# A Framework for Estimating the Bounds of Contingency Tables: Application to an Open Clinical Research Service

**A RENCI Technical Report**
TR-22-01


**Patrick L. Schmitt, Karamarie Fecho, Ashok Krishnamurthy, Priya Sharma, Hao Xu**
Renaissance Computing Institute, University of North Carolina at Chapel Hill, North Carolina

**Perry Haaland**
Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

**Bo Lan**
UNC Highway Safety Research Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

**Meghamala Sinha**
Oregon State University, Corvallis, Oregon

Corresponding author: Patrick L. Schmitt, Renaissance Computing Institute, University of North Carolina at Chapel Hill, 100 Europa Drive, Suite 540, Chapel Hill, NC, USA; Email: plschmittt@gmail.com.



How to cite this paper: Schmitt, P. L., Fecho, K., Haaland, P., Krishnamurthy, A., Lan, B., Sharma, P., & Sinha, M. A Framework for Estimating the Bounds of Contingency Tables: Application to an Open Clinical Research Service. RENCI Technical Report, TR-22-01, Renaissance ComputingInsti-tute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. doi: 10.7921/9ea1-h198. https://renci.org/technical-reports/tr-22-01.


renci

# A Framework for Estimating the Bounds of Contingency Tables: Application to an Open Clinical Research Service

Patrick L. Schmitt,[1*] Karamarie Fecho[1], Perry Haaland[2], Ashok Krishnamurthy[1], Bo Lan[3], Priya Sharma[1], Meghamala Sinha[4], Hao Xu[1]

[1]Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[2]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[3]UNC Highway Safety Research Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[4]Oregon State University, Corvallis, Oregon, USA

**Apart from the first author, all other authors are listed in alphabetical order*

## Abstract

In this paper, we provide a mathematical framework for the generation and application of contingency tables with bounds in situations where obtaining exact frequency distributions is not possible. We focus on the Integrated Clinical and Environmental Exposures Service (ICEES). ICEES is an open service that provides access to sensitive clinical data that have been integrated with public exposures data. The concept of bounded contingency tables is motivated by ICEES' privacy restrictions, which prohibit the release of electronic health record data on cohorts of fewer than 10 patients. While this service has unique limitations, the concept of bounded contingency tables is easily generalizable, and has been previously explored by others in the context of privacy restrictions imposed on open clinical data.

## Introduction

In this paper, we provide a mathematical framework for the generation and application of contingency tables with bounds in situations where obtaining exact frequency distributions is not possible. We focus on the Integrated Clinical and Environmental Exposures Service (ICEES). ICEES is an open service that provides access to sensitive clinical data that have been integrated with public exposures data (Fecho et al. 2019). The concept of bounded contingency tables is motivated by ICEES' privacy restrictions, which prohibit the release of electronic health record data on cohorts of fewer than 10 patients. While this service has unique limitations, the concept of bounded contingency tables is easily generalizable, and has been previously explored in the context of privacy restrictions by Dobra and Fienberg (2001).

## Contingency Tables

A contingency table is a specific way to represent a collection of realizations of random variables. Let $X_1, X_2$ be random variables on discrete supports $\mathcal{F}_1 = \{x_{11}, x_{12}, x_{13}\}$, $\mathcal{F}_2 = \{x_{21}, x_{22}\}$. Then an example of a contingency table is

| | $x_{11}$ | $x_{12}$ | $x_{13}$ |
|---|---|---|---|
| $x_{21}$ | 1 | 2 | 1 |
| $x_{22}$ | 0 | 0 | 4 |

where, for example, 4 is the number of observations of $X_1 = x_{13}$ and $X_2 = x_{22}$ simultaneously. Let $X_1, X_2, ..., X_n$ be random variables on supports (which we assume are discrete) $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n$. Then, the observed frequencies of events may be viewed as a function

$$\varphi : \prod_i \mathcal{F}_i \to \mathbb{N} \tag{1}$$

where $\varphi(x_1, \ldots, x_n)$ is the observed occurences of $X_1 = x_1, \ldots, X_n = x_n$. We call $\varphi$ a "frequency distribution". An n-variate contingency table is a specific way of representing $\varphi$. Namely, by listing rows of the form

| | | | | |
|---|---|---|---|---|
| $x_1$ | $x_2$ | $\cdots$ | $x_n$ | $y$ |

to represent the fact that $\varphi(x_1, x_2, \ldots, x_n) = y$. Naturally this requires $\prod_i |\mathcal{F}_i|$ rows unless rows with zeros are pruned or intervals in the supports are clustered together. For example, the first table in this format would be

| | | |
|---|---|---|
| $x_{11}$ | $x_{21}$ | 1 |
| $x_{11}$ | $x_{22}$ | 0 |
| $x_{12}$ | $x_{21}$ | 2 |
| $x_{12}$ | $x_{22}$ | 0 |
| $x_{13}$ | $x_{21}$ | 1 |
| $x_{13}$ | $x_{22}$ | 4 |

## Bounded Contingency Tables

Suppose that instead of knowing the frequencies $\varphi$ of random variables, we know only a series of lower and upper bounds on them. Let $\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_n$ again be the supports of the random variables $X_1, X_2, ..., X_n$. Then let $\mathcal{G}$ be a partition of $\prod_i \mathcal{F}_i$. Then suppose that there are functions $L, U : \mathcal{G} \to \mathbb{N}$ where

$$L(g) \leq \sum_{(x_1,...,x_n) \in g} \varphi(x_1, ..., x_n) \leq U(g) \qquad \forall g \in \mathcal{G} \tag{2}$$

thus, $U$ and $L$ provide upper and lower bounds on the sums of frequencies in collections of simultaneous events. In the context of a contingency table, $U$ and $L$ can be interpreted as providing upper and lower bounds on the sums of collections of rows of the table. These bounds can be exact, and each collection of rows can also be a single row, meaning that $g = \{(x_1, ..., x_n)\} \in \mathcal{G}$. In total we will call the objects $(\mathcal{G}, L, U)$ a "bounded contingency table".

More generally, if $S$ is any set, $\mathcal{G}$ a partition of $S$ and $U, L : \mathcal{G} \to \mathbb{N}$ are functions where $L \leq U$, we call $(\mathcal{G}, L, U)$ a bounded contingency table which "acts on $S$". we call $S$ the "rows" of the table, $\mathcal{G}$ the "row groupings", and $U, L$ the "bounds".

Given a bounded contingency table $T = (\mathcal{G}, L, U)$ and frequencies $\varphi$ we say that $\varphi$ satisfies $T$ if its frequencies are consistent with the bounds of $T$, i.e. if (2) is satisfied. Let

$$\mathcal{M}(T) = \{\varphi | \varphi \text{ satisfies } T\} \tag{3}$$

Then

$$|\mathcal{M}(T)| = \prod_{g \in \mathcal{G}} \sum_{k=L(g)}^{U(g)} \left(\!\!\binom{|g|}{k}\!\!\right) = \prod_{g \in \mathcal{G}} \sum_{k=L(g)}^{U(g)} \binom{|g| + k - 1}{k} \tag{4}$$

where $\left(\!\binom{|g|}{k}\!\right)$ is the multiset coefficient. This is because for each grouping of rows $g \in \mathcal{G}$ we may separately choose between $L(g)$ and $U(g)$ observations to 'deposit' in any of the $|g|$ rows, and the number of ways to deposit $k$ observations among $|g|$ rows is the multiset coefficient.

## Creation of Bounded Contingency Tables

The Integrated Clinical and Environmental Exposures Service (ICEES) provides open access to clinical data that have been integrated with public exposures data (Fecho et al. 2019). The data are accessible via an open application programming interface (OpenAPI), and the service supports functionalities that allow users to create cohorts and examine bivariate contingencies. However, the service is subject to regulatory constraints that prohibit the creation of cohorts of $< 10$ patients. We identify an approach for using ICEES cohort creation and bivariate functionalities to create multivariate contingency tables that support regression analysis and machine learning (Fecho et al. 2022). Due to the regulatory constraints, the multivariate approach results in data loss. Here, we consider the bounds of data loss in the context of ICEES.

Consider again an ordering of random variables $X_1, X_2, ..., X_n$ with supports $\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_n$ which we will assume are discrete. Suppose that there is an underlying multivariate frequency distribution given by the function

$$\varphi : \prod_{i=1}^{n} F_i \to \mathbb{N} \tag{5}$$

For any $k \leq n$ consider the function giving "incomplete" frequencies, that is, frequencies not broken down by the later variables

$$\varphi_k : \prod_{i=1}^{k} F_i \to \mathbb{N} \tag{6}$$

In this case $\varphi_n = \varphi$ is the total n-variate contingency table frequencies. Since these functions represent frequencies we can marginalize them

$$\varphi_1(x_1) = \sum_{y \in F_2} \varphi_2(x_1, y) \tag{7}$$

and so on. Through the ICEES API we can create a "cohort" which is a vector of the form

$$(c_1, c_2, \ldots, c_k) \tag{8}$$

for some $k \leq n$. From that cohort we can obtain a bivariate table, which gives us the value of every frequency of the form

$$\varphi_{k+2}(c_1, c_2, \ldots, c_k, x_{k+1}, x_{k+2}), \quad x_{k+1} \in F_{k+1}, x_{k+2} \in F_{k+2} \tag{9}$$

There is a threshold $T$ where if a cohort has less than $T$ observations associated with it, meaning that if

$$\varphi_k(c_1, c_2, \ldots, c_k) < T \tag{10}$$

then the cohort creation will fail and no information can be gained. In this case $T = 10$. If we want a complete contingency table, the first thing to try would be to look at every possible cohort of length $k = n - 2$ and create a bivariate table, which would give every frequency of the form

$$\varphi_n(c_1, \ldots, c_{n-2}, x_{n-1}, x_n) \tag{11}$$

And thus every frequency of length $n$, completing the table. However, many cohorts will be too small to make, and thus many bivariate tables will not be produced, leaving many rows of the contingency table blank. To remedy this we use the following algorithm to make a bounded contingency table:

1. Start with a bivariate table from an empty cohort (this can be made assuming the total number of all observations is greater than $T$). Then we know all frequencies of the form

   $$\varphi_2(x_1, x_2), \quad x_1 \in F_1, x_2 \in F_2 \tag{12}$$

   and have obtained a 2-variate contingency table.

2. Suppose the table is $k$-variate. This process will be done to move to a $k+1$-variate bounded contingency table until the table is a complete $n$-variate table. Assume that for every $k$-length cohort of the form $r = (r_1, r_2, \ldots, r_k)$ we either know $\varphi_k(r)$ or we know that $r \in g$ for some $g \in \mathcal{G}$ and we know the bounds $L(g), U(g)$ of this row grouping's combined frequencies. Apply this algorithm to every row $r = (r_1, r_2, \ldots, r_k)$:

   (a) If we know $\varphi(r)$, then this implies that we were able to construct a bivariate table from

$$(r_1, r_2, \ldots, r_{k-2}) \tag{13}$$

   Attempt now to create a bivariate table from $(r_1, r_2, \ldots, r_{k-1})$.

   (b) If we can construct a cohort in this way, a bivariate table can be made and we can obtain all frequencies of the form

$$\varphi_{k+1}(r_1, r_2, \ldots, r_{k-1}, x_k, x_{k+1}), \quad x_k \in F_k, x_{k+1} \in F_{k+1} \tag{14}$$

   Add these frequencies to the new $k+1$-variate table.

   (c) If a cohort cannot be made from $(r_1, \ldots, r_{k-1})$, then note that

$$\sum_{y \in F_{k+1}} \varphi_{k+1}(r_1, \ldots, r_k, y) \leq \varphi_k(r_1, \ldots, r_k) \tag{15}$$

   Hence, there is an upper bound on the sum of frequencies in rows in the set

$$g = \{(r_1, r_2, \ldots, r_k, y) | y \in F_{k+1}\} \tag{16}$$

   Add $g$ to $\mathcal{G}$ as a row grouping with an upper bound of $U(g) = \varphi_k(r)$ and a lower bound of $L(g) = 0$.

   (d) If the frequency $\varphi_k(r)$ is unknown but there are cumulative bounds, simply subdivide this row into rows of the form $(r_1, \ldots, r_k, y), y \in F_{k+1}$ and retain their prior bounds.

This process will end with an n-variate table composed of individual rows with known frequencies, and groupings of rows with unknown frequencies and upper bounds. The individual rows with known frequencies are still technically groupings of rows, just of size one each and with perfect bounds.

In the ICEES OpenAPI, the "supports" on contingency tables are not actual supports because they are not comprehensive as they do not account for undefined values. For example, logically speaking, a human must either have asthma or not, and ICEES tables only display patients that do or do not have asthma, but some patients do not have a value for "asthma" defined for them, and thus they are excluded when the "asthma" variable is introduced into the table. This is why equation (15) has a less than or equal to sign. Note that, given multiple bounded tables of this sort, an algorithm proposed by Buzzigoli and Giusti (1999), and modified by Dobra and Fienberg (2001), can be used to narrow the relevant bounds.

## Inference on Bounded Contingency Tables

Classical statistical inference takes place on observed frequencies $\varphi$ generally by making a test statistic $t(\varphi)$ and performing some decision rule on it (e.g. reject null hypothesis if $t(\varphi) > k_\alpha$). Given a bounded contingency table $T$, we have a set $\mathcal{M}(T)$ of possible underlying frequency distributions $\varphi \in \mathcal{M}$. We assume can as an uninformative prior that the actual frequencies obscured by the bounds of the table are uniformly distributed in $\mathcal{M}(T)$.

While it is difficult to mathematically describe this distribution of tables, one can algorithmically select a uniformly random table from $\mathcal{M}(T)$ by simply going through every grouping $g \in \mathcal{G}$, picking a uniform random number in $\{L(g), L(g)+1, \ldots, U(g)\}$, and allocating that number of observations one by one between every row in the grouping of rows uniformly. This is still problematic, though, because the number of possible underlying frequencies, $|\mathcal{M}(T)|$, can be massive even in small tables.

## Algebraic Properties of Frequencies

One can consider a large variety of operations on known frequencies and contingency tables. For example, if two frequency distributions $\varphi, \psi : S \to \mathbb{N}$ act on exactly the same domain then, under independence assumptions, it is reasonable to simply add them to get a more informative distribution $\varphi + \psi$. If one function $\varphi_n$ is composed of $n$-variate observations and another $\varphi_k$ where $k < n$ is composed of $k$-variate frequencies using the first $k$ supports that $\varphi_n$ uses, then one can define a $k$-variate aggregate table by marginalizing the excess variables

$$(\varphi_k + \varphi_n)(r) = \varphi_k(r) + \sum_{y_1,\ldots,y_{n-k}} \varphi_n(r_1, \ldots, r_k, y_1, \ldots, y_{n-k}) \tag{17}$$

More generally, given $\varphi_F$ acting on variables with supports $F = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n\}$, and another collection of supports $H = \{\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_k\}$ one can 'restrict' $\varphi_F$ to these new supports. First, for the sake of notational simplicity, order the supports in common between $F$ and $H$:

$$K = F \cap H = \{\mathcal{K}_1, \ldots, \mathcal{K}_j\} \tag{18}$$

Then, make a reduced lower-variate table $R(\varphi_F, H)$ by marginalizing the excess variables

$$R(\varphi_F, H) : \prod_i \mathcal{K}_i \to \mathbb{N} \tag{19}$$

$$R(\varphi_F, H)(k_1, \ldots, k_j) = \sum_{f_1,\ldots,f_{n-j}} \varphi_F(k_1, \ldots, k_j, f_1, \ldots, f_{n-j}) \tag{20}$$

To combine any two contingency tables that share any variables/supports in common, we can define a new $|F \cap H|$-variate table via

$$\varphi_F + \varphi_H = R(\varphi_F, H) + R(\varphi_H, F) \tag{21}$$

Notice that this extends the simple form of adding two tables $\varphi_F + \psi_F$ when they have the same supports because

$$F \subseteq H \Rightarrow R(\psi_F, H) = \psi_F \tag{22}$$

Hence,

$$R(\varphi_F, F) + R(\psi_F, F) = \varphi_F + \psi_F \tag{23}$$

Using this fact we can also rewrite the definition of adding frequencies as

$$\varphi_F + \varphi_H = R(\varphi_F, F \cap H) + R(\varphi_H, F \cap H) \tag{24}$$

Another property is that

$$R(\varphi_F + \psi_F, H) = R(\varphi_F, H) + R(\psi_F, H) \tag{25}$$

Moreover,

$$R(R(\varphi_F, H), L) = R(\varphi_F, H \cap L) \tag{26}$$

Using all of these above properties we see that for three collections of supports $F, H, L$ and frequency functions on those supports

$$
\begin{aligned}
(\varphi_F + \varphi_H) + \varphi_L &= R(R(\varphi_F, H) + R(\varphi_H, F), L) + R(\varphi_L, F \cap H) && (27)\\
&= R(R(\varphi_F, H), L) + R(R(\varphi_H, F), L) + R(\varphi_L, F \cap H) && (28)\\
&= R(\varphi_F, H \cap L) + R(\varphi_H, F \cap L) + R(\varphi_L, F \cap H) && (29)\\
&= R(\varphi_F, H \cap L) + R(R(\varphi_H, L), F) + R(R(\varphi_L, H), F) && (30)\\
&= R(\varphi_F, H \cap L) + R(R(\varphi_H, L) + R(\varphi_L, H), F) && (31)\\
&= \varphi_F + (\varphi_H + \varphi_L) && (32)
\end{aligned}
$$

So addition of distinct tables is associative.

Even more generally for collections of supports $H_1, H_2, \ldots, H_p$ and frequency functions $\varphi_{H_1}, \ldots, \varphi_{H_p}$ one may define their sum as a $|\bigcap_{j=1}^{p} H_i|$-variate table by

$$\sum_{i=1}^{p} \varphi_{H_i} = \sum_{i=1}^{p} R(\varphi_{H_i}, \bigcap_{j=1}^{p} H_j) \tag{33}$$

## Algebraic Properties of Bounded Contingency Tables

We can define the entropy of a bounded table $T = (\mathcal{G}, U, L)$ by the entropy of a uniformly distributed variable over $\mathcal{M}(T)$:

$$
\begin{aligned}
H(T) &= -\sum_{\varphi \in \mathcal{M}(T)} \frac{1}{|\mathcal{M}(T)|} \log \frac{1}{|\mathcal{M}(T)|} = \log |\mathcal{M}(T)| && (34)\\
&= \sum_{g \in \mathcal{G}} \log \sum_{k=L(g)}^{U(g)} \binom{|g| + k - 1}{k} && (35)
\end{aligned}
$$

Suppose that $\varphi : S \to \mathbb{N}$ is a frequency distribution. In most cases we would have $S = \prod_i \mathcal{F}_i$ for some supports $\mathcal{F}_i$. Let

$$\mathcal{N}(\varphi) = \{T | T \text{ is satisfied by } \varphi\} \tag{36}$$

Be the set of bounded contingency tables acting on the same set as $\varphi$ whose bounds are satisfied by the observed frequencies $\varphi$. Let $\mathcal{G}$ be a partition of $S$, so it is a possible row grouping for bounded tables in $\mathcal{N}(\varphi)$. Then let

$$\mathcal{N}(\varphi, \mathcal{G}) = \{T | T \text{ is satisfied by } \varphi \text{ and has groupings } \mathcal{G}\} \tag{37}$$

Such that

$$\mathcal{N}(\varphi) = \bigcup_{\mathcal{G} \text{ partitions } S} \mathcal{N}(\varphi, \mathcal{G}) \tag{38}$$

Clearly some bounded tables will be more useful than others. We seek to order these tables by how informative they are. The most obvious way to do this is order them by $H(T)$, however this is only useful for deciding which table you would prefer if you could only have one or the other. In other words, ordering bounded tables in this manner would be saying that for $T_1, T_2 \in \mathcal{N}(\varphi)$

$$T_1 \geq T_2 \Leftrightarrow |\mathcal{M}(T_1)| \leq |\mathcal{M}(T_2)| \tag{39}$$

But a stronger and more useful notion of ordering would be that

$$T_1 \geq T_2 \Leftrightarrow \mathcal{M}(T_1) \subseteq \mathcal{M}(T_2) \tag{40}$$

or, equivalently,

$$T_1 \geq T_2 \Leftrightarrow \mathcal{M}(T_1) \cap \mathcal{M}(T_2) = \mathcal{M}(T_1) \tag{41}$$

If this were the case, then having $T_1$ and $T_2$ at the same time would tell us absolutely nothing useful compared to simply having $T_1$. We are going to try and describe this 'informativeness' order on $\mathcal{N}(\varphi)$.

First, consider only the case of applying this ordering to $\mathcal{N}(\varphi, \mathcal{G})$ for some fixed row grouping $\mathcal{G}$. Let $T_1 = (\mathcal{G}, U_1, L_1)$ and $T_2 = (\mathcal{G}, U_2, L_2) \in \mathcal{N}(\varphi, \mathcal{G})$ be two bounded contingency tables. We will create a partial order on $\mathcal{N}(\varphi, \mathcal{G})$ by saying that

$$T_1 \leq T_2 \Leftrightarrow U_1 \geq U_2 \text{ and } L_1 \leq L_2 \tag{42}$$

So that a table is lesser than another table if its bounds are weaker.
This partial order forms a lattice because each two elements have a unique greatest lower bound $T_1 \wedge T_2$ and lowest greater bound $T_1 \vee T_2$

$$T_1 \wedge T_2 = (K, G, \max(U_1, U_2), \min(L_1, L_2)) \tag{43}$$

$$T_1 \vee T_2 = (K, G, \min(U_1, U_2), \max(L_1, L_2)) \tag{44}$$

Let $\mathbf{1}_{\mathcal{G}} = (\mathcal{G}, I, I)$ where $I(g) = \sum_{r \in g} \varphi(r)$ is the actual sum of frequencies for each grouping of outcomes. Then

$$T_1 \wedge \mathbf{1}_{\mathcal{G}} = \mathbf{1}_{\mathcal{G}}, \quad T_1 \vee \mathbf{1}_{\mathcal{G}} = T_1 \tag{45}$$

So on $\mathcal{N}(\varphi, \mathcal{G})$, these operations make an infinite lattice with a greatest upper bound of $\mathbf{1}_{\mathcal{G}}$ and where higher elements have lower entropy.

Moreover this is a distributive lattice as

$$T_1 \wedge (T_2 \vee T_3) = (T_1 \wedge T_2) \vee (T_1 \wedge T_3) \tag{46}$$

This poset is locally finite in the sense that intervals (sets of elements greater and lesser than two elements) are finite. We have

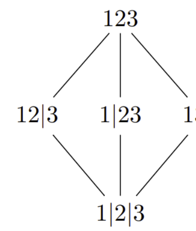
$$\#[T_1, T_2] = \prod_{g \in \mathcal{G}} (L_2(g) - L_1(g))(U_1(g) - U_2(g)) \tag{47}$$

Since in each row grouping $g \in \mathcal{G}$ we are choosing two numbers $u, l$ so that $L(g) \leq l \leq I(g) \leq u \leq U(g)$.

We return to the general case of $\mathcal{N}(\varphi)$. We seek to order tables so that if one is less than the other, then no new information about the underlying frequencies can be gained whatsoever. The ordering is simple when the tables have the same row groupings within $\mathcal{N}(\varphi, \mathcal{G})$, but more complex in general.
In combinatorics, the notation $\Pi(S)$ is used for the set of partitions of a set $S$. So

$$\mathcal{N}(\varphi) = \bigcup_{\mathcal{G} \in \Pi(S)} \mathcal{N}(\varphi, \mathcal{G}) \tag{48}$$

The set of partitions $\Pi(S)$ is has a natural lattice structure with the relation that for $\pi, \sigma \in \Pi(S)$ we have $\pi \leq \sigma$ if and only if every set in $\pi$ is a subset of a set in $\sigma$.




$\mathbf{\Pi_3}$

*The partition lattice on 3 elements*

However, in our case, when ordering by informativeness the tables with finer row groupings will usually be more informative because they will, in general, have fewer possible underlying frequency distributions. This makes our ordering somewhat opposite to the natural ordering of partitions.

First, note that the set of maximal elements of $\mathcal{N}(\varphi, \mathcal{G})$, that is, tables of the form $\mathbf{1}_\mathcal{G} = (\mathcal{G}, I_\mathcal{G}, I_\mathcal{G})$ for some row grouping $\mathcal{G}$, is a subposet of $\mathcal{N}(\varphi)$ which is isomorphic to the dual partition lattice $\Pi^*(S)$ (the partition lattice with reversed ordering) via the map

$$\{(\mathcal{G}, I_\mathcal{G}, I_\mathcal{G}) | \mathcal{G} \in \Pi(S)\} \to \Pi^*(S) \tag{49}$$

$$\mathbf{1}_\mathcal{G} = (\mathcal{G}, I_\mathcal{G}, I_\mathcal{G}) \mapsto \mathcal{G} \tag{50}$$

That is to say

$$\mathcal{M}(\mathbf{1}_{\mathcal{G}_1}) \subseteq \mathcal{M}(\mathbf{1}_{\mathcal{G}_2}) \Leftrightarrow \mathbf{1}_{\mathcal{G}_1} \geq \mathbf{1}_{\mathcal{G}_2} \Leftrightarrow \mathcal{G}_1 \geq \mathcal{G}_2 \tag{51}$$

In other words, if we look only at bounded tables with perfect bounds, then they are ordered in terms of usefulness by the fineness of the partitions of their row groupings. The absolute maximally informative bounded table on all of $\mathcal{N}(\varphi)$ is the 'trivial' table with both perfect bounds and where each row grouping is just a single row. This effectively gives us the exact frequencies $\varphi$, hence it is the most informative table possible.

We can visually represent a bounded contingency table as follows: first, enumerate the rows all the tables will act on $S = \{x_1, x_2, \ldots, x_n\}$. Then the figure below

$$\begin{array}{cc} 4 & 1 \\ \boxed{12} & \boxed{3} \\ 2 & 0 \end{array}$$

represents the table $T = (\mathcal{G}, U, L)$ such that $\mathcal{G} = \{\{x_1, x_2\}, \{x_3\}\}$ and $U(\{x_1, x_2\}) = 4$, $U(\{x_3\}) = 1$ and $L(\{x_1, x_2\}) = 2, L(\{x_3\}) = 0$.

In this context we can represent a frequency distribution $\varphi$ by a vector $\varphi = (a, b, c)$ to say that $\varphi(x_1) = a$,$\varphi(x_2) = b$ and so on. For example, a frequency satisfying the above bounded table's constraints would be $\varphi = (2, 1, 0)$ because $2 \leq 2 + 1 \leq 4$ and also $0 \leq 0 \leq 1$.

An example of ordering of these tables is

$$\begin{array}{ccc} 4 & & 5 \\ \boxed{12} & \geq & \boxed{12} \\ 2 & & 1 \end{array}$$

Using this notation we might define the 'concatenation' of two tables on disjoint supports via

$$\begin{array}{ccccc} 4 & & 1 & & \begin{array}{cc} 4 & 1 \end{array} \\ \boxed{12} & \bigoplus & \boxed{3} & = & \begin{array}{cc} \boxed{12} & \boxed{3} \end{array} \\ 2 & & 0 & & \begin{array}{cc} 2 & 0 \end{array} \end{array}$$

Formally, if we have two bounded tables $T = (\mathcal{G}_T, U_T, L_T), V = (\mathcal{G}_V, U_V, L_V)$ acting on disjoint supports (meaning that $\mathcal{G}_T$ partitions $S_T$ and $\mathcal{G}_V$ partitions $S_V$ where $S_T \cap S_V = \emptyset$), then we define

$$T \oplus V = (\mathcal{G}_{T\oplus V}, U_{T\oplus V}, L_{T\oplus V}) \tag{52}$$
$$\mathcal{G}_{T\oplus V} = \mathcal{G}_T \cup \mathcal{G}_V \in \Pi(S_T \cup S_V) \tag{53}$$
$$U_{T\oplus V}(g) = U_T(g) \text{ if } g \in S_T \text{ else } U_V(g) \tag{54}$$
$$L_{T\oplus V}(g) = L_T(g) \text{ if } g \in S_T \text{ else } L_V(g) \tag{55}$$

This operation is associative and commutative.

**Lemma 1.** Let $T$ be a bounded table decomposable as

$$T = \bigoplus_{i\in I} T_i \tag{56}$$

Then

$$\varphi \text{ satisfies } T \Leftrightarrow \varphi \text{ satisfies } T_i \text{ for all } i \in I \tag{57}$$

Put simply, a frequency function satisfies bounds if it satisfies the bounds of every block its divisible into. This is completely tautological, but worth nothing for proofs. This is useful because it can be extended to compare multiple tables in the following way:

**Theorem 2.** Suppose that $T, V$ are bounded contingency tables operating on $S = \bigcup_{i\in I} S_i$. Suppose that they can be decomposed as

$$T = \bigoplus_{i\in I} T_i, \quad V = \bigoplus_{i\in I} V_i \tag{58}$$

Where for all $i$, $T_i$ and $V_i$ operate on the same set $S_i$. Then

$$T \leq V \Leftrightarrow T_i \leq V_i \quad \forall i \tag{59}$$

*Proof.* For the rightward direction by contraposition suppose that there were $k$ so that for all $i \neq k$ we have $T_i \leq V_i$ but $T_k$ was not less informative $V_k$. So $\mathcal{M}(V_k) \nsubseteq \mathcal{M}(T_k)$. Then there would be a function $\varphi \in \mathcal{M}(V_k) \setminus \mathcal{M}(T_k)$ which broke the bounds of $T_k$ but satisfied the bounds of $V_k$. Extending this function in any way that satisfies the bounds of all $V_i$ where $i \neq k$ would yield

$$\varphi \in \mathcal{M}(V) \setminus \mathcal{M}(T) \Rightarrow \mathcal{M}(V) \nsubseteq \mathcal{M}(T) \tag{60}$$

And so it is not the case that $T \leq V$. For the leftward direction if for all $i$ there is $T_i \leq V_i$ then any function $\varphi$ which satisfies all of $V_i$'s bounds must also satisfy $T_i$'s bounds. So $V$ is more informative than $T$. Hence $T \leq V$. ☐

Ordering is reflexive, so for any table $V$ we must have $V \leq V$. From this we may also conclude that if $T$ has a disjoint support to that of $\varphi$'s domain then

$$T_1 \leq T_2 \Leftrightarrow T_1 \oplus V \leq T_2 \oplus V \tag{61}$$

Hence, there is an order isomorphism

$$\mathcal{N}(\varphi) \simeq \mathcal{N}(\varphi) \oplus V \tag{62}$$

By a combinatorial argument we also see that

$$\#\mathcal{M}(\bigoplus_{i\in I} T_i) = \prod_i \#\mathcal{M}(T_i) \tag{63}$$

We will now solve a special case of ordering. Consider the following tables ordered by informativeness

$$\begin{array}{cc} 1 & 1 \\ \boxed{1} & \boxed{2} \\ 0 & 0 \end{array} \geq \begin{array}{c} 2 \\ \boxed{12} \\ 0 \end{array}$$

Which we can see because $\varphi = (2,0)$ and $\varphi = (0,2)$ are satisfied by the rightmost table but not the leftmost, so the leftmost table is more informative. However on the other hand

$$\begin{array}{c} 1 \\ \boxed{12} \\ 0 \end{array} \geq \begin{array}{cc} 1 & 1 \\ \boxed{1} & \boxed{2} \\ 0 & 0 \end{array}$$

This is because $\varphi = (1,1)$ satisfies the rightmost table but not the leftmost. If we generalize this, we see that the upper bound on the left has to be less than or equal to the minimum of upper bounds on the right.

**Theorem 3.** Let $T_1, T_2 \in \mathcal{N}(\varphi)$ be bounded contingency tables where $\mathcal{G}_1 = \{g_1\}$ and $\mathcal{G}_2$ is a partition of $g_1$. Then a necessary and sufficient condition for $T_1$ to be more informative than $T_2$ is

$$T_1 \geq T_2 \Leftrightarrow \mathcal{M}(T_1) \subseteq \mathcal{M}(T_2) \tag{64}$$

$$U_1(g_1) \leq \min\{U_2(g_2) | g_2 \in \mathcal{G}_2\} \tag{65}$$

$$\text{and } L_1(g_1) \geq \max\{L_2(g_2) | g_2 \in \mathcal{G}_2\} \tag{66}$$

*Proof.* For the rightward direction by contraposition suppose that (65) and (66) are not satisfied. We aim to show that (64) is not satisfied. So we have supposed that

$$U_1(g_1) > \min\{U_2(g_2) | g_2 \in \mathcal{G}_2\} \text{ or} \tag{67}$$

$$L_1(g_1) < \max\{L_2(g_2) | g_2 \in \mathcal{G}_2\} \tag{68}$$

Assume WLOG (67). Pick $x \in g^*$ where $g^* \in \mathcal{G}_2$ is the minimizer of $U_2(g^*)$. Let $\gamma : S \to \mathbb{N}$ be a frequency function. Let

$$\gamma(x) = U_1(g_1), \quad \gamma = 0 \text{ otherwise} \tag{69}$$

Then

$$\sum_{z \in g^*} \gamma(z) = \gamma(x) = U_1(g_1) > U_2(g^*) \tag{70}$$

So $\gamma$ satisfies the bounds of $T_1$ but not $T_2$ because the bounds for $g^*$ aren't satisfied. In other words

$$\gamma \in \mathcal{M}(T_1) \setminus \mathcal{M}(T_2) \tag{71}$$

And so it is not the case that $\mathcal{M}(T_1) \subseteq \mathcal{M}(T_2)$. This concludes the rightward direction.
For the leftward direction suppose that (65) and (66) are satisfied. We aim to show that (64) is satisfied. Let $\psi$ satisfy $T_1$. We aim to show $\psi$ satisfies $T_2$ also. We have

$$\sum_{x \in g_1} \psi(x) \leq U_1(g_1) \leq \min\{U_2(g_2) | g_2 \in \mathcal{G}_2, g_2 \subseteq g_1\} \tag{72}$$

$$\Rightarrow \forall g_2 \in \mathcal{G}_2 \quad \sum_{x \in g_2} \psi(x) \leq U_1(g_1) \leq U_2(g_2) \tag{73}$$

hence $\psi$ satisfies the upper bounds all throughout $\mathcal{G}_2$. The case is equivalent for lower bounds. And so $\psi \in \mathcal{M}(T_2)$. $\square$

We will now show how tables can be decomposed so that the above theorem can be applied more generally.

**Definition 1.** We call a restriction of a partition $\mathcal{A} \in \Pi(S)$ to a subset $c \subseteq S$

$$R_c(\mathcal{A}) = \{a | a \in \mathcal{A}, a \cap c \neq \emptyset\} \tag{74}$$

And we call the overlap of a partition with a subset

$$O_c(\mathcal{A}) = \bigcup_{a \in \mathcal{A}, a \cap c \neq \emptyset} a = \bigcup R_c(\mathcal{A}) \tag{75}$$

We see that $R_c(\mathcal{A}) \in \Pi(O_c(\mathcal{A}))$.

**Theorem 4.** Given partitions $\mathcal{A}, \mathcal{C} \in \Pi(S)$, $\mathcal{A}$ refines $\mathcal{C}$ if and only if

$$O_c(\mathcal{A}) = c \quad \forall c \in \mathcal{C} \tag{76}$$

*Proof.* For the rightward direction suppose $\mathcal{A}$ refines $\mathcal{C}$. Then every set $a$ such that $a \cap c \neq \emptyset$ is a subset of $c$, and $\mathcal{A}$ covers all of $S$, hence

$$\bigcup_{a \in \mathcal{A}, a \cap c \neq \emptyset} a = c \tag{77}$$

For the leftward direction, by contraposition if $\mathcal{A}$ did not refine $\mathcal{C}$, then there would exist a set $a \in \mathcal{A}$ such that there was no superset of $a$ within $\mathcal{C}$. Pick any $c$ such that $a \cap c \neq \emptyset$, so $a$ has elements not in $c$. Then $O_c(\mathcal{C}) \setminus c \neq \emptyset$ hence $O_c(\mathcal{A}) \neq c$. □

**Lemma 5.** If $\mathcal{A}$ refines $\mathcal{C}$, then $\mathcal{A}$ is partitioned by the sets of its restrictions $R_c(\mathcal{A})$ i.e.

$$\{R_c(\mathcal{A}) | c \in \mathcal{C}\} \in \Pi(\mathcal{A}) \tag{78}$$

*Proof.* By observation $\mathcal{A}$ is the union of all $R_c(\mathcal{A})$, and no set could be in $R_c(\mathcal{A}) \cap R_d(\mathcal{A})$ since then there would be $a \cap c \neq \emptyset$ and $a \cap d \neq \emptyset$ hence $a \subseteq c, a \subseteq d$ but then $c \cap d = a \neq \emptyset$ and $\mathcal{C}$ is a partition. □

**Definition 2.** Given two partitions $\mathcal{A}, \mathcal{B} \in \Pi(S)$ and a third partition $\mathcal{C} \in \Pi(S)$, we call $\mathcal{C}$ a simultaneous decomposition of $\mathcal{A}, \mathcal{B}$ if $\mathcal{A}$ and $\mathcal{B}$ simultaneously refine $\mathcal{C}$, in formal terms

$$O_c(\mathcal{A}) = O_c(\mathcal{B}) = c \quad \forall c \in \mathcal{C} \tag{79}$$

Here is an example of a simultaneous decomposition of partitions:

$$\mathcal{A} = 1\ 2|3 \mid 4\ 5 \tag{80}$$
$$\mathcal{B} = 1|2\ 3 \mid 4|5 \tag{81}$$
$$\mathcal{C} = 1\ 2\ 3 \mid 4\ 5 \tag{82}$$

We can use a simultaneous decomposition to compare bounded tables. Let $T, V$ be bounded tables on the same rows, and let $\mathcal{C}$ be any simultaneous decomposition of $\mathcal{G}_T, \mathcal{G}_V$. Then for any $c \in \mathcal{C}$ we have that $R_c(\mathcal{A}), R_c(\mathcal{B}) \in \Pi(c)$. Now decompose $T$ and $V$ in the following way:

$$T = \bigoplus_{c \in \mathcal{C}} T_c = \bigoplus_{c \in \mathcal{C}} (R_c(\mathcal{G}_T), U_T, L_T) \tag{83}$$
$$V = \bigoplus_{c \in \mathcal{C}} V_c = \bigoplus_{c \in \mathcal{C}} (R_c(\mathcal{G}_V), U_V, L_V) \tag{84}$$

Then applying Theorem 2 we see that $T \leq V$ if and only if for all $c \in \mathcal{C}$ we have $T_c \leq V_c$.

**Definition 3.** Let $\mathcal{A}, \mathcal{B} \in \Pi(S)$. Then we say that $\mathcal{A}$ and $\mathcal{B}$ "don't partially overlap" if

$$\forall a \in \mathcal{A} \forall b \in \mathcal{B} \quad a \cap b \neq \emptyset \Rightarrow a \subseteq b \vee b \subseteq a \tag{85}$$

For example, the partitions 12|3 and 1|23 have partial overlapping. 12 and 12 don't have partial overlapping, and 12 and 1|2 also don't. The partitions 1|2|34 and 12|3|4 don't partially overlap either.

**Theorem 6.** Let $\mathcal{A}, \mathcal{B} \in \Pi(S)$ have no partial overlappings. Then let

$$\mathcal{C} = \{a \cup b | a \in \mathcal{A}, b \in \mathcal{B}, a \cap b \neq \emptyset\} \tag{86}$$

Then $\mathcal{C}$ is a simultaneous decomposition of $\mathcal{A}$ and $\mathcal{B}$, and for every $c \in \mathcal{C}$ we have that one of $R_c(\mathcal{A}), R_c(\mathcal{B})$ is just $\{c\}$ and another is a partition of $c$.

*Proof.* Because $\mathcal{A}, \mathcal{B}$ have no partial overlappings, for any $c = a \cup b \in \mathcal{C}$ we must have that either $a, b$ are disjoint, or one is a subset of another. Thus either $a \cup b = a$ or $a \cup b = b$. Suppose without loss of generality that $c = a \cup b = a$. Then

$$R_c(\mathcal{A}) = \{a \in \mathcal{A} | a \cap c \neq \emptyset\} = \{a\} = \{c\} \tag{87}$$

And $R_c(\mathcal{B})$ is composed of subsets of $c$ which cover $c$, so $R_c(\mathcal{B})$ is a partition of $c$. $\square$

Now, given any bounded tables whose row groupings don't have partial overlappings, we may decompose them according to the simultaneous decomposition in Theorem 6 and then use Theorems 2 and 3 to order the tables.

# Discussion

Herein, we provide a framework for the generation and application of bounded contingency tables in the context of ICEES as a driving use case. The framework we provide is not unique to ICEES, however, but rather is generalizable to any situation where bounds exist on contingency tables due to imprecise frequency distributions. The abstract formulation of bounded contingency tables as a partition with two functions also provides a novel combinatorial object that can be further studied.


## Acknowledgements

PLS would like to thank Stanley C. Ahalt of the Renaissance Computing Institute for collaboration and mentorship.


## Data Availability

ICEES is openly accessible at https://icees-asthma.renci.org/apidocs (Permalink URL: https://perma.cc/7RWE-78JL).

## Ethical Assurances

All study procedures were approved by the Institutional Review Board at the University of North Carolina at Chapel Hill (protocol #16-2978).


## Funding

This project was funded with awards from the National Center for Advancing Translational Sciences, National Institutes of Health (OT3TR002020, OT2TR003430, UL1TR002489, UL1TR002489-03S4). PLS also received support from the Renaissance Computing Institute through a student internship.


## Competing Interests

The authors have declared no competing interests.

## Bibliography


Fecho K,* Haaland P, Krishnamurthy A, Lan B, Ramsey S, Schmitt PL, Sharma P, Sinha M, Xu H. *An approach for open multivariate analysis of integrated clinical and environmental exposures data.* Inform Med Unlocked 2021;26:100733. doi.org/10.1016/j.imu.2021.100733. **Apart from first/lead author, all other authors are listed in alphabetical order*

Ahalt SC, Bizon C, Champion J, Cox S, Fecho K, Krishnamurthy A, Peden D, Pfaff E, Stillwell L, Tropsha A, Xu H. *A novel approach for exposing and sharing clinical data: the Translator Integrated Clinical and Environmental Exposures Service.* J Am Med Inform Assoc 2019;26(10):1064–1073. doi: 10.1093/jamia/ocz042.

Dobra A, Fienberg SE. *Bounding Entries in Multi-way Contingency Tables Given a Set of Marginal Totals.* Department of Statistics and Center for Automated Learning and Discovery, Carnegie Mellon University. October 8, 2001.

Buzzigoli L, Giusti A. *An Algorithm to Calculate the Lower and Upper Bounds of the Elements of an Array Given its Marginals.* Statistical Data Protection (SDP 1998) Proceedings, 131–147, 1999. Eurostat, Luxembourg.