Counting the Number of $r \times c$ Contingency Tables with Fixed Margins

MITCHELL GAIL and NATHAN MANTEL*

Exact and approximate methods are given for counting the number of $r \times c$ contingency tables with fixed margins. The approximate methods are extended to estimate the number of $r \times c \times s$ contingency tables with given first-order margins.

KEY WORDS: Enumeration of arrays with fixed margins; Exact conditional tests on contingency tables; Systems of equations in nonnegative integers.

1. INTRODUCTION

Consider an $r \times c$ contingency table with entries X_{ij}, row totals (margins) $m_i = \sum_j X_{ij}$, and column totals (margins) $t_j = \sum_i X_{ij}$, where $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$. We give exact and approximate methods for counting the number of tables consistent with the given margins. We became interested in this problem when trying to determine the exact conditional distribution of a function of the X_{ij}, which under independence, and conditional on the margins, follow the generalized hypergeometric distribution (Lehmann 1975, pp. 380–385). To determine whether an exact calculation was feasible, we needed to know how many tables were consistent with the given margins. Abramson and Moser (1973) and Good (1976) treat special cases.

2. EXACT ENUMERATION

The number of $r \times 2$ tables can be generated exactly from the recursive argument which follows. Simply stated, the number of possible tables with $i + 1$ rows can be built up from consideration of the number of tables with only i rows, so the entire $r \times 2$ problem can be solved by adding layer after layer until $i + 1 = r$. Let $N_i(t_1; m_1, m_2, \ldots, m_i)$ be the number of $i \times 2$ tables with margins t_1 and $m_1, m_2, \ldots, m_i$. Then

$$N_{i+1}(t_1; m_1, \ldots, m_{i+1}) = \sum_j N_i(t_1 - j; m_1, \ldots, m_i) \ , \quad (2.1)$$

where $j = 0, 1, \ldots, \min(m_{i+1}, t_1)$. Equation (2.1) gives the number of $(i + 1) \times 2$ tables with column total t_1 and margins $m_1, \ldots, m_{i+1}$. Equation (2.1) follows from the fact that the number of $(i + 1) \times 2$ tables with $X(i + 1, 1) = j$ is equal to the number of $i \times 2$ tables with column total $t_1 - j$.

To illustrate (2.1), consider the 4×2 table with $m_1 = 2$, $m_2 = 3$, $m_3 = 2$, and $m_4 = 1$. Without altering the problem, we relabel the margins $m_1 = 3$, $m_2 = 2$, $m_3 = 2$, $m_4 = 1$ in descending order. Table 1, based on (2.1), generates $N_4(t_1; 3, 2, 2, 1)$. The second column contains zeros for $t_1 > 3$, since no 1×2 tables with $m_1 = 3$ and $t_1 > 3$ are possible. The third column is obtained from the second by adding $m_2 + 1 = 3$ elements according to (2.1). The fifth column contains the numbers of possible 4×2 tables for $t_1 = 0, 1, 2, \ldots, 8$. Note that the distribution of $N_4(t; 3, 2, 2, 1)$ is symmetric. In particular, the number of 4×2 tables with column total $t_1 = 6$ and row margins 3, 2, 2, 1 is 9. The sum of the elements in column 5 is 72, the total number of possible 4×2 tables with unrestricted column totals, in agreement with the general result (3.6).

1. Computations for a 4 × 2 Table with Row Totals 3,2,2, and 1

T_1	$N_1(t_1;3)$	$N_2(t_1;3,2)$	$N_3(t_1;3,2,2)$	$N_4(t_1;3,2,2,1)$
8	0	0	0	1
7	0	0	1	4
6	0	0	3	9
5	0	1	6	14
4	0	2	8	16
3	1	3	8	14
2	1	3	6	9
1	1	2	3	4
0	1	1	1	1

For $c = 3$, the same reasoning used to obtain (2.1) leads to

$$N_{i+1}(t_1, t_2; m_1, \ldots, m_{i+1}) = \sum_{k_1} \sum_{k_2} N_i(t_1 - k_1, t_2 - k_2; m_1, \ldots, m_i) \ ,$$

where

$$0 \le k_\ell \le \min(m_{i+1}, t_\ell) \ , \quad \sum k_\ell \le m_{i+1} \ , \quad \text{and } \ell = 1 \text{ or } 2 \ . \quad (2.2)$$

The computational layout corresponding to Table 1 becomes a sequence of square arrays (t_1, t_2). The first such

[*] Mitchell Gail is a Medical Statistical Researcher at the Biometry Branch, National Cancer Institute, Landow Building C509, Bethesda, MD 20014. Nathan Mantel is Research Professor at George Washington University, 7979 Old Georgetown Road, Bethesda, MD 20014. This work was supported in part by Public Health Service Grant CA-15686 from the National Cancer Institute. The authors wish to thank Sylvan Green for programming equation (2.1), David Byar for suggesting this problem, and the editors for helpful comments.

© Journal of the American Statistical Association
December 1977, Volume 72, Number 360
Theory and Methods Section

2. *Computations for a 3 × 3 Table with Row Margins 3,2,1*

t_1 \ t_2	0	1	2	3	4	5	6
0	1	1	1	1	0	0	0
1	1	1	1	0	0	0	0
2	1	1	0	0	0	0	0
3	1	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0

t_1 \ t_2	0	1	2	3	4	5	6
0	1	2	3	3	2	1	0
1	2	4	5	4	2	0	0
2	3	5	5	③	0	0	0
3	3	4	3	0	0	0	0
4	2	2	0	0	0	0	0
5	1	0	0	0	0	0	0
6	0	0	0	0	0	0	0

t_1 \ t_2	0	1	2	3	4	5	6
0	1	3	5	6	5	3	1
1	3	8	12	12	8	3	0
2	5	12	15	12	5	0	0
3	6	12	12	6	0	0	0
4	5	8	5	0	0	0	0
5	3	3	0	0	0	0	0
6	1	0	0	0	0	0	0

array has zeros whenever $t_1 + t_2 > m_1$ and ones elsewhere. The (u, v)th element of the $(i + 1)$st array is obtained by summing over the triangular subset of the ith array with vertices (u, v), $(u - m_{i+1}, v)$, and $(u, v - m_{i+1})$. Table 2 illustrates this method for a 3 × 3 table with row margins 3, 2, and 1. (Elements corresponding to lattice points outside the arrays shown have value zero.) For example, the circled element of the second array is the sum of elements within the triangle shown in the first array. Similarly, the triangular summation mask used to obtain the third array from the second has vertices (u, v), $(u - 1, v)$, and $(u, v - 1)$. The third array contains the exact number of 3 × 3 tables with row margins 3, 2, 1 for all possible fixed column margins. The sum of all elements in the array, 180, is thus the total number of such 3 × 3 tables without restriction on column totals, as in (3.6). The generalization of (2.2) to $r \times c$ tables is straightforward. Successive $c - 1$ dimensional arrays are computed by summing over integer lattice points of the simplex $\sum k_\ell \leq m_{i+1}$, where $\ell = 1, 2, \ldots, (c - 1)$.

3. A NORMAL APPROXIMATION TO THE NUMBER OF $r \times c$ TABLES

First consider the $r \times 2$ case. The recursion (2.1) is reminiscent of a succession of convolutions of random variables. Let $Y_i = X_{i1}$ have independent uniform discrete distributions on $0, 1, \ldots, m_i$. Then the sample space $Y_1, \ldots, Y_r$ has $\prod_{i=1}^r (m_i + 1)$ equally likely points which are in one-to-one correspondence with the set of $r \times 2$ contingency tables having row totals $m_1, \ldots, m_r$. If $T_1 = \sum Y_i$, then $P(T_1 = t) = N_r(t;\, m_1, \ldots, m_r)/\prod_i (m_i + 1)$, as the numerator is the number of favorable cases.

This result leads to a simple normal approximation for $N_r(T_1;\, m_1, \ldots, m_r)$. A sufficient condition for the Liapounov form of the central limit theorem (Hogg and Craig 1970, p. 362) is $\lim_{r\to\infty} (\sum m_i^3)(\sum m_i^2)^{-\frac{3}{2}} = 0$, since the third absolute moment of Y_i is a polynomial in m_i of order three. Under this condition, $(T_1 - \mu)/\sigma$ tends to a standard normal distribution where $\mu = \sum EY_i = \sum m_i/2$ and $\sigma^2 = \sum m_i(m_i + 2)/12$. Thus we have the approximation

$$\begin{aligned} N(t;\, m_1, \ldots, m_r) &= \prod_i (m_i + 1) P[T_1 = t] \\ &\doteq \prod_i (m_i + 1)[\Phi((t - \mu + \tfrac{1}{2})/\sigma) \\ &\qquad - \Phi((t - \mu - \tfrac{1}{2})/\sigma)] \ , \end{aligned} \tag{3.1}$$

where Φ is the standard normal distribution function. For most purposes, an approximation in terms of the normal density,

$$N(t;\, m_1, \ldots, m_r) \doteq [\prod_i^r (m_i + 1)](2\pi\sigma^2)^{-\frac{1}{2}} \exp(-Q/2) \ , \tag{3.2}$$

is sufficiently accurate, where $Q = (t - \mu)^2/\sigma^2$.

To illustrate, consider the 6 × 2 table with row margins 3, 9, 15, 18, 36, and 42. The recursive method outlined in Section 2 required 1.36 seconds of DEC10 system central processor computer time to calculate the entire distribution of $N_6(T_1;\, 42, 36, 18, 15, 9, 3)$ for $T_1 = 0, 1, \ldots, 123$. In particular, exactly $N_6(73;\, 42, 36, 18, 15, 9, 3) = 339{,}314$ tables have $T_1 = 73$. To apply (3.2), we compute $\mu = 123/2$, $\sigma^2 = 328.75$, and $Q = (73 - 123/2)^2/328.75 = 0.40228$. Hence (3.2) yields $43 \times 37 \times 19 \times 16 \times 10 \times 4 \times .01799 = 348{,}118$ tables in good agreement with the exact calculation. The calculation based on (3.1) yields 348,044, which is about the same as that given by (3.2).

Finally, we give a normal approximation for the number of $r \times c$ tables with column totals $t_1, t_2, \ldots, t_c$. This approximation is based on the fact that (3.1) and its generalizations can be interpreted as successive convolutions of independent multivariate discrete distributions. We regard the row vectors $(X_{i1}, X_{i2}, \ldots, X_{i,c-1})$ as independent multivariate vectors. The m_i indistinguishable elements in the ith row may be allocated into c categories in $\binom{m_i+c-1}{c-1}$ ways, each equally likely (Feller 1957, p. 39). For $c = 3$, these outcomes can be represented by the lattice points of the triangular array (X_{i1}, X_{i2}), with $X_{i1} + X_{i2} \leq m_i$. More generally, the outcomes can be represented as the lattice points of the simplex defined by $\sum_j^{c-1} X_{ij} \leq m_i$. The number of such outcomes with

$X_{i1} = x_1$ is $\binom{m_i - x_1 + c - 2}{c-2}$ and the number with $X_{i1} = x_1$ and $X_{i2} = x_2$ is $\binom{m_i - x_1 - x_2 + c - 3}{c-3}$. By symmetry, the probability mass function (pmf) of X_{ij} is

$$p(X_{ij} = x_j) = \binom{m_i - x_j + c - 2}{c - 2}\binom{m_i + c - 1}{c - 1}^{-1} \quad \text{for} \quad x_j = 0, 1, \ldots, m_i \ , \qquad (3.3)$$

and the joint pmf of X_{ij} and X_{ik} is

$$p(X_{ij} = x_j, X_{ik} = x_k) = \binom{m_i - x_j - x_k + c - 3}{c - 3}\binom{m_i + c - 1}{c - 1}^{-1} \quad \text{for} \quad x_j + x_k \leq m_i \ . \qquad (3.4)$$

The moments can be computed from (3.3) or (3.4), but Riordan (1958, pp. 103–104) used elegant enumerator generating function methods to show $E(X_{ij}) = m_i/c$, $\text{var}(X_{ij}) = m_i(m_i + c)(c - 1)/(c + 1)c^2$ and

$$\text{cov}(X_{ij}, X_{ik}) = -m_i(m_i + c)/(c + 1)c^2 \ .$$

The random vector $(T_1, T_2, \ldots, T_{c-1})$ is regarded as the sum of independent row vectors. Thus

$$E(T_j) = \sum m_i/c \ ,$$

$$\sigma^2 \equiv \text{var}(T_j) = \sum_i \text{var}(X_{ij}) = [\sum m_i(m_i + c)](c - 1)/(c + 1)c^2 \ ,$$

and

$$\text{cov}(T_j, T_k) = \sum_i \text{cov}(X_{ij}, X_{ik}) = -\sigma^2/(c - 1) \ .$$

Therefore, the multivariate normal approximation to the distribution of $(T_1, \ldots, T_{c-1})$ is known. Indeed, because $T_1, T_2, \ldots, T_{c-1}$ are equicorrelated and have a common variance, the appropriate multivariate normal density is

$$((c - 1)/2\pi\sigma^2 c)^{(c-1)/2} c^{\frac{1}{2}} \exp\,(-Q/2) \ , \qquad (3.5)$$

where $Q = ((c - 1)/\sigma^2 c)(\sum_{j=1}^{c} t_j^2 - S^2/c)$ with S the grand total of the $r \times c$ table.

To illustrate, we approximate the number of 4×3 tables with row margins 20, 10, 5, 5, and $(t_1, t_2) = (11, 10)$. We note $E(T_1) = E(T_2) = 13.3333$, $\sigma^2 = \text{var}\ T_1 = \text{var}\ T_2 = 37.2222$, and $\text{cov}(T_1, T_2) = -18.6111$. From (3.6), the total number of tables with unrestricted column totals is $\binom{22}{2}\binom{12}{2}\binom{7}{2}\binom{7}{2} = 6{,}723{,}486$. From (3.5), the multivariate normal density evaluated at $(t_1 = 11, t_2 = 10)$ is 0.0031930. Thus the number of tables with row margins 20, 10, 5, 5 and column totals 11, 10, 19 is approximately $6{,}723{,}486 \times 0.0031930 = 21{,}469$ in good agreement with the exact result, 22,245, obtained from (2.2). The approximation performs best when there are many rows and when many of the row margins are large, which is precisely the situation in which exact enumeration becomes costly. For completeness, we note that the approximation for arbitrary c is obtained by multiplying (3.5) times

$$\prod_{i=1}^{r} \binom{m_i + c - 1}{c - 1} \ , \qquad (3.6)$$

the total number of tables with given row margins. This approximation has the great advantage of computational ease, even for the case of large c, which requires an enormous number of calculations for exact solution. As a final practical point, if $c > r$, one should usually relabel rows as columns before using the normal approximation, since convergence to normality is improved by convoluting as many rows as possible.

These approximations afford a ready estimate of the number of $r \times c$ tables with fixed row totals and with an arbitrary subset of column totals fixed. Suppose, for example, that two column totals, which we take without loss of generality to be t_1 and t_2, are fixed. Then the number of $r \times c$ tables with given row totals and with t_1 and t_2 fixed is approximated by the product (3.6) times the marginal density of (T_1, T_2), evaluated at (t_1, t_2). The exact number of such tables can also be worked out from the $(c - 1)$ dimensional array of exact solutions to the original problem with all column totals fixed. For example, with $c = 3$, the number of tables with $m_1 = 3$, $m_2 = 2$, and $m_1 = 1$ and with $t_1 = 2$ is exactly 49, which is obtained by summing over the row $t_1 = 2$ in the third array in Table 2. Generally, the exact number of $r \times c$ tables with fixed row totals and a subset of fixed column totals may be computed by summing appropriate terms in the $(c - 1)$ dimensional array of exact solutions to the original problem with all row and column totals fixed.

We have extended the normal approximation to estimate the number of $r \times c \times s$ tables with given first order margins $x_{i\cdot\cdot}$, $x_{\cdot j\cdot}$, and $x_{\cdot\cdot k}$. Without loss of generality reorient the tables so that $r \leq \max(s, c)$. The three-way table can be viewed as a two-way table with r rows and cs columns. As before, if only the row totals $x_{i\cdot\cdot}$ are fixed, there are $\prod_{i=1}^{r} \binom{x_{i\cdot\cdot} + cs - 1}{cs - 1}$ possible tables, and each of these random "column" totals, $X_{\cdot jk}$, has expectation $x_{\cdot\cdot\cdot}/cs$, variance

$$\sigma^2 = [(cs - 1)/(cs)^2(cs + 1)][\sum x_{i\cdot\cdot}(x_{i\cdot\cdot} + cs)] \ ,$$

and common correlation coefficient $-1/(cs - 1)$. To determine what fraction of these tables with $x_{i\cdot\cdot}$ fixed has margins $x_{\cdot j\cdot}$ and $x_{\cdot\cdot k}$, we regard $X_{\cdot j\cdot}$ as the random sum of s "column" totals in the $r \times (cs)$ table and $X_{\cdot\cdot k}$ as the random sum of c such "column" totals. It can be shown that

$$\sigma_1^2 \equiv \text{var}(X_{\cdot j\cdot}) = \sigma^2 s^2(c - 1)/(cs - 1) \ , \qquad \text{cov}(X_{\cdot j\cdot}, X_{\cdot j'\cdot}) = -\sigma_1^2/(c - 1) \ ,$$

$$\sigma_2^2 \equiv \text{var}(X_{\cdot\cdot k}) = \sigma^2 c^2(s - 1)/(cs - 1) \ , \qquad \text{cov}(X_{\cdot\cdot k}, X_{\cdot\cdot k'}) = -\sigma_2^2/(s - 1) \ ,$$

and $\text{cov}(X_{\cdot j\cdot}, X_{\cdot\cdot k}) = 0$. Thus the approximate number of tables with fixed $x_{i\cdot\cdot}$ and with $X_{\cdot j\cdot} = x_{\cdot j\cdot}$ and $X_{\cdot\cdot k} = x_{\cdot\cdot k}$ is the previous product giving the number of tables with $x_{i\cdot\cdot}$ fixed times

$$((c - 1)/2\pi\sigma_1^2)^{(c-1)/2} c^{\frac{1}{2}} \exp\,(-Q_1/2) \times ((s - 1)/2\pi\sigma_2^2)^{(s-1)/2} s^{\frac{1}{2}} \exp\,(-Q_2/2) \ ,$$

where $Q_1 = ((c - 1)/\sigma_1^2 c)(\sum x_{\cdot j\cdot}^2 - x_{\cdot\cdot\cdot}^2/c)$ and $Q_2 = ((s - 1)/\sigma_2^2 s)(\sum x_{\cdot\cdot}{}^{2}{}_{k} - x_{\cdot\cdot\cdot}^2/s)$. By similar devices we can obtain normal approximations to the number of $r \times c \times s$ tables with fixed second-order margins, $x_{ij\cdot}$, $x_{i\cdot k}$, and $x_{\cdot jk}$.

[*Received November 1976. Revised June 1977.*]

REFERENCES

Abramson, Morton, and Moser, W.O.J. (1973), "Arrays with Fixed Row and Column Sums," *Discrete Mathematics*, 6, 1–14.

Feller, W. (1957), *An Introduction to Probability Theory and Its Applications, Vol. I*, New York: John Wiley & Sons.

Good, Irving J. (1976), "On the Application of Symmetric Dirichlet Distributions and Their Mixtures to Contingency Tables," *The Annals of Statistics*, 4, 1159–1189.

Hogg, Robert V., and Craig, A.T. (1970), *Introduction to Mathematical Statistics*, New York: The MacMillan Co.

Lehmann, Erich L. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco: Holden-Day.

Riordan, J. (1958), *An Introduction to Combinatorial Analysis*, New York: John Wiley & Sons.