NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

The number of $2 \times c$ tables with given margins

by

Øyvind Bakke and Mette Langaas

PREPRINT
STATISTICS NO. 11/2012

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL
http://www.math.ntnu.no/preprint/statistics/2012/S11-2012.pdf
Øyvind Bakke has homepage: http://www.math.ntnu.no/∼bakke
E-mail: bakke@math.ntnu.no
Address: Department of Mathematical Sciences, Norwegian University of Science
and Technology, N-7491 Trondheim, Norway.

The number of $2 \times c$ tables with given margins

Øyvind Bakke Mette Langaas

Abstract

We provide an elementary proof of a formula for the number of possible $2 \times c$ contingency tables with given row and column sums. Further, we show that the number of $r \times c$ contingency tables with given row sums is maximal as a function of column sums when column sums are as equal as possible. If only the sum of all table entries is given, the number of tables is maximal when also row sums are as equal as possible. The knowledge of those numbers is useful for determining which method to use for statistical testing of association in a contingency table.

1 Introduction

Contingency tables are used in statistics to summarize data that are simultaneously classified according to two categorical variables. For example, a 2×3 contingency table may display frequencies of subjects of a case–contol study, classified into three genotypes (columns) and into diseased or healthy (rows).

There are numerous methods of performing statistical tests of whether the two variables are associated. One method proposed for testing association in genome-wide association studies is exact permutation testing (also called conditional testing) (Tian et al., in prep.). In such tests, probabilities of all contingency tables having the same row and column margins as the given table, and a test statistic not less than that of the given table, are added to get a p-value.

To choose between an exact permutation test and other tests, the number of tables with given margins (Theorem 1) is of interest, as is the worst-case number of such tables when only row sums are known (Theorem 2). For example, the number of cases (diseased) and controls (healthy) may be known ahead of a study, but not genotypes. The number of tables with given margins is often surprisingly low. For a study of 1000 cases and 1000 controls, the number is at most 334,334 for 2×3 tables.

Gail and Mantel (1977) gave approximate formulas for the number of contingency tables of any size with given margins, as well as exact recursive formulas requiring summation of a large number of terms. There are also algorithms to find approximate numbers of contingency tables with given margins (see Barvinok et al., 2010; Dyer et al., 1997). Diaconis and Gangolli (1995) gave a review of literature on counting tables with given margins, and many newer references were given by Greselin (2003).

2 The number of tables

Note. The binomial coefficient $\binom{t}{s}$ is taken to be 0 when $t < s$.

Theorem 1. *Let r and c be positive integers. We consider $r \times c$ tables of nonnegative integers.*

(a) *The number of tables having total sum of entries N is $\binom{N+rc-1}{rc-1}$.*

(b) *The number of tables having row sums $(n_1, \ldots, n_r)$ is $\prod_{i=1}^{r} \binom{n_i+c-1}{c-1}$.*

(c) *For $r = 2$, the number of tables having row sums (n_1, n_2) and column sums $(m_1, \ldots, m_c)$ is*

$$\sum_{S \subset \{1,\ldots,c\}} (-1)^{|S|} \binom{n_1 + c - 1 - |S| - \sum_{j \in S} m_j}{c-1},$$

where the sum is over all proper subsets (including the empty set) of $\{1, \ldots, c\}$, and $|S|$ denotes the cardinality of S. For 2×2 tables ($c = 2$), this number becomes $n_1 + 1 - \max(0, n_1 - m_1) - \max(0, n_1 - m_2) = \min(m_1, n_1) - \max(0, n_1 - m_2) + 1$, and for 2×3 tables ($c = 3$), it becomes $\binom{n_1+2}{2} - \binom{n_1-m_1+1}{2} - \binom{n_1-m_2+1}{2} - \binom{n_1-m_3+1}{2} + \binom{n_1-m_1-m_2}{2} + \binom{n_1-m_1-m_3}{2} + \binom{n_1-m_2-m_3}{2}$.

Proof. The statements of (a) and (b) follow from a well-known counting result: The number of ways to select, with repetition, n of k distinct objects is $\binom{n+k-1}{k-1}$, which is the number of ways to fill nonnegative integers having sum n in k cells.

The result of (c) is attributed to Mann (1994) by Diaconis and Gangolli (1995). The original proof is unknown to us, but the result follows readily by the inclusion–exclusion principle. Let T be the set of all $1 \times c$ tables $(x_1, \ldots, x_c)$ having row sum n_1. For $j = 1, \ldots, c$, let A_j be the subset of T consisting of all such tables with $x_j \geq m_j + 1$. Then the set of possible first rows of a $2 \times c$ table having the given margins is $T \cap \overline{A_1} \cap \cdots \cap \overline{A_c}$. The second row is determined by the first row, so by the inclusion–exclusion principle the number of tables is $\left|T \cap \overline{A_1} \cap \cdots \cap \overline{A_c}\right| = |T| + \sum_{\varnothing \subset S \subseteq \{1,\ldots,c\}} (-1)^{|S|} \left|\bigcap_{j \in S} A_j\right|$.

By (b) with $r = 1$, $|T| = \binom{n_1+c-1}{c-1}$. For $S \subseteq \{1, \ldots, c\}$, $\bigcap_{j\in S} A_j$ consists of all $1 \times c$ tables $(x_1, \ldots, x_j)$ having row sum n_1 and with $x_j \geq m_j + 1$ for all $j \in S$. But the number of such tables is the same as the number of $1 \times c$ tables $(x'_1, \ldots, x'_c)$ having row sum $n_1 - \sum_{j\in S} m_j - |S|$, by the one-to-one correspondence $x'_j = x_j - m_j - 1$ for $j \in S$ and $x'_j = x_j$ otherwise, which gives the terms stated in the Theorem. It only remains to observe that $\bigcap_{j=1}^{c} A_j = \varnothing$, which explains why we need only consider proper subsets $S \subset \{1, \ldots, c\}$ in the sum (and even fewer subsets if the lesser of the row sums n_1 and n_2 is substituted for n_1). □

Although we only provide a formula for $r = 2$ rows when both row and column sums are given, the formula can still be of use to find the number $t(\mathbf{n}; \mathbf{m})$ of $r \times c$ tables having row sums $\mathbf{n} = (n_1, \ldots, n_r)$ and column sums $\mathbf{m} = (m_1, \ldots, m_c)$ when $r > 2$. Choose a k, $1 < k < r$. Then $t(\mathbf{n}; \mathbf{m}) = \sum_{\mathbf{m}'} t(n_1, \ldots, n_k; \mathbf{m}')\, t(n_{k+1}, \ldots, n_r; \mathbf{m} - \mathbf{m}')$, where $\mathbf{m}'$ runs through all vectors such that $\mathbf{m}'$ and $\mathbf{m} - \mathbf{m}'$ are possible column sum vectors for two tables having row sums $(n_1, \ldots, n_k)$ and $(n_{k+1}, \ldots, n_r)$, respectively (or, equivalently, the possible rows of a $2 \times c$ table having row sums $n_1 + \cdots + n_k$ and $n_{k+1} + \cdots + n_r$, respectively). This way, if necessary recurring into even smaller tables, the counting is broken down to tables having one or two rows. Of course, the table may be transposed if convenient.

This algorithm is due to David des Jardin and to John Mount (Diaconis and Gangolli, 1995, p. 27). We used the algorithm in combination with Theorem 1(c) on what was called "the hardest problem to date", a 5×3 table having a total sum of entries of 135 (Diaconis and Gangolli, 1995, pp. 26, 28). It took a few lines of R (R Development Core Team, 2012) code 25 minutes to arrive at the correct number of $1.225914 \cdot 10^{15}$ with the default precision of R on a standard desktop PC.

Theorem 2. *Let r and c be positive integers. We consider $r \times c$ tables of nonnegative integers having total sum of entries N.*

(a) *The number of tables, as a function of row and column sums, is maximal when the sums of any two rows differ by at most one and the sums of any two columns differ by at most one. For $2 \times c$ tables, this maximal number is given by Theorem* 1(c).

(b) *The number of tables having given row sums, as a function of column sums, is maximal when the sums of any two columns differ by at most one. For $2 \times c$ tables, this maximum number is given by Theorem* 1(c), *and if n is the lesser of the two row sums, then the maximum number is $n + 1$ for 2×2 tables and $\binom{n+2}{2} - 3\binom{n-m+1}{2} + r \max(n - m, 0)$ for 2×3 tables, where m and r are the unique integers such that $N = 3m + r$ and $0 \leq r < 3$.*

Proof. Let $T(\mathbf{n};\mathbf{m})$ denote the set of tables having row sums $\mathbf{n} = (n_1, \ldots, n_r)$ and column sums $\mathbf{m} = (m_1, \ldots, m_c)$, and let $t(\mathbf{n};\mathbf{m}) = |T(\mathbf{n};\mathbf{m})|$ denote the number of tables in $T(\mathbf{n};\mathbf{m})$.

If $c = 1$, the statement is trivial, so assume $c \geq 2$. We start by proving (b). The main part of the proof is to prove the following claim: If $m_i < m_j$, then $t(\mathbf{n};\mathbf{m}') \geq t(\mathbf{n};\mathbf{m})$, where $\mathbf{m}'$ is obtained from $\mathbf{m}$ by replacing m_i by m_i+1 and m_j by m_j-1. We assume without loss of generality that $m_1 < m_2$ and prove the claim for $i = 1,\ \ j = 2$.

First assume $c = 2$. For $r = 1$ the claim is trivial, so assume $r \geq 2$. There is a one-to-one correspondence between the subset of tables in $T(\mathbf{n},\mathbf{m})$ having its upper left entry less than n_1 and the the subset of $T(\mathbf{n},\mathbf{m}')$ having its upper right entry less than n_1. If the two upper entries are denoted (x, y), then the correspondence is given by $(x, y) \to (x + 1, y - 1)$ and leaving the rest of the table unchanged.

There are tables in $T(\mathbf{n};\mathbf{m})$ having $(x, y) = (n_1, 0)$ if and only if $n_1 \leq m_1$. The number of such tables is $t(n_2, \ldots, n_r; m_1 - n_1, m_2)$. Similarly, there are tables in $T(\mathbf{n};\mathbf{m}')$ having upper entries $(0, n_1)$ if and only if $n_1 < m_2$. The number of such tables is $t(n_2, \ldots, n_r; m_1 + 1, m_2 - 1 - n_1)$. Note that if tables are "lost" (there are tables having upper entries $(n_1, 0)$ in $T(\mathbf{n};\mathbf{m})$), then also tables are "gained" (there are tables having upper entries $(0, n_1)$ in $T(\mathbf{n};\mathbf{m}')$), since in that case $n_1 \leq m_1 < m_2$. Then the net gain is $t(n_2, \ldots, n_r; m_1 + 1, m_2 - 1 - n_1) - t(n_2, \ldots, n_r; m_1 - n_1, m_2)$.

The absolute value of the difference of column sums corresponding to the first term is $|m_2 - m_1 - n_1 - 2|$ and to the second $m_2 - m_1 + n_1$. Since $m_2 - m_1 > 0$, the first absolute value is less than or equal to the second (it is equal if $m_2 - m_1 = 1$ and less if $m_2 - m_1 \geq 2$). By repeatedly adjusting column sums $(m_1 - n_1, m_2)$ by adding 1 to the first and subtracting 1 from the second, eventually column sums m_1+1 and m_2-1-n_1 will be reached, in either order. By induction on the number of rows, $t(n_2, \ldots, n_r; m_1 + 1, m_2 - 1 - n_1) \geq t(n_2, \ldots, n_r; m_1 - n_1, m_2)$, proving the claim for $c = 2$.

Next, consider $c > 2$. Then $t(\mathbf{n}; m_1, \ldots, m_c) = \sum_{\mathbf{n}'} t(\mathbf{n}'; m_1, m_2)\, t(\mathbf{n} - \mathbf{n}'; m_3, \ldots, m_c)$, where $\mathbf{n}'$ runs through all vectors such that $\mathbf{n}'$ and $\mathbf{n} - \mathbf{n}'$ are possible row sum vectors for two tables having column sums (m_1, m_2) and $(m_3, \ldots, m_c)$, respectively, that is, the possible columns of a $r \times 2$ table having column sums $m_1 + m_2$ and $m_3 + \cdots + m_c$, respectively.

If $m_1 < m_2$, $t(\mathbf{n}'; m_1 + 1, m_2 - 1) \geq t(\mathbf{n}'; m_1, m_2)$ for all $\mathbf{n}'$ by the case $c = 2$. So, by comparing terms in the above sum, $t(\mathbf{n}; m_1 + 1, m_2 - 1, m_3, \ldots, m_c) \geq t(\mathbf{n}; m_1, m_2, m_3, \ldots, m_c)$, and the claim is proved.

By repeatedly replacing pairs m_i, m_j of column sums, where $m_i < m_j$, with new column sums m_i+1, m_j-1 until the sums of any two columns differ

by at most one, we eventually arrive at a maximum value of $t(\mathbf{n}, \mathbf{m})$ over all possible column sums $\mathbf{m}$. Note that this number is indeed a global maximum, since we would reach the same vector of column sums, up to permutation of vector entries, regardless of how the column sums were initially given. This concludes the proof of the main statement of (b).

By Theorem 1(c), the number of 2×2 tables having row sums (n_1, n_2) and column sums (m_1, m_2) is $n_1+1-\max(0,\, n_1-m_1)-\max(0,\, n_1-m_2)$. We can assume without loss of generality that $n_1 \leq n_2$. Then $2n_1 \leq n_1 + n_2 = m_1 + m_2$. If $|m_1 - m_2| \leq 1$, both $n_1 \leq m_1$ and $n_1 \leq m_2$, so that both subtracted terms vanish.

For the 2×3 case, writing $N = 3m + r$ with $0 \leq r < 3$, the maximum is attained when at least one column sum is m and the remaining r column sums are $m + 1$. Then $2n \leq 3m + r \leq 4m + 2$, so that $n - 2m \leq 1$, and the three last terms of the formula given in Theorem 1(c) vanish, assuming without loss generality that $n = n_1$. It is easily verified that the remaining terms are equal to $\binom{n+2}{2} - 3\binom{n-m+1}{2} + r \max(n - m, 0)$.

Part (a) is proven by applying (b) to any chosen vector of row sums, choosing column sums that differ by at most one. Then (b) is applied again to the transposed table, adjusting the row sums of the original table to obtain the maximum number of tables. Again, this number is indeed a global maximum, since we would reach the same vectors of row and column sums, up to permutation of vector entries, regardless of how the row sums were initially chosen. □

References

A. Barvinok., Z. Luria, A. Samorodnitsky, A. Yong, An approximation algorith for counting contingency tables, Random Struct. Algor. 37 (2010), 25–66.

P. Diaconis, A. Gangolli, Rectangular arrays with fixed margins, in: D. Aldous, P. Diaconis, J. Spencer, J.M. Steele (Eds.), Discrete Probability and Algorithms, Springer-Verlag, 1995, pp. 15–41.

M. Dyer, R. Kannan, J. Mount, Sampling contingency tables, Random Struct. Algor. 10 (1997), 487–506.

M. Gail, N. Mantel, Counting the number of $r \times c$ contingency tables with fixed margins, J. Am. Stat. Assoc. 72 (1977), 859–862.

F. Greselin, Counting and enumerating frequency tables with given margins, Statistica & Applicazioni 1 (2003), 87–104.

B. Mann, unpublished results (1994).

R Development Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org/.

J. Tian, C. Xu, H. Zhang, Y. Yang, Exact MAX tests in case–control association analysis using R: package MaXact, in prep.