coincides with the standard Bonferroni bounds [2] for the cell counts, can be found in a variety of sources (e.g., see Kwerel [23], Warmuth [32], or Rüschendorf [28]). Setting $k = 2$ in (6) yields the 1-dimensional bounds for 2-way tables given in equation (2), which we rewrite following Mardia [25] as

$$\frac{1}{2}\{n_{i+} + n_{+j} - \mid n_{i+} - n_{+j} \mid\} \geq n_{ij} \geq \frac{1}{2}\{n_{i+} + n_{+j} - n + \mid n_{i+} + n_{+j} - n \mid\}. \tag{7}$$

Note that the lower bound in (6) involves the sum of all of the 1-way margins associated with the cell $(i_1, i_2, \ldots, i_k)$. Because we will have occasion to use such sums again in other bounds we label this as $S_{1[i_1 i_2 \ldots i_k]}$, and rewrite (6) as

$$\begin{aligned} &\min\{n_{i_1+\cdots+}, n_{+i_2+\cdots+}, \ldots, n_{+\cdots+i_k}\} \\ &\quad \geq \quad n_{i_1 i_2 \ldots i_k} \geq \max\{S_{1[i_1 i_2 \ldots i_k]} - n(k-1), 0\}. \end{aligned} \tag{8}$$

We note that the upper and lower bounds cannot be reached simultaneously and thus they do not add up to the corresponding marginal totals.

In Table 5 we illustrate the upper and lower Fréchet bounds for the cell counts in Table 1, where we use only the information in the 1-way margins, which are included here (in the lower layer) and denoted by an * for reference. The upper and lower bounds are the same in both layers, because the gender totals both exceed all other marginal totals. By applying the Fréchet bounds for 2-way tables separately for each layer of the table, we get the sharper bounds in Table 6. This approach fixes the 2-way margins for Race $\times$ Income and Race $\times$ Gender in the full 3-way table. Now there are some non-zero lower bounds in the first row of each layer only, as expected, and the bounds for the two layers differ.

The statistical and probabilistic literature on multidimensional distributions with given 1-dimensional marginals is now quite extensive and functions that describe such distributions are referred to as *copulas*, e.g., see DallÁglio et al. [9]. Further, the Fréchet bounds in this case are sharp, i.e. they are attained and thus there exist copulas which achieve them (see Rüschendorf [27] for details). For us, this means that there exist contingency tables with the appropriate marginal totals achieving the specified upper and lower bounds.

Other results for Fréchet bounds for contingency tables with given 1-way margins are also relatively easy to derive. For example, there are at most $(\sum_{j=0}^{k} I_j) - 1$ distinct upper bounds for

Gender = Male

Race	Income Level $\leq \$10,000$	$> \$10000$ and $\leq \$25000$	$> \$25000$	Total*
White	304,0	215,0	223,0	-
Black	44,0	44, 0	44, 0	-
Chinese	5,0	5, 0	5, 0	-
Total*	-	-	-	356

Gender = Female

Race	Income Level $\leq \$10,000$	$> \$10000$ and $\leq \$25000$	$> \$25000$	Total*
White	304, 0	215, 0	223, 0	693
Black	44, 0	44, 0	44, 0	44
Chinese	5, 0	5, 0	5, 0	5
Total*	304	135	54	386

Table 5: Fréchet bounds for entries in Table 1 given all 1-way margins. (The totals given in the table are for the 1-way margins.)

the $\prod_{j=0}^{k} I_j$ cells in the table, and the minimum in the lower bound can be nonzero only for those cells in the largest, row, column, layer, etc., provided that the corresponding sum is greater than $(k-1)n/k$. In Table 5 we see that only those cells in row 1 (Race = White) could have non-zero lower bounds because only their 1-way total is greater than $2(742)/3 = 495$, but because the other margins are more evenly spread, the realized lower bounds are still zero. Finally, if all of the cell counts are positive, the upper bound always exceeds the lower bound. For proofs of these results when $k = 2$, see Gusfield [19]. The generalizations for $k > 2$ are direct.

4. m-Dimensional Fréchet Marginal Bounds for k-Way Tables

The bounds associated with the fixing of the m-way margins of a k-way table are more complex than those of the preceding subsection and for the lower bounds we need to proceed through iteration, beginning with $(k-1)$-dimensional bounds, and doing successive substitution. Moreover the bounds take a different form depending on whether k is even or odd.

We begin with the upper bound, which is straight forward and intuitively exactly what one might guess it to be:

$$\min\{n_{i_1 i_2 \ldots i_{k-1}+}, n_{i_1 i_2 \ldots i_{k-2}+i_k}, \ldots, n_{+i_2 \ldots i_k}\} \geq n_{i_1 i_2 \ldots i_k}, \tag{9}$$

where the minimum is taken over all $(k-1)$-dimensional margins associated with the $(i_1, i_2, \ldots, i_k)$ cell. Clearly the upper bound gets tighter as m increases. Unfortunately, despite the claim in Warmuth [32], the upper bound does not always corresponds to an actual k-dimensional distribution function with the specified m-dimensional marginal distributions. Thus there does not always exists a contingency table specified by this upper bound with the appropriate marginal totals. Rüschendorf [28] gives alternative upper bounds based on the Bonferroni inequality which may be sharper, at least in some circumstances, and to which we return in Section 5..

Gender = Male

Race	Income Level $\leq \$10,000$	$> \$10000$ and $\leq \$25000$	$> \$25000$	Total*
White	107, 80	80, 53	169, 142	329
Black	23, 0	23, 0	23, 0	23
Chinese	4, 0	4, 0	4, 0	4
Total*	107	80	169	356

Gender = Female

Race	Income Level $\leq \$10,000$	$> \$10000$ and $\leq \$25000$	$> \$25000$	Total*
White	197, 175	135, 113	54, 32	364
Black	21, 0	21, 0	21, 0	21
Chinese	1, 0	1, 0	1, 0	1
Total*	197	135	54	386

Table 6: Upper and lower Fréchet bounds for entries in Table 1 using Race $\times$ Income and Race $\times$ Gender margins from the "conditional independence" model.

The upper bound idea in equation (9) extends immediately to bounds based on other collections of possible overlapping marginals and we describe these extensions and their relationship to the theory of loglinear models briefly in Section 7. below.

We now need some additional notation. Let $I_m = (i_1, i_2, \ldots, i_m)$ for $i_1 < i_2 < \cdots < i_m$ where $m = 1, 2, \ldots, k-1$. Further, let

$$S_{m[i_1 i_2 \ldots i_k]} = \sum_{I_m} n_{i_1 i_2 \ldots i_m + \cdots +}, \tag{10}$$

where the summation in (10) is over similar ordered m-tuples defining all of the m-dimensional marginal totals of $n_{i_1 i_2 \ldots i_k}$. Thus $S_{1[i_1 i_2 \ldots i_k]}$ is, as we defined above, the sum of all 1-dimensional marginal totals, etc. These sums of marginal totals play a crucial role in the lower bounds. Further, we let

$$\bar{n}_{i_1 i_2 \ldots i_k} = n - S_{1[i_1 i_2 \ldots i_k]} + S_{2[i_1 i_2 \ldots i_k]} - \ldots + (-1)^k n_{i_1 i_2 \ldots i_k}, \tag{11}$$

i.e., $\bar{n}_{i_1 i_2 \ldots i_k}$ is the *diagonally complementary* count opposite $n_{i_1 i_2 \ldots i_k}$ in the 2^k table formed by collapsing all of the remaining categories for each of the k variables in the table into a single complementary category (i.e., not i_j for $j = 1, 2, \ldots, k$). We can define marginals of the $\bar{n}_{i_1 i_2 \ldots i_k}$,

$$\bar{S}_{m[i_1 i_2 \ldots i_k]}, \tag{12}$$

by summing over subscripts as usual (although there is the need to remove a common multiplication factor resulting from the use of multiple collapsed cells). Because the $\{\bar{n}_{i_1 i_2 \ldots i_k}\}$ themselves form a table of counts, the bounds described in this paper hold for them as well.

The lower bound for the cell entries given by Warmuth [32] now depends on whether k is even or odd, and they result from a fairly direct application of Poincaré's theorem which is a basic inclusion/exclusion alternation result. Suppose first that $k = 2p$, i.e., k is even. Then we have that

$$n_{i_1 i_2 \ldots i_k} \geq \max\{S_{1[i_1 i_2 \ldots i_k]} - S_{2[i_1 i_2 \ldots i_k]} + \cdots + S_{\overline{k-1}[i_1 i_2 \ldots i_k]} - n, 0\}. \tag{13}$$

For $k = 2p + 1$, i.e., k is odd, and we have

$$\begin{aligned} n_{i_1, i_2, \ldots, i_k} \quad \geq \quad & \max\{n - S_{1[i_1 i_2 \ldots i_k]} + S_{2[i_1 i_2 \ldots i_k]} - \cdots \\ & + S_{\overline{k-1}[i_1 i_2 \ldots i_k]} - \min_j\{\bar{n}_{i_1 i_2 \ldots i_{j-1} + i_{j+1} \ldots i_k}\}, 0\}. \end{aligned} \tag{14}$$

To get the lower bound for fixed m-dimensional margins where $m \leq (k-1)$, we substitute for the marginal sums in the lower bounds alternating between versions equations (13) and (14) applied to successive margins of lower dimension. In Section 6., we illustrate this approach for 3-way tables. As with the upper bound, Rüschendorf [28] gives alternative lower bounds for fixed m-dimensional marginals derived based on a Bonferroni approach which may be sharper, at least in some circumstances. We now turn to this topic.

5. m-Dimensional Bonferroni Marginal Bounds

Unless we are in the special case of margins that correspond to decomposable loglinear models the bounds in the preceding section are not sharp, i.e., there is not necessarily a set of extremal tables corresponding to the bounds (see Section 7.). Rüschendorf [28] gives alternative bounds based on an approach that uses the same Bonferroni inequalities as are commonly used in statistics to derive simultaneous confidence intervals. Bonferroni's classic paper [2] appeared in 1936, prior to Fréchet [16] and Hoeffding [21], so perhaps we should be referring to Bonferroni-Fréchet-Hoeffding bounds. Galambos and Simonelli [18] give a detailed modern treatment of Bonferroni inequalities and their application.

The basic Bonferroni inequalities come from an inclusion-exclusion identity from probability. For independent sets A_i for $i = 1, 2, \ldots, k$, with $\bar{A}_i$ representing the complementary event, following Galambos and Simonelli [18] we can write:

$$P(\bigcap_{i=1}^{k} \bar{A}_i) = 1 - \sum_{j=1}^{n} kP(A_j) + \sum_{1 \leq i,j \leq k} P(A_i \cap A_j) - \cdots. \tag{15}$$

Then, for $m \leq k$, if we let

$$Q_k = \sum\nolimits^{*} P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_m}) \tag{16}$$

where the summation is over all $i_1 \leq i_2 \leq \cdots \leq i_m$, equation (15) becomes

$$P(\bigcap_{i=1}^{k} \bar{A}_i) = 1 - Q_1 + Q_2 - Q_3 + \cdots (-1)^k Q_k. \tag{17}$$

Clearly, we can set bounds on $P(\bigcap_{i=1}^{k} \bar{A}_i)$ by stopping at Q_m in equation (17). If m is even, then the result is a lower bound, whereas if m is odd, we get an upper bound.

Now if we reinterpret the Bonferroni equation in terms of counts instead of probabilities, we can rewrite equation (17) for the entry in the $(i_1, i_2, \ldots, i_k)$ cell, i.e., $n_{i_1 i_2 \cdots i_k}$, as

$$n_{i_1 i_2 \ldots i_k} = \sum_{m=0}^{k} (-1)^m \bar{S}_{m[i_1 i_2 \ldots i_k]} \tag{18}$$

where $\bar{S}_{m[i_1 i_2 \ldots i_k]}$ is defined above in equation (12) above. Truncating (18) at $\bar{S}_{m[i_1 i_2 \ldots i_k]}$ when m is even (odd) produces an upper (lower) bound. Suppose m is even (odd), then we also have a lower (upper) bound by truncating (18) at the $(m-1)$st term. Ruschendorf gives a result that sharpens this bound using a supremum over the class of minimal spanning trees involving subsets of the Qs, which turns out to be very helpful.

There is an intimate link between these bounds and maximum likelihood estimates under the loglinear model whose minimal sufficient statistics are the m-dimensional marginal totals, especially when $m = k-1$ and k is odd. We illustrate the link explicitly in the next section.

6. Re-examining Bounds for 3-Way Tables

¿From Section 3., the Fréchet bounds for a 3-way table of counts with given 1-way margins are:

$$\min\{n_{i++}, n_{+j+}, n_{++k}\} \geq n_{ijk} \geq \max\{n_{i++} + n_{+j+} + n_{++k} - 2n, 0\}. \tag{19}$$

Following Warmuth [32], we can derive (19) by substitution in the following bounds from Section 4.. Since $k = 3$ is odd we need to use equations (9) and (14):

$$\min\{n_{ij+}, n_{i+k}, n_{+jk}\} \geq n_{ijk}, \tag{20}$$

$$n_{ijk} \geq \max\{n - S_{1[ijk]} + S_{2[ijk]} - \min\{\bar{n}_{ij+}, \bar{n}_{+jk}, \bar{n}_{i+k}\}, 0\}. \tag{21}$$

Suppose the minimum in the lower bound in equation (21) occurs for the margins for variables 1 and 2 adding across 3, i.e.,

$$\min\{\bar{n}_{ij+}, \bar{n}_{+jk}, \bar{n}_{i+k}\} = n_{ij+} - n_{i++} - n_{+j+} + n. \tag{22}$$

Then the upper and lower bounds for fixed 2-dimensional margins take the form:

$$\min\{n_{ij+}, n_{i+k}, n_{+jk}\} \geq n_{ijk} \geq \max\{-n_{++k} + n_{i+k} + n_{+jk}, 0\}. \tag{23}$$

Recall that the Fréchet bounds have a useful substitution property as we move from higher dimensions to lower ones. Thus, by substituting in equation (23) using the bounds for 2-way margins given 1-way margins, we get equation (19).

In Table 7 we give the upper and lower Fréchet bounds for entries in Table 1 using all three 2-way margins. The upper values are somewhat tighter than those in Table 6, which were based only on a pair of 2-way margins, and there is a non-zero lower bound for the count of "2" in the last row of the first layer.

Since $k = 3$ and we are working with all margins of dimension $m = 2$, the Bonferroni approach from the previous section in this situation yields an extra component for the upper bound of the form:

$$n - n_{i++} - n_{+j+} - n_{++k} + n_{ij+} + n_{+jk} + n_{i+k} = n_{ijk} + \bar{n}_{ijk}. \tag{24}$$

This bound is explicitly the one applicable for 2^3 tables. When the size of a dimension (variable) exceeds 2, we need to consider all possible 2^3 tables that can be formed by partitioning the categories of these variables and then compute upper bounds of the form (24) for the sum including the (i, j, k) cell in each such partition.

The lower bounds from the simple Bonferroni approach involve only 1-way margins and thus are not helpful since the Fréchet lower bounds have already incorporated information from 2-way

Gender = Male

Race	Income Level $\leq \$10,000$	$> \$10000$ and $\leq \$25000$	$> \$25000$	Total*
White	107, 80	80, 53	169, 142	329
Black	21, 0	14, 0	9, 0	23
Chinese	1, 0	2, 0	2, 0	4
Total*	107	80	169	356

Gender = Female

Race	Income Level $\leq \$10,000$	$> \$10000$ and $\leq \$25000$	$> \$25000$	Total*
White	197, 175	135, 113	54, 32	364
Black	21, 0	14, 0	9, 0	21
Chinese	1, 0	1, 0	1, 0	1
Total*	197	135	54	386

Table 7: Upper and lower Fréchet bounds for entries in Table 1 using all three 2-way margins from the "no 2nd-order interaction" model.

margins. But the sharpening approach adds additional lower bounds by utilizes the margins associated with the three possible conditional independence models produced by dropping a single interaction term from the no 2nd-order interaction model.

The Bonferroni upper bound component in (24) is directly related to the a result on the existence of maximum likelihood estimates under the no 2nd-order interaction logliner model for a 2^3 contingency table. It is well-known that there is a special problem of "non-existence" for this model that occurs when the (1,1,1) and (2,2,2) cells contain sampling zeros and the remaining 6 cell counts are positive (e.g., see Haberman [20]). The bound in (24) would in this case be 0, and thus it is telling us that there exists only one table with the three sets of 2-way margins, i.e., the original table which contains sampling zeros in the (1,1,1) and (2,2,2) cells. It then follows that the 2-way margins reveal the contents of the remaining cells, and thus disclosure for all cells is exact. The same result holds for all pairs of diagonally complementary cells containing sampling zeros, and generalizes to all three-way tables via the partitioning idea mentioned above. There is a similar connection for the bounds of a k-way table given its $(k-1)$-way margins and the "non-existence" of the corresponding no kth-order interaction loglinear model.

In Table 8 we have the full sharp bounds on the counts in Table 1, given the three two-way marginals. The extra upper and lower values that come from the Bonferroni bound results narrow the differences between the upper and lower bounds in 5 of the 18 cells in the table.

7. Decomposable and Graphical Loglinear Models

The explicit role of marginal totals in the calculation of the bounds in the multi-way cases, suggests that their is a natural link with results on loglinear models. This is indeed the case. For example, the lower bound in equation (23) is working with the minimal sufficient statistics on a linear scale in a way that parallels the way that maximum likelihood estimates (MLEs) work

Gender = Male

Race	Income Level $\leq \$10,000$	$> \$10000$ and $\leq \$25000$	$> \$25000$	Total*
White	107, 85	79, 64	168, 158	329
Black	21, 0	14, 0	9, 0	23
Chinese	1, 0	2, 1	2, 1	4
Total*	107	80	169	356

Gender = Female

Race	Income Level $\leq \$10,000$	$> \$10000$ and $\leq \$25000$	$> \$25000$	Total*
White	197, 175	135, 120	54, 44	364
Black	21, 0	14, 0	9, 0	21
Chinese	1, 0	1, 0	1, 0	1
Total*	197	135	54	386

Table 8: Combined Fréchet and Bonferroni bounds for entries in Table 1 using all three two-way margins from the "no 2nd-order interaction" model.

with them multiplicatively. There is also a link here to the recent work on the exact distribution of a contingency table under a loglinear model given its minimal sufficient statistics, e.g., see Diaconis and Sturmfels [10] and Fienberg et al. [13].

Using very different language, Rüschendorf [28] describes bounds for "decomposable (regular) system[s]." These turn out to involve margins that would correspond to a *directly estimable* or *decomposable* loglinear model (e.g., see Bishop et al. [1] and Lauritzen [24]). Rüschendorf's approach presents a characterization of the class of all distributions with margins that correspond to decomposable MLEs as a mechanism for computing Fréchet-like bounds similar to those described above which are sharp. In the nondecomposable case, as in the preceding section, he uses bounds constructed from implied decomposable models. It remains to be seen if and how these results simplify for the class of graphical loglinear models, which includes the decomposable models as special cases.

Rüschendorf [29] has also made links between the theory of upper and lower bounds and generalizations of the iterative proportional scaling algorithm, which is widely used in contingency table estimation for loglinear models.

8. Alternative Approaches to Calculating Bounds

Buzzigoli and Giusti [4] have suggested a simple alternation scheme for computing the upper and lower bounds for a k-way contingency table given all $(k-1)$-way margins. Their heuristic argument is quite simple. For a given cell to take its maximum possible value, all others with which it is added to form a given marginal total should take their minimum possible values. Similarly, for a given cell to take its minimum possible value, all others with which it is added to form a given marginal total should take their maximum possible values. Thus they begin with a simple set of bounds and then alternately adjust these to take into account the estimated upper and lower constraints from the preceding step.

Let $n^U_{i_1\ldots i_k}$ and $n^L_{i_1\ldots i_k}$ be the desired *U*pper and *L*ower bounds, respectively. Begin with the upper Fréchet bounds from (9) and a lower bound of zero:

$$
\begin{aligned}
n^U_{i_1\ldots i_k} &= \min(n_{+i_2\ldots i_k}, n_{i_1+i_3\ldots i_k}, \ldots, n_{i_1\ldots i_{k-1}+}), && (25)\\
n^L_{i_1\ldots i_k} &= 0. && (26)
\end{aligned}
$$

Then the shuttle algorithm alternates between

$$
\begin{aligned}
n^U_{i_1\ldots i_k} = \min\{n_{+i_2\ldots i_k} &- \sum_{i\neq i_1} n^L_{ii_2\ldots i_k}, n_{i_1+i_3\ldots i_k} - \sum_{i\neq i_2} n^L_{i_1ii_3\ldots i_k},\\
&\ldots, n_{i_1\ldots i_{k-1}+} - \sum_{i\neq i_k} n^L_{i_1\ldots i_{k-1}i}\}, && (27)\\
n^L_{i_1\ldots i_k} = \max\{0, n_{+i_2\ldots i_k} &- \sum_{i\neq i_1} n^U_{ii_2\ldots i_k}, n_{i_1+i_3\ldots i_k} - \sum_{i\neq i_2} n^U_{i_1ii_3\ldots i_k},\\
&\ldots, n_{i_1\ldots i_{k-1}+} - \sum_{i\neq i_k} n^U_{i_1\ldots i_{k-1}i}\}. && (28)
\end{aligned}
$$

Buzzigoli and Giusti show that this algorithm produces the bounds in (5) for the $2 \times 2 \times 2$ table with given 2-way totals, and they conjecture that it works more generally, stopping after a finite set of adjustments to the initial estimates.

We have applied the shuttle algorithm to the the $3 \times 3 \times 2$ table of counts in Table 1 and have verified the computation of the bounds reported in Table 8. The alternation of Bonferroni bounds between the upper and lower limits depending upon whether k is even or odd may have something to do with whether or not the algorithm does in fact converge in a finite number of steps.

Roehrig et. al [26] describe a network algorithm approach to calculating upper and lower bounds. For 3-way tables their approach computes precisely the bounds presented here and achieved by the shuttle algorithm, but further work is required to determine how generalizations of their method to k-way tables relate to the formal bounds described in this paper.

9. Discussion

Many proposals for disclosure limitation deal with queries that arrive sequentially, and thus it is important to ask the relevance of the results on bounds described in the foregoing section for disclosure limitation in such circumstance (e.g., see Buzzigoli and Giusti [3]). In fact the results on upper and lower bounds apply directly. Suppose that an agency has responded to a sequence of queries, by releasing g different but possibly overlapping sets of marginal totals, involving k variables having determined that the risk of disclosure is acceptable. Now the agency receives a new query, for the $(g + 1)$st set of marginal totals involving a different subset of the k variables (and possibly some additional ones). To determine whether the new request is safe the agency need only compute the upper and lower bounds associated with holding the $(g + 1)$ different margins fixed.

An interesting and related problem is the extent to which one can draw inferences about "unreleased" marginal tables from that contained in overlapping released margins. To address this problem, one needs to proceed as above for each new margin of interest, and then collapse the upper and lower bounds so computed to get bounds for the entries in the unobserved margins.

The bounds for each cell entry in a contingency table described in this paper are essentially each computed separately, and thus they cannot all be achieved simultaneously. The bounds represent values associated with extremal tables that lie on the boundaries of a convex polytope and we typically get an upper bound occurring simultaneously with lower bounds for other cells, etc. This in fact is the whole idea underlying the Buzzigoli-Giusti "shuttle" algorithm.

Cell suppression algorithms typically proceed by first identifying unsafe cells using these types of marginal bounds, and then chooses an "optimal" pattern of complementary suppressions. The bounds initially computed are then not the ones of interest. What we want to do is reformulate the bounding problem for the incomplete contingency table resulting from the choice of suppressed cells. How to do this in a sensible fashion remains an open research problem.

Acknowledgements

Preparation of this paper was supported in part by the U.S. Bureau of the Census and by Statistics Netherlands. We thank Adrian Dobra, who provided comments and corrections on an earlier draft. Formal proofs for some of the results described here will appear in a subsequent paper now in preparation.

References

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

[2] Bonferroni, C.E. (1936). Teoria statistica delle classi e calcolo delle probabilitá. *Publicazioni del R. Instituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 1–62.

[3] Buzzigoli, L. and Giusti, A. (1996a). Disclosure control methods for 'linked' tables: some experiments with a matrix language. *Proceedings of Third International Seminar on Confidentiality*, Bled, Slovenia, pp.175–183.

[4] Buzzigoli, L. and Giusti, A. (1998). An algorithm to calculate the lower and upper bounds of the elements of an array given its marginals. Paper presented at Conference on Statistical Data Protection '98, Lisbon, Portugal.

[5] Chowdhury, S.D., Duncan, G.T., Krishnan, R., Roerig, S.F., and Mukherjee, S. (1996). Disclosure detection in multivariate categorical databases: An optimization approach. Unpublished manuscript, Heinz School of Public Policy and Management, Carnegie Mellon University.

[6] Cox, L. (1980). Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association*, **75**, 377–385.

[7] Cox, L. (1995). Network models for complementary cell suppression. *Journal of the American Statistical Association*, **90**,1453–1462. Addendum (1996), **91**, 1757.

[8] Dalenius, T. and Reiss, S.P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, **6**, 73–85.

[9] Dall'Aglio, G., Kotz, S., and Salinetti, G (eds.) (1991). *Advances in Probability Distributions with Given Marginals.* Kluwer, Dordrecht, Netherlands.

[10] Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, **26**, 363–397.

[11] Duncan, G.T., and Fienberg, S.E. (1997). Obtaining information while preserving privacy: A Markov perturbation method for tabular data. Paper to be presented at Conference on Statistical Data Protection '98, Lisbon, Portugal.

[12] Fienberg, S.E. (1994). Conflicts between the needs for access to statistical information and demands for confidentiality. *Journal of Official Statistics*, **10**, 115–132.

[13] Fienberg,S.E., Meyer, M.M., Makov, U.E., and Steele, R.J. (1997). Notes on generating the exact distribution for a contingency table given its marginal totals. Unpublished manuscript.

[14] Fienberg, S.E., Steele, R.J., and Makov, U.E. (1996). Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: data swapping and log-linear models. *Proceedings of Bureau of the Census 1996 Annual Research Conference*. US Bureau of the Census, Washington, DC, 87–105.

[15] Fischetti, M. and Salazar, J.J. (1996). Models and algorithms for the cell suppression problem. *Proceedings of Third International Seminar on Confidentiality*, Bled, Slovenia, pp. 114–122.

[16] Fréchet, M. (1940). *Les Probabilitiés, Associées a un Système d'Événments Compatibles et Dépendants*, Premiere Partie. Hermann & Cie, Paris.

[17] Fréchet, M. (1951). Sur les tableau de corrélation dont le marge sont données. *Ann. Univ. Lyons Sect. A, Ser. 3*, **14**, 53–77.

[18] Galambos, J., and Simonelli, I. (19996). *Bonferroni-type Inequalities with Applications*. Springer-Verlag, New York.

[19] Gusfield, D. (1988). A graph theoretic approach to statistical data security. *SIAM Journal on Computing*, **17**, 552–571.

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

[21] Hoeffding, W. (1940). Scale-invariant correlation theory. *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin,* **5(3)**, 181–233. Translated in N.I. Fisher and P.K. Sen, eds., *The Collected Works of Wassily Hoeffding*, (1994), Springer-Verlag, New York, 57–108.

[22] Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall, London.

[23] Kwerel, S.M. (1988). Fréchet bounds. In S. Kotz and N.L. Johnson (eds.), *Encyclopedia of Statistical Sciences*, Vol. 3, Wiley, New York, pp. 202–209.

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

[25] Mardia, K.V. (1970). A translation family of bivariate distributions and Fréchet bounds. *Sankhya, Series A*, **32**, 119–122.

[26] Roehrig, S.F., Padman, S., Duncan, G., and Krishnan, R. (1998). Disclosure detection in multiple linked categorical datafiles: A unified network approach. Paper presented at Conference on Statistical Data Protection '98, Lisbon, Portugal.

[27] Rüschendorf, L. (1981). Sharpness of Fréchet bounds. *Zeitschrift Wahrscheinlichkeitstheorie verw. Gebiete*, **57**, 293–302.

[28] Rüschendorf, L. (1991). Bounds for distributions with multivariate margins. In K. Mosler and M. Scarsini, eds., *Stochastic Orders and Decision under Risk*, IMS Lecture Notes–Monograph Series, Vol. 19, 285–310.

[29] Rüschendorf, L. (1996). Developments on Fréchet bounds. In L. Rüschendorf, B. Schweizer, and M.D. Taylor, eds., *Distributions with Fixed Marginals and Related Topics.* IMS Lecture Notes–Monograph Series, Vol. 28, 273-296.

[30] Rüschendorf, L., Schweizer, B., and Taylor, M.D., eds. (1996). *Distributions with Fixed Marginals and Related Topics.* IMS Lecture Notes–Monograph Series, Vol. 28, v.

[31] de Vries, R.E. (1993). Disclosure control of tablular data using subtables. Unpublished Report. Statistics Netherlands.

[32] Warmuth, W. (1988). Marginal Fréchet-bounds for multidimensional distribution functions. *Statistics*, **19**, 283–294.