The Annals of Statistics
1996, Vol. 24, No. 1, 64–68

ON UNIFORM GENERATION OF TWO-WAY TABLES WITH FIXED MARGINS AND THE CONDITIONAL VOLUME TEST OF DIACONIS AND EFRON

BY R. B. HOLMES AND L. K. JONES[1]

Massachusetts Institute of Technology and University of Massachusetts–Lowell

Two efficient Monte Carlo algorithms are proposed for uniformly generating two-way contingency tables with fixed margins. These permit some improvements on recent work of Diaconis, Efron and Gangolli, especially concerning estimates of the total number of such tables.

1. Introduction. In [1] Diaconis and Efron motivate the use of the uniform distribution on the set of $I \times J$ two-way tables with the same margins [row sums $\mathbf{r} = (r_1, r_2, \ldots, r_I)$, column sums $\mathbf{c} = (c_1, c_2, \ldots, c_J)$] for the alternative hypothesis to that of tables chosen multinomially with independent row and column categories [i.e., with Pr(cell(i, j)) proportional to $r_i c_j$]. Consider, for example, Table 1, showing eye color versus hair color (from [4]).

Now recall that, for a table $\mathbf{p}$ and sample size n, the chi-square statistic is defined as

$$S = \sum_i \sum_j \frac{\left(p_{ij} - r_i c_j / n\right)^2}{r_i c_j / n}.$$

For Table 1 the chi-square statistic $S = 138.29$ with nine degrees of freedom, indicating strong rejection of the hypothesis of independence. The conditional volume test statistic of Diaconis and Efron is (an estimate of)

$$\varepsilon(S|\mathbf{r}, \mathbf{c}) = \frac{\#\{\text{tables with margins } \mathbf{r}, \mathbf{c} \text{ and } \chi^2\text{-statistic} \le S\}}{N^{(n)}(\mathbf{r}, \mathbf{c})}, \tag{1}$$

where S is the χ^2-statistic of the given table, and $N^{(n)}(\mathbf{r}, \mathbf{c})$ is the total number of tables with margins $\mathbf{r}$ and $\mathbf{c}$. (In [1] the tables are normalized by dividing each entry by the total sample size n.) For Table 1 a value of 0.093 was reported in [1] as an estimate of (1). Recently Diaconis and Gangolli [2] gave an estimate of 0.154 for (1), using a Markov chain Monte Carlo algorithm. Our unbiased estimate based on more than 10,000 tables from the uniform distribution is 0.149 (actually, 1537 tables out of 10,317 had $\chi^2 \le 138.29$). Hence we cannot reject the alternative hypothesis that the table was selected uniformly from the set of tables with the same margins.

Received November 1994; revised April 1995.
[1]Partially supported by NSF Grant DMS-92-02161.
AMS 1991 *subject classifications*. 62G99, 62P99.
Key words and phrases. Two-way tables, uniform distribution, Monte Carlo algorithm, rejection algorithm, conditional volume test, Fisher–Yates distribution.

TABLE 1

	Hair color				
Eye color	**Black**	**Brunette**	**Red**	**Blond**	**Total**
Brown	68	119	26	7	220
Blue	20	84	17	94	215
Hazel	15	54	14	10	93
Green	5	29	14	16	64
Total	108	286	71	127	592

The most accurate approach of Diaconis and Efron [1] was first to give an approximate formula for $N^{(n)}(\mathbf{r}, \mathbf{c})$ (obtained from multiplying their (3.13) by (3.14), but without normalization, in [1]), namely,

$$(2)\quad N^{(n)}(\mathbf{r}, \mathbf{c}) \sim \left(\frac{2n + IJ}{2}\right)^{(I-1)(J-1)} \left(\prod_{i=1}^{I} \bar{r}_i\right)^{J-1} \left(\prod_{j=1}^{J} \bar{c}_j\right)^{K-1} \frac{\Gamma(JK)}{\Gamma(J)^I \Gamma(K)^J},$$

where $K = (J + 1)/J\|\bar{\mathbf{r}}\|^2 - 1/J$, $\bar{\mathbf{r}} = (1 - w)/I + w\mathbf{r}/n$, $\bar{\mathbf{c}} = (1 - w)/J + w\mathbf{c}/n$ and $w = 1/(1 + IJ/2n)$. They then estimated the numerator in (1) in an unbiased fashion by sampling tables $\mathbf{p}$ with the same fixed margins but under the assumption that the row and column categories are conditionally independent, and by averaging

$$\left(g_1(\mathbf{p} \mid \mathbf{r}, \mathbf{c})\right)^{-1} I_{\{\chi^2\text{-statistic for } \mathbf{p} \le S\}},$$

where

$$(3)\qquad g_1(\mathbf{p} \mid \mathbf{r}, \mathbf{c}) = \frac{\left(\prod_{i=1}^{I} r_i!\right)\left(\prod_{j=1}^{J} c_j!\right)}{\left(\prod_{j=1}^{J}\prod_{i=1}^{I} p_{ij}!\right)(n!)}$$

is the Fisher–Yates probability function associated with the independence hypothesis. [Here tables are generated by labeling n distinct objects with I colors and J numbers such that each color i (resp., number j) is used r_i (resp., c_j) times; p_{ij} is then the number of objects colored i and numbered j. Equation (3) follows by noting that it equals the number of labelings with p_{ij} of color i and number j divided by the total number of labelings.] Even if the values used for $N^{(n)}(\mathbf{r}, \mathbf{c})$ are accurate, the above Monte Carlo estimate of (1) puts excessive weight on the tables generated from the tails of the Fisher–Yates distribution. Although exact recursive calculations of $N^{(n)}(\mathbf{r}, \mathbf{c})$ are reported in [2], exact calibration of the χ^2-statistic requires running through the entire set of tables and is relatively slow. Methods of aperiodic Markov chains are described in [2] for estimating (1). These have an asymptotically negative exponential bias in terms of the Monte Carlo sample size, but the actual bias for a fixed table and Monte Carlo sample size is unknown. For large n this method may be too slow.

Our method produces a table from precisely the uniform distribution. In fact, our first algorithm can be modified for the case of $n \to \infty$ to produce a matrix of probabilities with fixed given margins from the uniform distribu-

tion. Our second algorithm deals with the situation of small margin proportions. We also show in this case that the estimate by Diaconis and Efron [1] of $N^{(n)}(\mathbf{r}, \mathbf{c})$ may be inaccurate. Finally, we show how to calibrate Fisher's exact test as in [3], but assuming the uniform alternative hypothesis.

2. Rejection algorithms for uniform generation of tables with fixed margins.

2A. *The naive rejection algorithm.* Assume that the row sums r_i are arranged in increasing order. We generate a sequence of $I - 1$ vectors, each with J nonnegative integer components, as follows:

1. Choose the first vector to have component sum equal to r_1, and choose it uniformly from the set of all such vectors. This is easily done by picking $J - 1$ positions without replacement from $r_1 + J - 1$ possible positions, ordering these positions as $p_1 < p_2 < \cdots < p_{J-1}$, and then producing the vector $(p_1 - 1, p_2 - p_1 - 1, \ldots, p_{J-1} - p_{J-2} - 1, r_1 + J - 1 - p_{J-1})$. For example, take $J = 3$ and $r_1 = 5$, so that $r_1 + J - 1 = 7$. If the positions picked happen to be $p_1 = 2$ and $p_2 = 3$, then the first row vector is $(1, 0, 4)$.
2. Check whether, for any j, the jth entry exceeds the column sum c_j. If so, reject the construction and restart at 1. If not, generate a second vector with component sum equal to r_2 exactly as in 1. Again check that, for each j, the sum of the jth entries of the vectors constructed so far do not exceed c_j; if so, reject and restart at 1.
3. Continue in this manner until $I - 1$ vectors have been generated successfully without any rejection. Clearly such a set of $I - 1$ vectors may be uniquely extended to form a table with the desired margins.

Of course, this algorithm may equally well be applied to columns instead of rows.

We now claim that the tables so produced have the desired uniform distribution. This follows by noting that $I - 1$ vectors generated as above with no rejection will be uniformly generated from the set of all such $I - 1$ vectors. Since there is a 1–1 correspondence between those for which no rejection could occur and the tables with the desired margins, steps 1–3 do indeed generate tables uniformly.

We note that "too many" rejections may occur if one or more column (row) sums are "small" compared to the row (column) sum at hand. This was observed in the experiment below. To increase the acceptance rate regardless of the relative size of row (column) sums, we propose a revised algorithm in Section 2B.

To generate probability tables with prescribed margins, we let the positions in steps 1–3 be the $J - 1$ order statistics from the uniform distribution on $(0, r_i)$ and use $(p_1, p_2 - p_1, \ldots, p_{J-1} - p_{J-2}, r_i - p_{J-1})$ as the vector. The proof of uniformity is similar.

2B. *The revised rejection algorithm.* As in Section 2A we generate $I-1$ vectors with the ith vector chosen uniformly from those with sum r_i, but with their jth components bounded by c_j. Clearly the accepted collections of $I-1$ vectors which extend to yield a table will again have the uniform distribution, but with higher acceptance rate, and hence also will the tables so produced.

It remains to demonstrate how to uniformly generate nonnegative integer vectors with given row sums r_i and bound c_j on the jth components. To this end let $N_{\max}$ be the maximum row sum among the first $I-1$ rows. Compute $\phi_q^j \equiv$ (coefficient of x^q) in the polynomial

$$\prod_{k=1}^{j} (1 + x + x^2 + \cdots + x^{c_k}), \tag{4}$$

for $j = 1, 2, \ldots, J-1$ and $q = 0, 1, \ldots, N_{\max}$. Here ϕ_q^j is the number of possible j-dimensional vectors with nonnegative integer components summing to q such that the kth component is less than or equal to c_k. To generate a J-dimensional vector $\mathbf{v}$ with nonnegative integer components whose sum is N, we first define a probability function on the set $\{0, 1, \ldots, c_J \wedge N\}$ whose value at x is proportional to ϕ_{N-x}^{J-1} and select v_J randomly from this distribution, and then successively select v_{J-k} randomly from the distribution on $\{0, 1, \ldots, c_{J-k} \wedge N - (v_J + \cdots + v_{J-k+1})\}$ whose value at x is proportional to

$$\phi_{N-(v_J+\cdots+v_{J-k+1})-x}^{J-k-1}.$$

For $k = J-1$, we choose $v_1 = N - (v_J + \cdots + v_2)$.

To run this algorithm efficiently, the ϕ's are precomputed by recursive integer convolution of the coefficient sequences of the polynomials (4) in their definition and are stored in an array for lookup.

Again we may work with columns instead of rows as in the naive algorithm. Code is available upon request.

3. Applications. Using either the naive or revised algorithm, $N^{(n)}(\mathbf{r}, \mathbf{c})$ may be estimated by first calculating the number of collections of $I-1$ vectors satisfying just the sum constraints and then multiplying by the acceptance rate. For example, in 100,000 iterations of the naive algorithm applied rowwise to Table 1, there were 10,468 successes (no rejections in steps 1–3). Reordering the rows in order of increasing marginal sum, it is easily seen that there are $\binom{67}{3} \cdot \binom{96}{3} \cdot \binom{218}{3}$ collections of three vectors yielding the estimate 1.220×10^{15} for $N^{(n)}(\mathbf{r}, \mathbf{c})$. Formula (2) gives 1.235×10^{15}, while the actual value is 1.226×10^{15} (as the result of an exhaustive calculation reported in [2]). With none of the row proportions terribly small the naive acceptance rate is large enough and formula (2) is quite accurate.

On the other hand consider Table 2, with both a row and a column of small relative proportion. Applying the naive algorithm 100,000 times using rows (resp., columns) yielded only 36 (resp., 18) successful tables, for an acceptance rate of 0.036% (resp., 0.018%). On the other hand, repeating with the revised

TABLE 2

					Total
	—	—	—	—	9
	—	—	—	—	49
	—	—	—	—	182
	—	—	—	—	478
	—	—	—	—	551
Total	9	309	355	596	1269

algorithm yielded 9702 (9.7%) tables and 12,536 (12.5%), after switching rows and columns. Hence in this case the revised algorithm is more than 250 times as efficient as the naive algorithm. Also, $N^{(n)}(\mathbf{r}, \mathbf{c})$ was estimated from the revised algorithm using the number of vector collections calculated from the ϕ_q^j arrays and the acceptance rates. The resulting estimates were 3.346×10^{16} and 3.365×10^{16}, after switching rows and columns, while (2) yielded 1.319×10^{17} using rows and 4.126×10^{16} after switching rows and columns, so that the Diaconis–Efron approximation appears to be in error by at least 20%.

Finally, we remark that our algorithm may be used to calibrate Fisher's exact test for a two-way table under the alternative uniform hypothesis. (The null hypothesis assumes the Fisher–Yates distribution, and the test statistic is the likelihood ratio.) We simply record the fraction of uniformly randomly generated tables $\mathbf{p}$ for which $g_1(\mathbf{p}|\mathbf{r}, \mathbf{c}) \geq g_1(\overline{\mathbf{p}}|\mathbf{r}, \mathbf{c})$ for the given table $\overline{\mathbf{p}}$ or, equivalently, those for which

$$\prod_{i,j} p_{ij}! \leq \prod_{i,j} \overline{p_{ij}}!.$$

Acknowledgment. The authors are grateful to Professor Persi Diaconis for bringing this problem to their attention.

REFERENCES

[1] DIACONIS, P. and EFRON, B. (1985). Testing for independence in a two-way table: new interpretations of the chi-square statistic. *Ann. Statist.* **13** 845–874.

[2] DIACONIS, P. and GANGOLLI, A. (1995). Rectangular arrays with fixed margins. In *Discrete Probability and Algorithms* (D. Aldous, P. Diaconis, J. Spencer and J. M. Steele, eds.) Springer, New York.

[3] MEHTA, C. and PATEL, N. (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *J. Amer. Statist. Assoc.* **78** 427–434.

[4] SNEE, R. (1974). Graphical display of two-way contingency tables. *Amer. Statist.* **38** 9–12.

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE
OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS 02173

DEPARTMENT OF MATHEMATICAL SCIENCES
UNIVERSITY OF MASSACHUSETTS–LOWELL
1 UNIVERSITY AVENUE
LOWELL, MASSACHUSETTS 01854