Bounded Contingency Tables

Patrick Schmitt

March 25, 2022

Abstract

In this paper we describe the generation and use of contingency tables with bounds in situations where obtaining exact frequency distributions (in this case due to privacy restrictions) are impossible.

Introduction

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) $\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_n$. Then, the observed frequencies of events may be viewed as a function

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

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

x_1	x_2	$\cdots$	x_n	y

to represent the fact that $\varphi(x_1, x_2, \dots, x_n) = y$. 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 φ 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, L and U can be interpreted as providing lower and upper 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".

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

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. The normal contingency table is then representing

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

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' of some length k which is a vector of the form

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

and from that cohort we can obtain a bivariate table, which gives us the value of every frequency of the form

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

The problem is that there is a threshold which we will call 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 they have 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 old 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 API 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. This is also why the initial ordering can change the resulting table's bounds.

Inference on Bounded Contingency Tables

Classical statistical inference takes place on observed frequencies φ generally by making a test statistic $t(\varphi)$ and performing some decision rule on it (e.g. reject null hypothesis if $t(\varphi) > 5$). Given a bounded contingency table, we have a set $\mathcal{M}$ of possible underlying frequencies/tables $\varphi \in \mathcal{M}$. We assume as an uninformative prior that the actual frequencies obscured by the bounds of the table are uniformly distributed in $\mathcal{M}$.

While it is difficult to mathematically describe this distribution of tables, one can easily algorithmically select a uniformly random table from $\mathcal{M}$ 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 points between every row in the grouping of rows uniformly. This is still problematic, though, because the number of possible underlying frequencies, $|\mathcal{M}|$, can be massive even in small tables.

Algebraic Properties of Frequencies

One can consider a large variety of operations on known frequencies/standard contingency tables as well as bounded contingency tables. For example, if two frequencies φ, ϕ act on exactly the same supports then, under independence assumptions, it is reasonable to simply add them to get a more informative table $\varphi + \phi$. If one function φ_n is composed of n-variate observations and another φ_k where $k < n$ is composed of k-variate frequencies using the first k supports that φ_n uses, one can define a k-variate aggregate table by marginalizing the excess variables

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

While the ordering of variables in a contingency table has been emphasized because it is important to generating bounded contingency tables, it is of course not important when all frequencies are known. One can permute the ordering of variables without any information being lost.

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

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

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

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

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

To combine any two contingency tables which share any variables/supports in common at all 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

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

And 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

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

And 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.

And so 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

As for bounded contingency tables, in line with the previous description of treating them as representing a uniform random variable over $\mathcal{M}$, the set of possible underlying frequencies dictated by the bounds, we can define the entropy of a table $T = (\mathcal{G}, U, L)$ by the entropy of a uniformly distributed variable over $\mathcal{M}$:

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

Suppose that φ is a function of observed frequencies. Let

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

Be the set of bounded contingency tables on the same supports as φ whose bounds are satisfied by the observed frequencies φ. We can also specify the set of bounded tables with some particular grouping of rows (as long as the grouping $\mathcal{G}$ partitions the domain of φ, that is $\prod_i \mathcal{F}_i$)

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

So that

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

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 $\mathcal{N}(\varphi, \mathcal{G})$ 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 $T_1 \vee T_2$

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

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

It follows that

$$H(T_1 \wedge T_2) \geq H(T_1), H(T_2) \tag{45}$$

$$H(T_1 \vee T_2) \leq H(T_1), H(T_2) \tag{46}$$

Between bounded tables defined on $\mathcal{G}$, these operations are both commutative and associative. Additionally, as $\max(L_1, \min(L_1, L_2)) = L_1$ in general

$$T_1 \wedge (T_1 \vee T_2) = T_1 = T_1 \vee (T_1 \wedge T_2) \tag{47}$$

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

$$T_1 \wedge \mathbf{1} = \mathbf{1}, \quad T_1 \vee \mathbf{1} = T_1 \tag{48}$$

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

Moreover this is a distributive lattice as

$$T_1 \wedge (T_2 \vee T_3) = (T_1 \wedge T_2) \vee (T_1 \wedge T_3) \tag{49}$$

This poset is locally finite in the sense that intervals (sets of elements greater and lesser than two elements) are finite. Given a table $T = (\mathcal{G}, U, L) \in \mathcal{N}(\varphi, \mathcal{G})$ the number of tables in $\mathcal{N}(\varphi, \mathcal{G})$ greater than T according to this ordering will be

$$\#[T, \mathbf{1}] = \prod_{g \in \mathcal{G}} (I(g) - L(g))(U(g) - I(g)) \tag{50}$$

$$\text{where } I(g) = \sum_{r \in g} \varphi(r) \tag{51}$$

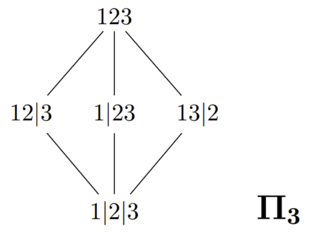
Since in each row grouping $g \in \mathcal{G}$ we are choosing two numbers u, l so that $L(g) \leq l \leq I(g) \leq u \leq U(g)$. In general if $T_1 \leq T_2$ then

$$\#[T_1, T_2] = \prod_{g \in \mathcal{G}} (L_2(g) - L_1(g))(U_1(g) - U_2(g)) \tag{52}$$

Now we will go back 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. Hence if we rename the rows these bounded tables act on to $S = \prod_i \mathcal{F}_i$ then

$$\mathcal{N}(\varphi) = \bigcup_{\mathcal{G} \in \Pi(S)} \mathcal{N}(\varphi, \mathcal{G}) \tag{53}$$

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 π is a subset of a set in σ.



The partition lattice on 3 elements

However, in our case, when ordering by informativeness the tables with finer row groupings will 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 $\Pi^*(S)$ via the canonical isomorphism

$$\{(\mathcal{G}, I_\mathcal{G}, I_\mathcal{G}) | \mathcal{G} \in \Pi(S)\} \to \Pi^*(S) \tag{54}$$

$$\mathbf{1}_\mathcal{G} = (\mathcal{G}, I_\mathcal{G}, I_\mathcal{G}) \mapsto \mathcal{G} \tag{55}$$

That is to say

$$\mathcal{M}(\mathbf{1}_{\mathcal{G}_1}) \subseteq \mathcal{M}(\mathbf{1}_{\mathcal{G}_2}) \Leftrightarrow \mathbf{1}_{\mathcal{G}_1} \geq \mathbf{1}_{\mathcal{G}_2} \Leftrightarrow \mathcal{G}_1 \geq \mathcal{G}_2 \tag{56}$$

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 φ, hence it is the best possible bounded table to have.

To go further, we must be able to see the structure of tables more easily. 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

$$\underset{2}{\overset{4}{\boxed{12}}}\underset{0}{\overset{1}{\boxed{3}}}$$

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 function φ by a vector $\varphi = (a, b, c)$ to say that $\varphi(x_1) = a$,$\varphi(x_2) = b$ and so on. So for example, a frequency satisfying the above bound 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

$$\underset{2}{\overset{4}{\boxed{12}}} \geq \underset{1}{\overset{5}{\boxed{12}}}$$

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

$$\underset{2}{\overset{4}{\boxed{12}}} \bigoplus \underset{0}{\overset{1}{\boxed{3}}} = \underset{2}{\overset{4}{\boxed{12}}}\underset{0}{\overset{1}{\boxed{3}}}$$

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

$$T \oplus V = (\mathcal{G}_{T\oplus V}, U_{T\oplus V}, L_{T\oplus V}) \tag{57}$$

$$\mathcal{G}_{T\oplus V} = \mathcal{G}_T \cup \mathcal{G}_V \in \Pi(S_T \cup S_V) \tag{58}$$

$$U_{T\oplus V}(g) = U_T(g) \text{ if } g \in S_T \text{ else } U_V(g) \tag{59}$$

$$L_{T\oplus V}(g) = L_T(g) \text{ if } g \in S_T \text{ else } L_V(g) \tag{60}$$

This operation is associative and commutative. This gives us a

Lemma 0.1. Let T be a bounded table decomposable as

$$T = \bigoplus_{i\in I} T_i \tag{61}$$

Then

$$\varphi \text{ satisfies } T \Leftrightarrow \forall i \in I \quad \varphi \text{ is satisfactory on } T_i \tag{62}$$

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. From this fact we may also conclude that if T has disjoint support to that of φ's domain then

$$T_1 \leq T_2 \Leftrightarrow T_1 \oplus V \leq T_2 \oplus V \tag{63}$$

so

$$\mathcal{N}(\varphi) \cong \mathcal{N}(\varphi) \oplus V \tag{64}$$

By a combinatorial argument we also see that

$$\#\mathcal{M}(\bigoplus_{i\in I} T_i) = \prod_i \#\mathcal{M}(T_i) \tag{65}$$

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

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

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

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

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 0.2. 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

$$T_1 \geq T_2 \Leftrightarrow \mathcal{M}(T_1) \subseteq \mathcal{M}(T_2) \tag{66}$$

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

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

Proof. For the rightward direction by contraposition suppose that (67) and (68) are not satisfied. We aim to show that (66) is not satisfied. So we have supposed that

$$U_1(g_1) > \min\{U_2(g_2)|g_2 \in \mathcal{G}_2\} \text{ or} \tag{69}$$

$$L_1(g_1) < \max\{L_2(g_2)|g_2 \in \mathcal{G}_2\} \tag{70}$$

Assume WLOG (69). 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

$$\gamma(x) = U_1(g_1), \quad \gamma = 0 \text{ otherwise} \tag{71}$$

Then

$$\sum_{z \in g^*} \gamma(z) = \gamma(x) = U_1(g_1) > U_2(g^*) \tag{72}$$

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

$$\gamma \in \mathcal{M}(T_1) \setminus \mathcal{M}(T_2) \tag{73}$$

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 (67) and (68) are satisfied. We aim to show that (66) is satisfied. Let ψ satisfy T_1. We aim to show ψ satisfies T_2 also. We have

$$\sum_{x \in g_1} \psi(x) \leq U_1(g_1) \leq \min\{U_2(g_2)|g_2 \in \mathcal{G}_2, g_2 \subseteq g_1\} \tag{74}$$

$$\Rightarrow \forall g_2 \in \mathcal{G}_2 \quad \sum_{x \in g_2} \psi(x) \leq U_1(g_1) \leq U_2(g_2) \tag{75}$$

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

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

Definition 0.1. We call a restriction of a partition $\mathcal{A} \in \Pi(S)$ to a subset $c \subseteq S$

$$R_c(\mathcal{A}) = \{a | a \in \mathcal{A}, a \cap c \neq \emptyset\} \tag{76}$$

And we call the overlap of a partition with a subset

$$O_c(\mathcal{A}) = \bigcup_{a \in \mathcal{A}, a \cap c \neq \emptyset} a = \bigcup R_c(\mathcal{A}) \tag{77}$$

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

Theorem 0.3. Given partitions $\mathcal{A}, \mathcal{C} \in \Pi(S)$, $\mathcal{A}$ refines $\mathcal{C}$ if and only if

$$O_c(\mathcal{A}) = c \quad \forall c \in \mathcal{C} \tag{78}$$

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

$$\bigcup_{a \in \mathcal{A}, a \cap c \neq \emptyset} a = c \tag{79}$$

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

$$\{R_c(\mathcal{A}) | c \in \mathcal{C}\} \in \Pi(\mathcal{A}) \tag{80}$$

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 0.2. 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

$$O_c(\mathcal{A}) = O_c(\mathcal{B}) = c \quad \forall c \in \mathcal{C} \tag{81}$$

Here is an example of a simultaneous decomposition of partitions:

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

$$\mathcal{B} = 1 | 2\ 3 \mid 4 | 5 \tag{83}$$

$$\mathcal{C} = 1\ 2\ 3 \mid 4\ 5 \tag{84}$$

Theorem 0.5. If T, V are bounded tables, and if $\mathcal{C}$ is any simultaneous decomposition of $\mathcal{G}_T, \mathcal{G}_V$, so that $\mathcal{G}_T$ is partitioned by $\{R_c(\mathcal{G}_T) | c \in \mathcal{C}\}$ and so we may simultaneously decompose the tables according to $\mathcal{C}$

$$T = \bigoplus_{c \in \mathcal{C}} T_c = \bigoplus_{c \in \mathcal{C}} (R_c(\mathcal{G}_T), U_T, L_T) \tag{85}$$

$$V = \bigoplus_{c \in \mathcal{C}} V_c = \bigoplus_{c \in \mathcal{C}} (R_c(\mathcal{G}_V), U_V, L_V) \tag{86}$$

Where each T_c, V_c is a table acting on all the rows in $c \in \mathcal{C}$. Then $T \geq V$ if and only if $\mathcal{M}(T) \subseteq \mathcal{M}(V)$ if and only if for all $c \in \mathcal{C}$ we have $\mathcal{M}(T_c) \subseteq \mathcal{M}(V_c)$

Proof. Use lemma 0.2. □

We will now describe a specific type of decomposition.

Definition 0.3. For any set S let $\mathcal{A}, \mathcal{B}$ be partitions of S, that is to say $\mathcal{A}, \mathcal{B} \in \Pi(S)$. Then we say that $\mathcal{A}$ and $\mathcal{B}$ "don't partially overlap" if

$$\forall a \in \mathcal{A} \forall b \in \mathcal{B} \quad a \cap b \neq \emptyset \Rightarrow a \subseteq b \vee b \subseteq a \tag{87}$$

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 $\mathcal{P} = 1|2|34$ and $\mathcal{Q} = 12|3|4$ don't partially overlap either. If we make a third partition $\mathcal{D} = 12|34$ then 'restrict' $\mathcal{P}$ and $\mathcal{Q}$ to each of the sets, so for $12 \in \mathcal{D}$ we get 12 for $\mathcal{Q}$ and $1|2$ for $\mathcal{P}$, we see that these restrictions are going to have that one is a refinement of another.

Theorem 0.6. Every non partially overlapping pair of partitions $\mathcal{A}, \mathcal{B} \in \Pi(S)$ has a simultaneous decomposition $\mathcal{C} \in \Pi(S)$ where for each $c \in \mathcal{C}$ one of $R_c(\mathcal{A}), R_c(\mathcal{B})$ is just $\{c\}$ and another is a partition of $c \in \mathcal{C}$. We will call $\mathcal{C}$ a refinement decomposition, because each pair of restrictions has that one refines another.

Proof. Let $\mathcal{A}, \mathcal{B} \in \Pi(S)$ have no partial overlapping. Then for any $a \in \mathcal{A}$ let

$$S(a, \mathcal{B}) = \bigcup_{b \in \mathcal{B}, b \subseteq a} b \subseteq S \tag{88}$$

Notice that either $S(a, \mathcal{B}) = \emptyset$, or $S(a, \mathcal{B}) = a$. This is because $\mathcal{B}$ is a partition, if the set of $b \in \mathcal{B}$ such that $b \subseteq a$ is nonempty, it must cover all of a, and it cant be bigger than a because otherwise there'd be partial overlappings. Consider also

$$S(b, \mathcal{A}) = \bigcup_{a \in \mathcal{A}, a \subseteq b} a \tag{89}$$

Now if $S(a, \mathcal{B}) \neq \emptyset$ then for every $b \subseteq a$ we have $S(b, \mathcal{A}) = \emptyset$ and vice versa. Let $\mathcal{C} = \emptyset$. Pick any $a \in \mathcal{A}$ such that $S(a, \mathcal{B}) \neq \emptyset$, or failing that pick any $b \in \mathcal{B}$ such that $S(b, \mathcal{A}) \neq \emptyset$. Then add a (or b) to $\mathcal{C}$. Then remove every set in both $\mathcal{A}$ and $\mathcal{B}$ which intersects the chosen a or b. Repeat this process until $\mathcal{C}$ covers S. First, we aim to show that the sets of $\mathcal{C}$ are disjoint and hence that $\mathcal{C} \in \Pi(S)$ is a partition. By contradiction if we did not pick disjoint sets, then one set would be a subset of another (since no partial overlaps are allowed between $\mathcal{A}, \mathcal{B}$ nor within themselves), and so $c \subseteq d \in \mathcal{C}$. But it is impossible that $c \subseteq d$ since if this were the case then either $S(c, \mathcal{B}) = \emptyset$ or $S(c, \mathcal{A}) = \emptyset$ respectively (depending on where c came from) and so c could never have been chosen.
Now we will show $\mathcal{C}$ simultaneously decomposes $\mathcal{A}, \mathcal{B}$. For any $c \in \mathcal{C}$ we see that

$$O_c(\mathcal{A}) = \bigcup_{a \in \mathcal{A}, a \cap c \neq \emptyset} a = O_c(\mathcal{B}) = c \tag{90}$$

Because every set in $\mathcal{A}$ must either be wholly contained within c or not contained at all, and $\mathcal{A}$ is a partition, the union of those sets is exactly equal to c. Hence $\mathcal{C}$ is a decomposition. Now to finish the proof we take any $c \in \mathcal{C}$ and obtain $R_c(\mathcal{A}), R_c(\mathcal{B})$. If $c \in \mathcal{C}$ was originally picked from $\mathcal{A}$, so $c \in \mathcal{A}$, then clearly

$$R_c(\mathcal{A}) = \{a | a \in \mathcal{A}, a \cap c \neq \emptyset\} = \{c\} \tag{91}$$

And so

$$R_c(\mathcal{B}) = \{b | b \in \mathcal{B}, b \cap c \neq \emptyset\} \tag{92}$$

Must be a collection of subsets of c because otherwise there would be partial overlaps. And the case is vice versa for if $c \in \mathcal{B}$ was picked from $\mathcal{B}$. □

Using theorem 0.7 with 0.6 and 0.3, we can decompose any two bounded contingency tables whose row groupings don't have partial overlappings, and then order the blocks individually, thus giving the ordering of the tables as a whole.

Bibliography

1. Fecho, K., Xu, H., Sinha, M., Sharma, P., Schmit, P., Ramsey, S., et al. (2021). *An Approach for Open Multivariate Analysis of Integrated Clinical and Environmental Exposures Data*. Informatics in Medicine Unlocked, 26. doi:https://doi.org/10.1016/j.imu.2021.100733.

2. Weisstein, Eric W. "Multichoose." From MathWorld–A Wolfram Web Resource. https://mathworld.wolfram.com/Multichoose.html

3. Martin, Jeremy. *Lecture Notes on Algebraic Combinatorics.* University of Kansas, 2010, https://jlmartin.ku.edu/LectureNotes.pdf.