Statistica & Applicazioni
Vol. I, n. 2, 2003

Counting and enumerating frequency tables with given margins

Francesca Greselin[§]

Summary:
The problem of finding the number of rectangular tables of non-negative integers with given row and column sums occurs in many interesting contexts, mainly in combinatorial problems (counting magic squares, enumerating permutation by descents, etc.) and in statistical applications (studying contingency tables with given margins, testing for independence, etc.). In the present paper a new recursive argument is presented to produce a general expression for the number of m×n tables with given margins. The result has the same expressive force of the one presented by Gail and Mantel (1977), but, remarkably, the counting approach suggests, quite naturally, also a recursive algorithm to explicitly generate the entire class of tables. This work is a necessary step for studying a new measure of association, based on the relative position that a given table assumes in its class, endowed by an association ordering.

Keywords: *Contingency tables, frequency tables, enumeration of contingency tables with given margins, counting contingency tables with given margins*.

1. Introduction

Consider the integral m×n table **X** whose entries are count data for a two-way classification scheme. We shall be concerned first with the problem of counting up, then of enumerating all contingency tables with given margins.

These classical problems arise when trying to introduce a new approach to measure association, namely evaluating the suitability to consider, as an association index, the relative position a distribution assumes in the set of all

[§] Dipartimento di Metodi Quantitativi per le Scienze Economiche e Aziendali, Università degli Studi di Milano Bicocca, Piazza dell'Ateneo Nuovo 1, Milano, I-20126, Italy (e-mail: francesca.greselin@unimib.it).

contingency tables with given margins, endowed by a partial or total ordering of dependence (Greselin F. and Zenga M. 2001, 2002). The new association measure requires the knowledge of *how many* tables are consistent with the given margins, and also *what they exactly are* as in the context of Fisher-Yates test (and, more generally, exact inferential methods for contingency tables) in order to rank each of them with respect to a chosen association ordering.

2. Preliminaries and notation

Given the m×n table **X** with m and n greater than one, let x_{ij} be its entry in row i and column j. Throughout, x_{ij} are assumed to be non negative integers. Let **1** be a column vector of ones and define the row totals $\mathbf{r} = \mathbf{X1}$, the column totals $\mathbf{c} = (\mathbf{X}^T\mathbf{1})^T$, and the total population size $N = \mathbf{c1} = \mathbf{r}^T\mathbf{1}$.

Let $\Sigma_{m,n}(\mathbf{r},\mathbf{c})$ denote the reference set of all m×n tables satisfying the same marginal constraints **r** and **c** as **X**:

$$\Sigma_{m,n}(r,c)=\{\mathbf{A} \mid \mathbf{A}=\{a_{ij}\} \text{ is } m\times n,\ a_{ij}\epsilon N,\ \mathbf{A1} = \mathbf{r};\ (\mathbf{A}^T\mathbf{1})^T = \mathbf{c}\} \qquad (1)$$

Similarly, $\#_{m,n}(\mathbf{r},\mathbf{c})$ will designate the number of matrices $\mathbf{A} \in \Sigma_{m,n}(\mathbf{r},\mathbf{c})$:

$$\#_{m,n}(\mathbf{r},\mathbf{c}) = \text{Card } \{ \Sigma_{m,n}(r,c) \} \qquad (2)$$

As usual, the following representation for table **A** can be chosen:

a_{11}	a_{12}	…	a_{1j}	…	a_{1n}	r_1
…	…		…		…	…
a_{i1}	a_{i2}	…	a_{ij}	…	a_{in}	r_i
…			…		…	…
…			…		…	…
a_{m1}	a_{m2}	…	a_{mj}	…	a_{mn}	r_m
c_1	c_2	…	c_j	…	c_n	

3. Brief review of the literature on determining the cardinality of $\Sigma_{m,n}$ (r,c)

The problem of counting all tables with given margins has received a fair amount of attention in the literature on combinatorics. It may be noted that their number is equal to the number of ways that m types of objects can be distributed in n boxes, so that there are altogether r_i objects of type i and there are c_j objects in box j.

Most of the literature was primarily concerned with finding exact values for $\#_{m,n}(r,c)$ in various special circumstances.

A number of authors studied the particular case of square matrices n×n with all marginal totals equal to t, i.e. *stochastic matrices*, in statistics, or *magical arrays*, in discrete mathematics. MacMahon (1915, 1960), Anand, Dumir and Gupta (1966), Stein and Stein (1970), Smith (1971), Gupta and Nath (1972), Nath and Iyer (1972), Abramson and Moser (1973) solved in a closed form only simple cases. For example, Stein and Stein gave the exact values of $\#_{n,n}(r,c)$ for a finite set of values of n and t. Just to give an idea of the size of this set, all the 5×5 tables with margins $\mathbf{r} = \mathbf{c} = (15,15,15,15,15)$ are $1{,}9208\ldots\times10^{50}$. Everett and Stein (1971) provided an asymptotic formula for the number of integer stochastic matrices.

Carlitz (1966, 1971, 1972), Grimson (1971, 1972) and Gupta (1968, 1971) considered a more specific case: the enumeration of symmetric square matrices. Stanley (1973) proved that $\#_{n,n}(r,c)$ is a polynomial on t (the row and column sum) of degree $(n-1)^2$, then Dahmen and Micchelli (1998) re-proved this result, that was indeed conjectured by Gupta (1968), while Jackson and Van Rees (1975) presented a simplified computation for the coefficients of that polynomial. A comprehensive review can be found in Stanley (1986, 1997).

Another thread of investigations is based on the special circumstance of matrices of zeros and ones, analyzed by O'Neil (1969a), Brualdi (1980) and Snijders (1991), and works quoted therein. Kemperman and Kuba (1998) investigated the slightly broader case of two valued matrices, and they solved the 2×n and 3×n cases.

Referring to rectangular matrices, but yet restricting to the situation in which all whose rows sum to a given non-negative integer t and whose columns sum to a non-negative integer s, Edmonds (1977) solved the 2×n and the 3×n cases, showing that $\#_{m,n}(r,c)$ is a polynomial on t of degree (n-1) and 2(n-1) respectively. Mirsky (1968) determined the necessary and sufficient conditions for the existence of integral matrices whose elements, row-sums, and column-sums, all lay between prescribed bounds. Agresti and Wackerly (1977) and Agresti, Wackerly and Boyett (1979) gave maxima for several table dimensions m and n and N population size.

Finally, referring to the general case, Leti (1970) gave an exact expression for $\#_{m,n}(\mathbf{r},\mathbf{c})$ by an iterating process. A straightforward approach to table enumeration makes use of a recurrence: Gail and Mantel (1977) carry that out, leading to an exact and expressive formula. Macdonald (1979) provided an algorithm for computing $\#_{m,n}(\mathbf{r},\mathbf{c})$ and a formula for this number, based on complete symmetric functions. James and Kerber (1981), relating $\Sigma_{m,n}(\mathbf{r},\mathbf{c})$ to the number of double cosets, gave a group-theoretic result, expressing $\#_{m,n}(\mathbf{r},\mathbf{c})$ in terms of cardinality of Young subgroups. Mount (1994, 1999) made a great improvement in deriving closed expressions for the number of tables. He showed that the counting function is a piecewise polynomial of row and column sums (already given by Blackley (1964) in a different way), he proposed a technique for inferring the polynomials, and obtained an algorithm whose run-time depends on the order of the matrix only and not on the magnitude of the components of **r** (or **c**, supposing m and n fixed).

In addition to the attempts aimed to the analytical formulation of $\#_{m,n}(\mathbf{r},\mathbf{c})$, useful approximations and asymptotic expressions for the number of tables - in its general formulation - were investigated by O'Neil (1969b), Békéssy et al. (1972), Bender (1974), Good (1976), Good and Crook (1977) and, more recently, by Diaconis and Efron (1985). Mount (1999) proposed a new estimate and compared his approximate evaluations to the results by Gail and Mantel and Diaconis and Efron on a set of tables with their respective exact values of $\#_{m,n}(\mathbf{r},\mathbf{c})$.

The most up-to-date review over all these topics - that have seen active development also in recent years - is offered in Diaconis and Gangolli (1995).

4. Counting the number of m×n tables with given margins

A well-known elementary result that represents a first step in approaching the more complex problem is the cardinality $C_k(N)$ of the class composed by all k-tuples of non negative integers with given sum N.

4.1 *Cardinality of k-tuples of integers with given sum*

Let N be the value of the given sum. So, for these k-tuples:

- the same integer value can be assumed by two or more elements in the k-tuple;
- two k-tuples with the same elements are considered different if their elements are placed in different orders;
- no element may assume a value greater than N.

Remarks:

- For k = 0, clearly $C_0(N) = 0$.
- For k = 1 there is only one 1-tuple, i.e. the integer N: $C_1(N) = 1$.
- For k = 2 there is a 2-tuple (i, N-i) for every integer i: $0 \leq i \leq N$, therefore $C_2(N) = 1+N$.
- For N = 0 there is only one possible k-tuple, composed by all null

elements, so $C_k(0) = 1$, for every k.

- For N > 0 each k-tuple contains at least a non-null element. If $0 < N < k$ all the k-tuples have at least k-N (and no more than k-1) null elements.

PROPOSITION:
The number $C_k(N)$ of k-tuples of integer numbers whose sum is a given integer N is:

$$C_k(N) = \binom{N+k-1}{k-1} \tag{3}$$

The thesis can easily be proved by induction on k, using a well-known property of binomial coefficients.

Remark 1:
An algorithm that would generate all k-tuples with given sum N can be developed recursively using the relation between k-tuples and (k-1)-tuples; the process stops when arriving at k=1, whose solution is known.
Remark 2:
$C_k(N)$ corresponds to the number of ways of distributing N like objects into k unlike cells.

4.2 *Cardinality of matrices of integers with given row and column sums*

Before proceeding to get an expression for $\#_{m,n}(r,c)$, the literature on the complexity of computing this cardinality will be recalled.

4.2.1 *Problem hardness*

Counting tables with given margins is #P-hard (even in the case when m or n is 2, using the familiar notion of #P-completeness, introduced by Valiant, 1979) as proved by Dyer, Kannan and Mount (1997). They obtained this relevant result by polinomially reducing the computation of the number $\#_{2,n}(\mathbf{r},\mathbf{c})$ of these matrices to the computation of the number $\nu_{2,n}(\mathbf{r},\mathbf{c})$ of integer points in a specific 2×n politope $P(\mathbf{r},\mathbf{c})$ with integer vertices, then they employed Dyer and Frieze's result (1991) about the #P-hardness of computing the $P(\mathbf{r},\mathbf{c})$ volume to obtain the thesis.

4.2.2 *A recursive method for counting*

Before approaching a general expression for $\#_{m,n}(r,c)$, some remarks are needed. For the nature of the problem:

$$\#_{m,n}(\mathbf{r},\mathbf{c}) = \#_{n,m}(\mathbf{c},\mathbf{r}). \tag{4}$$

Furthermore, the quantity $\#_{m,n}(\mathbf{r},\mathbf{c})$ is invariant to whatever reordering of the elements of vectors **r** and **c**. Without any loss of generality, one can pose $m \le n$, and for any $\mathbf{A} \in \Sigma_{m,n}(\mathbf{r},\mathbf{c})$:

- each row (column) of **A** is an n-tuple (m-tuple) with given sum, as in section 4.1;
- A can be composed by entries having the same repeated value;
- the arrangement of the entries is meaningful, hence two tables that differ only by the arrangement of their elements must be considered different;
- $a_{ij} \le \min(r_j,c_i)$, $\forall i,j$: $1 \le i \le m$, $1 \le j \le n$, as r_j and c_i are the maximum values in the row and column;
- the total sum of the entries of **A** is equal to the sum of the elements of vectors **r** and **c** (by definition):

$$\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij} = \sum_{i=1}^{m} r_i = \sum_{j=1}^{n} c_j. \tag{5}$$

The case m = n = 0 has no interest, so let us pose m = 1. Then $\#_{1,n}(\mathbf{r},\mathbf{c}) = 1$, for there is only one matrix whose entries are exactly those of vector **c**.

For m = n = 2, once the value of any of the four entries of **A** is assigned, the three others are determined by the margin constraints. All matrices with given margins are generated giving to a_{11} all the range of its possible values:

$$\max(0, c_1 - r_2) \le a_{11} \le \min(r_1, c_1). \tag{6}$$

The resulting expression for $\#_{2,2}$ is therefore:

$$\#_{2,2}(\mathbf{r},\mathbf{c}) = \min(r_1,c_1) - \max(0, c_1 - r_2) + 1. \tag{7}$$

Let us now discuss the case m = 2 and n > 2.

First of all, without loss of generality, let a_{11} be chosen in the range (6). The value of a_{21} is determined by the constraint on first column:

a_{11}	a_{12}	. . .	. . .	. . .	a_{1n}	r_1
$a_{21} = c_1 - a_{11}$	a_{22}					r_2
c_1	c_2	....	c_i	....	c_n	

For each possible value of a_{11} one has to consider all matrices with 2 rows and (n-1) columns:

a_{12}	. . .	. . .	. . .	a_{1n}	$r_1' = r_1 - a_{11}$
a_{22}					$r_2' = r_2 - a_{21}$
$c_1' = c_2$	. . . .	$c'_{i-1} = c_i$	. . . .	$c'_{n-1} = c_n$	

with new marginal constraints **r′** and **c′**, defined by:

$$\begin{cases} c_i' = c_{i+1} & \text{for } 1 \le i \le n-1 \\ r_1' = r_1 - a_{11} & \\ r_2' = r_2 - a_{21} = r_2 - (c_1 - a_{11}) & \end{cases} \tag{8}$$

Note that **c′** is (n-1)-dimensional (while vector **c** has n components), whereas **r′** still has two components as **r**.

Let now $a_{11}^{(k)}$ denote the k-th integer value for a_{11} obtained by ordering all possible values for a_{11}, with an index k ranging from 1 to $\#_{2,2}(\mathbf{r},\mathbf{c})$. In relation with $a_{11}^{(k)}$, let us consider the vectors **r′** e **c′**, directly determined by $a_{11}^{(k)}$, and denoted with $\mathbf{r'}^{(k)}$ and $\mathbf{c'}^{(k)}$. So we have obtained the relation:

$$\#_{2,n}(\mathbf{r},\mathbf{c}) = \sum_{k=1}^{\#_{2,2}(\mathbf{r},\mathbf{c})} \#_{2,n-1}(\mathbf{r'}^{(k)}, \mathbf{c'}^{(k)}) \tag{9}$$

The **general case**: **m > 2, n > 2.**

Let $\#_{m-1,n}(\mathbf{r'},\mathbf{c'})$ denote the number of tables with m-1 rows, with the constraints given by the vectors **r′** and **c′**, obtained from **r** and **c** by the following relations:

$$\begin{cases} c_i' = c_i - a_{1i} & \text{for } 1 \le i \le n \\ r_j' = r_{j+1} & \text{for } 1 \le j \le m\text{-}1 \end{cases} \tag{10}$$

For each n-tuple of non negative integer values assigned to the first row elements of the former table:

a_{11}	. . . .	. . . .	a_{1i}	. . . .	a_{1n}	r_1
a_{21}	. . . .	. . . .	a_{2i}	. . . .	a_{2n}	
						. . .
						r_j
						. . .
a_{m1}			a_{mi}		a_{mn}	r_m
c_1	. . . .	. . . .	c_i	. . . .	c_n	

one obtain $\#_{m-1,n}(\mathbf{r'}, \mathbf{c'})$ different solutions: as many as the number of tables with m-1 rows and satisfying constraints **r′**, **c′**:

a_{21}	…	…	a_{2i}		a_{2n}	$r'_1=r_2$
…			…			…
						$r'_{j-1}=r_j$
						…
						$r'_{m-1}=r_m$
$c'_1=c_1-a_{11}$		….	$c'_i=c_i-a_{1i}$	….	$c'_n=c_n-a_{1n}$	

Observe that **c** is an n-dimensional vector, and so is **c′**, whereas if **r** has m components then **r′** has only (m-1).

Let us denote with $\mathbf{R}=\{a_{11}, a_{12}, \ldots, a_{1n}\}$ the generic first row in matrix **A**. Such rows are vectors in $\mathbf{N}^n$, so we can arrange them in the usual lexicographic order, and we can call $\mathbf{R}^{(k)}$ the k-th vector in this ordered sequence. We will denote with $R_n(\mathbf{r}, \mathbf{c})$ the unknown number of these first rows in **A.** Consider now the vectors **r′** and **c′**, directly determined by $\mathbf{R}^{(k)}$, and denote them with $\mathbf{r}'^{(k)}$ and $\mathbf{c}'^{(k)}$. We can state the relation:

$$\#_{m,n}(\mathbf{r},\mathbf{c}) = \sum_{k=1}^{R_n(\mathbf{r},\mathbf{c})} \#_{m-1,n}(\mathbf{r}'^{(k)},\mathbf{c}'^{(k)}) \tag{11}$$

that recursively gives the number of m×n tables with given margins **r** and **c**.

This expression gives the solution of the problem for m>2; the recursive process ends when it arrives to the case m=2, for which the solution is known.

It remains to draw $R_n(\mathbf{r},\mathbf{c})$, i.e., the number of n-tuples that can be the first rows for **A**. The elements a_{1j} for j=1,…,n in such first rows must satisfy the following inequalities:

$$\begin{cases} a_{1j} \le c_j & \text{for } 1 \le j \le n \\ \sum_{j=1}^{n} a_{1j} = r_1 & \end{cases} \tag{12}$$

To obtain a recursive expression for $R_n(\mathbf{r},\mathbf{c})$, an explicit definition for $R_2(\mathbf{r},\mathbf{c})$ will be useful. Let a_{11} and a_{12} be the values of the generic first row in a two columns table, with a generic number of rows. Observe that, given a value to a_{11}, a_{12} is determined by the second constraint: $a_{12} = c_1 - a_{11}$. Each first row of **A** corresponds to an integer value assigned to a_{11}, satisfying the following inequalities:

$$\begin{cases} a_{11} \le \min\left(r_1,\, c_1\right) \\ a_{11} \ge \max\left(0,\ c_1 - \sum_{i=2}^{m} r_i\right) = \max\left(0,\ r_1 + c_1 - N\right) \end{cases} \tag{13}$$

The second inequality comes from the fact that the minimum value for a_{11} is reached when all the a_{i1} assume their respective maximum value r_i, their sum being given as:

$$\sum_{i=1}^{m} a_{i1} = c_1. \tag{14}$$

Hence:

$$R_2(\mathbf{r},\mathbf{c}) = \min\left(r_1,\, c_1\right) - \max\left(\ 0,\ r_1 +\ c_1 - N\right) + 1. \tag{15}$$

Finally, let us consider $R_n(\mathbf{r},\mathbf{c})$, focusing once more on the range of possible values for a_{11}. As done before, let $a_{11}^{(k)}$ denote the k-th ordered value of a_{11} in its range given by (13), and let the vectors $\mathbf{r}'$ and $\mathbf{c}'$, be directly determined by $a_{11}^{(k)}$, with $\mathbf{r}'^{(k)}$ and $\mathbf{c}'^{(k)}$. Therefore:

$$R_n(\mathbf{r},\mathbf{c}) = \sum_{k=1}^{R_2(\mathbf{r},\mathbf{c})} R_{n-1}(\mathbf{r}'^{(k)}, \mathbf{c}'^{(k)}) \tag{16}$$

where the vectors $\mathbf{r}'^{(k)}$ and $\mathbf{c}'^{(k)}$ are defined by:

$$\begin{cases} c_i'^{(k)} = c_{i+1}^{(k)} & \text{for } 1 \le i \le n-1 \\ r_1'^{(k)} = r_1^{(k)} - a_{11}^{(k)} & \\ r_j'^{(k)} = r_j^{(k)} & \text{for } 2 \le j \le m \end{cases} \tag{17}$$

so completing the determination of the cardinality of the considered class of tables.

5. Generation of the entire class of tables with given margins

5.1 *Brief review*

The problem of generating the isomarginal family of bivariate distributions has been widely analized: a number of authors, since Klotz (1967), Stein and

Stein (1970) with a branching algorithm, successively extended and clearly described by Good and Crook (1977), then Goodall (1968), March (1972), Boulton and Wallace (1973), Hancock (1975), Baker (1977), Pagano et al. (1981), and also Mehta and Patel (1983) with the network algorithm, all them have suggested algorithms for exhaustively stepping through the set $\Sigma_{m,n}(\mathbf{r},\mathbf{c})$, one table at a time. The main purpose of these works was directed toward the foundation of exact inferential methods for contingency tables. Most of these algorithms begin at some canonically constructed initial table and proceed by making small changes to cell entries so that the tables evolve monotonically in some linear order. The construction of an algorithm for the enumeration of the tables with given margins is very much dependent upon the chosen representation for $\Sigma_{m,n}(\mathbf{r},\mathbf{c})$ and in the literature three alternatives are presented:

- a tree in which each leave is a completely filled matrix,
- a network consisting of nodes and arcs in which each table is a path (i.e. a sequence of arcs) from the source node to the sink node,
- a convex subset of lattice points in $R^{m\times n}$.

For a detailed survey of algorithms for total enumeration, see Verbeek and Kroonenberg (1985).

Of course, the set $\Sigma_{m,n}(\mathbf{r},\mathbf{c})$ is large, and complexity considerations (for a thorough discussion see section 9 in Diaconis and Gangolli, 1995) may rule the algorithms out of success. The number of possible tables increases factorially fast as m, n, or the total population size N increases. However, in statistical applications, one is frequently in the situation in which many subjects are classified into a small number of categories, so m and n are given and N can vary. In these circumstances the problem is hence theoretically tractable: we were pleasantly surprised at how often the complete enumeration of $\Sigma_{m,n}(\mathbf{r},\mathbf{c})$ with respect to real bivariate statistical data is available, in a reasonable computing time.

5.2 *Generating from counting*

The nature of the present approach to the problem of counting can lead to gear an algorithm for generating the entire set of bivariate distributions with given margins. Recursion is not only a particularly powerful means in mathematical definitions but also in formulating algorithms. It is well known that recursive algorithms are primarily appropriate when the problem to solve, or the data structure to process, are already defined in recursive terms. The structure of the algorithm can hence be represented by a tree, whose definition is now briefly recalled.

A tree structure with base type T is either (Wirth,1976):

- the empty structure;

- a node of type T with a finite number of associated disjoint tree structures of base type T, called subtrees, represented by the following graph:

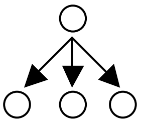

Figure 1. *Graph representation of a tree, with a root and three subtrees.*

Let us consider the tree:

- whose base type is a table with m rows and n columns;
- whose root[1] is the m×n table with unassigned internal values;
- whose nodes at level 1 correspond to the tables in which the first element a_{11}, is given. The i-th branch that exits from the root, from left to right, points to the i-th table whose value of a_{11} is the i-th among the $R_n(\mathbf{r},\mathbf{c})$ possible values;
- each node at level *l*, with $0 < l < (n-1)$, has as many ordered descendants as the possible values of $a_{1(l+1)}$. This position completely defines level *l*+1 of the tree and corresponds to the partial construction of a double table, determining a portion of its first row.

Now, giving a value to a_{1i} for i = 1,…, n-1 also means having given the value of a_{1n} (because of the constraint on c_1) so that nodes at level n-1 refers to all the tables with first row assigned and satisfying the **r** and **c** constraints. Their number is $R_n(\mathbf{r}, \mathbf{c})$. Scanning these first rows from left to right in level n-1 of the tree, they appear in lexicographic order as vectors in space $\mathbf{N}^n$.

To understand how the algorithm proceeds (choosing an example without any statistical meaning, just to be able to draw it in a page), let **c** = (3,2,2) and **r** = (3,1,3) be the two marginal distributions representing the row and the column constraints, with m = n = 3.

In the following representation, the tree is built up to level m – 1 = 2, and all possible first rows for the tables appear:

[1] *level*, *descendant* node, *root* and *ancestor* definitions are here recalled:
The *root* of a tree is defined to be at *level* 0. A node y which is directly below node x is called a (direct) *descendant* of x; if x is at *level* i, then y is said to be at *level* i+1. Inversely, node x is said to be the (direct) *ancestor* of y. Cfr. N. Wirth (1976).

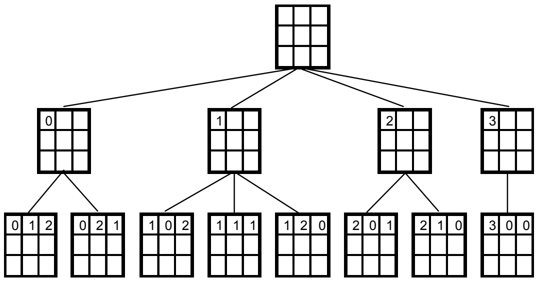

Figure 2. *Tree representation for the generation of all bivariate tables with* **c** *= (3,2,2) and* **r** *= (3,1,3). In this figure the tree is only partially built up to level m – 1 = 2. All possible first rows for the tables appear in level 2.*

Our goal now is to accomplish the definition of the tree. Let i assume an integer value in (1,…,m-2), where $i \cdot (n-1) \leq l < (i+1) \cdot (n-1)$, hence the generic level l can be expressed as: l = i · (n-1) + j with j opportune integer in (0,1,…,(n – 2)). Each node at level l has as many descendants as the possible values of $a_{(i+1)(j+1)}$. So nodes at level l+1 are completely defined. Observe that the range of values for $a_{(i+1)(j+1)}$ is determined by the new constraints **r′** and **c′** given by (10), or the analogue obtained after i rows of recursion.

Finally, as the assignment of all values $a_{(i+1)(j+1)}$ with j = 0,1,…,(n–2) means also the assignment of $a_{(i+1)(n)}$ for the constraint $c_{(i+1)}$, the nodes at level $l = (i+1) \cdot (n-1)$ correspond to all the tables with i+1 rows completely determined and satisfying the **r** and **c** constraints.

Reaching level l = (m-1)(n-1) and considering the constraint represented by **r**, all the m×n tables are wholly defined.

Throughout this representation, the cardinality of the class of tables with given margins is the number of leaves[2] of the corresponding tree, i.e., the number of elements appearing at level (m-1)(n-1). In other words, it equals the cardinality of nodes of the tree whose path length[3] is (m-1)(n-1).

In the previous example, the whole tree for **c**=(3,2,2) and **r**=(3,1,3), is:

[2] If a node in a tree has no descendants, it is called a terminal element or a *leaf*; and an element which is not terminal is an *interior node*.

[3] The number of branches or edges which have to be traversed in order to proceed from the root to a node x is called the *path length of x*.

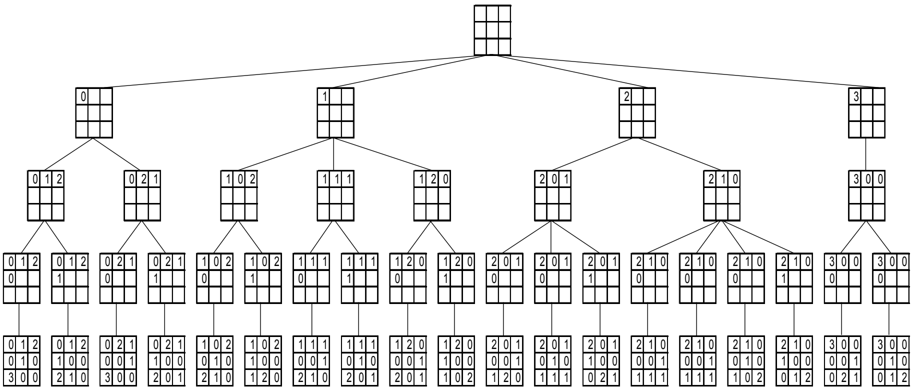

Figure 3. *Complete tree representation for the generation of all bivariate tables with **c** = (3,2,2) and **r** = (3,1,3).*

An algorithm that generates all the tables with given margins can be easily built on such tree structure, following its preorder traversal[4] and enumerating all tables, from the first to the last one, in lexicographic order.

A software program has been developed according to the algorithm and run on a 1800 MHz Pentium PC. The following table reports the running time of the program, for values of m, n and N in some range of statistical association analysis.

Table 1. *Computing time to generate some classes of tables with given margins, for different dimensions m and n, and constraints **r** and **c**.*

m	n	N	**r**	**c**	Time to generate $\Sigma_{m,n}(\mathbf{r},\mathbf{c})$ (in seconds)
3	2	1280	(620,660)	(384,768,128)	0.35
4	2	672	(310,362)	(176,152,158,186)	22.383
3	3	432	(256,64,112)	(144,176,112)	147.192
3	2	720	(180,540)	(240,240,240)	0.22
3	3	134	(27,51,56)	(61,45,28)	14.45
4	2	300	(120,180)	(54,66,141,39)	18.57
3	4	56	(15,9,21,11)	(12,26,18)	28.44
4	4	125	(12,34,21,58)	(15,37,44,29)	398997.76

[4] Three main orderings, conveniently expressed in recursive terms, emerge naturally from the structure of a tree. Referring to the general tree in which R denotes the root and A and B denote the left and right subtrees, the three orderings are:

1. Preorder: R,A,B (visit root before the subtrees)
2. Inorder: A,R,B
3. Postorder: A,B,R (visit root after the subtrees).

The results are very encouraging because the program, while generating the set of matrices in $\Sigma_{m,n}(\mathbf{r},\mathbf{c})$, also computes a variety of statistical indices (contingencies, association measures and ranking of tables in partial association ordering) for each of them in a reasonable amount of time. As remarked above, the number of tables with given margins grows factorially with m and n and the computing time become heavy for m and n ≥ 4.

Further research can be devoted to the enumeration of tables of higher dimension. The particular form of the algorithm extends quite readily to the enumeration of three and higher dimensional tables conditional on their margins. The coding of these extensions would be but a direct extension of the present code: whenever the present algorithm generates an m×n table $\{a_{ij}\}$, a new procedure can consider it as the first slice of a m×n×p table, so that all three-way tables can be built up generating all p×n tables with given first column $\{a_{ij1}\}=\{a_{ij}\}$ and given margins. The recursive structure of the method can afford successively four-way tables and so on. These considerations attain to the feasibility of the extension of the software to cope with contingency tables formed by cross-classifying three or more categorical variables.

With reference to the complexity of the same problem, once more, as the cardinality of the class of tables increases exponentially with the number of categories in each variable, we expect that the needed computing time becomes rapidly critical.

6. Concluding remarks

The present paper gives a recursive method to obtain the number of frequency tables with given margins. This approach for counting, naturally suggested an algorithm to generate all the matrices. The software based on this algorithm provides a tool to produce the class of all considered tables. Our aim is to exploit it in order to investigate how different association orderings can order contingency tables in the same class. Furthermore, a possible significant extension of this work might be to evaluate the behavior of a new category of association indices, based on the relative position a contingency table assumes in the context of its class. A sort of property of multiplicative invariance of those measures, for example, could be very useful to face the hardness of the problem of generating the entire class of tables.

Acknowledgement

The present work has been developed within a current research about descriptive analysis of contingency tables, led by professor M.Zenga, at Dipartimento di Metodi Quantitativi per le Scienze Economiche e Aziendali, Faculty of Economics, University of Milano Bicocca, partially supported by Italian M.I.U.R. grants (prot.200137492-001).

The author is especially grateful to M.Zenga and to G.Mauri for their helpful suggestions that have considerably encouraged the presentation of the paper.

References

Abramson M., Moser W.O.J. (1973). Arrays with Fixed Row and Column Sum. *Discrete Mathematics*, 6, 1-14.

Agresti A., Wackerly D. (1977). Some exact conditional tests of independence for R×C cross-classification tables. *Psychometrika*, 42, 111-125.

Agresti A., Wackerly D., Boyett J. (1979). Exact conditional tests for cross-classification: Approximation of attained significance levels. *Psychometrika,* 44, 75-83.

Anand H., Dumir V.C., Gupta H. (1966). A combinatorial distribution problem. *Duke Math. J.,* 33, 757-769.

Balmer D.W. (1988). Recursive enumeration of r×c tables for exact likelihood evaluation. *Applied Statistics*, 37, 290-301.

Baker R.J. (1977). Exact distributions derived from two-way tables. *J. Roy. Statist. Soc,.* Ser. C, 26, 199-206.

Békéssy A., Békéssy P., Komlòs J. (1972). Asymptotic enumeration of regular matrices. *Studia Sci. Math. Hung*., 7, 343-353.

Bender E.A. (1974). The asymptotic number of non-negative integer matrices with given row and column sums. *Discrete Math*., 10, 217-223.

Blackley S. (1964). Combinatorial remarks on partitions of a multipartite number. *Duke Math. J.,* 31, 335-340.

Boulton D., Wallace C. (1973). Occupancy of a regular array. *Computing J.,* 16, 57-63.

Brualdi R.A. (1980). Matrices of zeros and ones with fixed row and column sum vectors. *Linear algebra and its applications,* 33, 159-231.

Carlitz L. (1966). Enumeration of symmetrical arrays. *Duke Math. J.*, 33, 771-782.

Carlitz, L. (1971). Enumeration of symmetrical arrays. *Duke Math. J.,* 38, 717-731.

Carlitz L. (1972). Enumeration of 3×3 arrays. *Fibonacci Q.,* 10, 489-498.

Dahmen W., Micchelli C.A. (1998). Diophantine equations and multivariate splines. *Trans. Amer. Math. Soc.,* 308, 509-532.

Diaconis P., Efron B. (1985). Testing for independence in a two-way table: new interpretations of the Chi-Square statistic. *Annals of Statistics*,13, 3, 845-874.

Diaconis P., Gangolli A. (1995). Rectangular arrays with fixed margins, in: *Discrete Probability and Algorithm.* D. Aldous, P. Diaconis, J. Spencer and J.M. Steele Eds., Springer-Verlag, Berlin / New York, 15-41.

Dyer M., Frieze A. (1991). Computing the volume of convex bodies: a case where randomness provably helps. *Proceedings of Symposia in Applied Mathematics,* vol.44.

Dyer M., Kannan R., Mount J. (1997). Sampling contingency tables. *Random Structures and algorithms,*10, 487-506.

Edmonds F.C.(1977) Enumeration of arrays of a given size. *Discrete Math.,* 18, 1, 1-22.

Everett C.J., Stein P.R. (1971). The asymptotic number of integer stochastic matrices. *Discrete Math.,* 1, 55-72.

Gail M., Mantel N. (1977). Counting the number of r×c contingency tables with fixed margins. *Jour. Amer. Statist. Assoc.,* 72, 859-862.

Good I. J.(1976), On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Stat.,* 4, 1159-1189.

Good I.J., Crook J. (1977). The enumeration of arrays and a generalization related to contingency tables. *Discrete Math*., 19, 23-65.

Goodall D.W. (1968). Contingency tables and computers. *Biometrie – Paximetrie*, 9, 113-119.

Greselin F., Zenga M. (2001). Partial and total ordering relations in the Fréchet class. *Rapporto N°XXXVIII/2001 del Dipartimento di Metodi Quantitativi per le Scienze Economiche e Aziendali*. Università di Milano-Bicocca.

Greselin F., Zenga M. (2002). Measures of association in the Fréchet class. *Atti della XLI Riunione Scientifica della SIS.* Cleup, Padova.

Grimson R.C. (1971). Some results on the enumeration of symmetric arrays. *Duke Math. J.*, 38, 711-715.

Grimson R.C. (1972). Enumeration of symmetric arrays with different row sums. *Rend. Sem. Mat. Univ. Padova*, 48, 105-112.

Gupta H. (1968). Enumeration of symmetric matrices. *Duke Math. J*., 35, 653-660.

Gupta H. (1971). On the enumeration of symmetric matrices. *Duke Math. J.*, 38, 709-710.

Gupta H., Nath G.L. (1972). Enumeration of stochastic cubes. *Notices of the Amer. Math. Soc.,* 19, A-568.

Jackson D.M., Van Rees G.H.J. (1975). The enumeration of generalized double stochastic non-negative integer square matrices. *SIAM Jour. Comp*., 4, 475-477.

James G.D., Kerber A. (1981). *The representation theory of the symmetric group.* Addison Wesley, Reading, MA.

Kemperman J.H.B., Kuba A. (1998). Reconstruction of two-valued matrices from their two projections. *International Journal of Imaging systems and technology,* 9 (2-3), 110-117.

Klotz J.H. (1967). The Wilcoxon, ties, and the computer. *J. Amer. Statist. Assoc.,* 61, 772-787; Corr. 62 1520-1521.

Leti G. (1970). La distribuzione delle tabelle della classe di Fréchet. *Metron,* vol. XXVIII, n.1-4.

Macdonald J. (1979). *Symmetric Functions and Hall Polynomial*. Clarendon Press, Oxford.

MacMahon P.A. (1915, 1960). *Combinatory Analysis.* University Press, Cambridge, England, 2 volumes, or reprinted as 2 volumes in 1 by Chelsea.

March D. (1972). Exact probabilities for r×c contingency tables. *Communications of the ACM*, 15, 991-992.

Mehta C., Patel N. (1983). A network algorithm for performing Fisher's exact test in r×c contingency tables. *Jour. Amer. Statist. Assoc*., 78, 427-434.

Mirsky L. (1968). Combinatorial theorems and integral matrices. *Jour. Comb. Theory,* 5, 30-44.

Mount J. (1994). *Application of convex sampling to optimization and contingency table generation / counting*. PhD thesis, Dept. of Computer Science, Carnegie Mellon University, t.r. number CMU-CS-95-152.

Mount J. (1999). Fast unimodular counting. (a copy of this work is available by e-mail from the author: jmount@esgear.com).

Nath G.B., Iyer P.V.K. (1972). Note on the combinatorial formula for nHr. *J. Austral. Math. Soc.,* 14, 264-268.

O'Neil P.E. (1969a). *Asymptotic enumerations for 0-1 matrices.* Ph.D. thesis, Ch. 2, The Rockefeller University, unpublished.

O'Neil P.E. (1969b). Asymptotics and random matrices with row-sum and column-sum restrictions. *Bull. of Amer. Math Soc.* 75, 1276-1282.

Pagano M. and Taylor Halvorsen K. (1981). An algorithm for finding the exact significance levels of r×c contingency tables. *Jour. Amer. Statist. Assoc.,* 78, 931-934.

Smith D.A. (1971). The number of 4×4 magic squares. *Notices Amer. Math. Soc., 18*, 90-98.

Snijders T.A.B. (1991). Enumeration and simulation methods for 0-1 matrices with given marginals. *Psychometrika,* 56, 397-417.

Stanley R.P. (1973). Linear homogeneous diophantine equations and magic labelling of graphs. *Duke Math. J.,* 40, 607-632.

Stanley R.P. (1986). *Enumerative combinatorics*. Wadsworth, Monterey, CA - reprinted in Cambridge studies in advanced mathematics by Cambridge Univ. Press, 1997.

Stein M.L., Stein P.R. (1970). *Enumeration of stochastic matrices with integer elements.* Los Alamos Scientific Laboratory Publ. Nr. LA-4434.

Valiant L.G. (1979). The complexity of computing the permanent. *Theoretical computer science*, 8, 189-201.

Verbeek A., Kroonenberg P. M. (1985). A survey of algorithms for exact distributions of test statistics in r×c contingency tables with fixed margins. *Computational Statistics and Data analysis*, 3, 159-185.

Wirth N. (1976). *Algorithms + Data Structures = Programs*. Prentice-Hall Inc., New Jersey.