# LOG-LINEAR MODELS FOR DYAD FORMATION

by

Kinley Larntz
and
Sanford Weisberg*

Technical Report No. 231

August 20, 1974

*Kinley Larntz and Sanford Weisberg are assistant professors, both with the School of Statistics, University of Minnesota, St. Paul, Minnesota 55101. The authors wish to thank R. Dennis Cook for his assistance on Example 2 and R. E. Sykes and J. C. Fox for suggesting the problem and providing the data for Example 1. This work was supported by Grants from the Graduate School, University of Minnesota.

# SUMMARY

We consider the analysis of $k \times k$ upper triangular contingency tables in which both margins have the same polytomy. Such a table will arise when $k$ subjects form pairs (dyads) in which the pair $(i,j)$ cannot be distinguished from $(j,i)$ and individuals cannot form dyads with themselves. We develop appropriate multiplicative models for the data and give two examples.

## 1. INTRODUCTION

In this article we consider the analysis of a k x k upper triangular contingency table with the main diagonal deleted in which the same polytomy occurs on each margin. Such a table can arise, for example, when frequencies of dyads formed by subjects i and j are observed and (i,j) cannot be distinguished from (j,i). Our notation for this problem is shown in Table 1.

For the triangular table, we first fit a "random pairing" model which corresponds to the usual independence hypothesis in an r x c contingency table. Under this model, differences in dyad frequencies can be attributed solely to individual differences. We next consider some multiplicative models [7] in which factors that are common to some dyads but not others may influence the frequency of dyad formations.

Table 1 goes about here

Table 2 goes about here

For example, we shall consider data given in Sykes, Larntz, and Fox [11]. Several sets of k = 6 U. S. Navy recruits were observed during their first few days in training camp. The recruits made up three pairs of bunkmates housed in adjacent two tier bunks. An observer made periodic visits to the bunkhouse and recorded each instance in which two of the recruits were talking to each other. One data set is given here as Table 2. The investigators were interested in assessment of the effects of proximity and race on frequency of dyad formation. They hypothesized that, beyond individual differences in propensity to interact, recruits would tend to interact most often with bunkmates and least often with other recruits bunked relatively far away. They also hypothesized that intra-racial interaction would be more frequent than inter-racial interaction.

In Section 2, we consider the random pairing model. A probability mechanism for generating the table is discussed; we then derive maximum likelihood estimators for the expected dyad counts and give a simple iterative procedure for calculating them. After a brief discussion of goodness-of-fit tests, we give two examples. In Section 3, the random pairing model is generalized to include one additional multiplicative factor. In Section 4, a general model, permitting any number of multiplicative factors is then presented. Section 5 discusses extensions to multi-dimensional tables and some computational considerations.

## 2. A RANDOM PAIRING MODEL

The random pairing situation can be seen to arise from a single urn with a large number of balls such that a proportion $\pi_i$, are marked i, $i = 1, \ldots, k$, $\Sigma\pi_i = 1$. Two balls are drawn at random with replacement with the result given by $(i,j)$, $i \leq j$. This corresponds to formation of a dyad by individuals i and j. If $i < j$, the count in the (i,j) cell, $n_{ij}$ is increased by one. If $i = j$, no changes are recorded. A model similar to this, except allowing for diagonal cells, was given by Mantel and Crittenden [10].

From the urn model, we see that the probability that a randomly chosen dyad is (i,j) is given by

$$\Pr(\ (i,j)\ ) = \begin{cases} 0 & \text{if} \quad i \geq j \\ \dfrac{\pi_i\pi_j}{\sum\sum_{i<j} \pi_i\pi_j} & i < j\ . \end{cases} \tag{2.1}$$

Hence in a data set with $N = \sum\sum_{i<j} n_{ij}$ observed dyads, the expected frequency of the (i,j) dyad, say $m_{ij}$, is given by

$$m_{ij} = N \frac{\pi_i\pi_j}{\sum\sum_{i<j} \pi_i\pi_j} = N\, p_i p_j \tag{2.2}$$

where $p_i = \pi_i / \sqrt{\sum\sum_{i<j} \pi_i\pi_j}$. If we take logarithms of (2.2) we get

$$\begin{aligned} \log m_{ij} &= \log N + \log p_i + \log p_j \\ &= u + u_{1(i)} + u_{1(j)} \end{aligned} \tag{2.3}$$

with each u-term identified with the term directly above it. The subscript "1" occurs on both main effect u-terms in (2.3) because the same polytomy exists on both margins. In equation (2.3), we recognize a log-linear model for the expected dyad counts.

## 2.1 Maximum Likelihood Estimation

The likelihood function for $p_1, \ldots, p_k$ given $n_{12}, n_{13}, \ldots, n_{k-1,k}$ is given by

$$\begin{aligned} \mathrm{Lik}(p_1,\ldots,p_k) &\propto (p_1p_2)^{n_{12}} (p_1p_3)^{n_{13}} \cdots (p_{k-1}p_k)^{n_{k-1,k}} \\ &= \prod_{i=1}^{k} p_i^{(\sum_{j<i} n_{ji} + \sum_{j>i} n_{ij})} \end{aligned} \tag{2.4}$$

Writing $N_i = \sum_{j<i} n_{ji} + \sum_{j>i} n_{ij}$ (i.e., $N_i$ is the total number of observed dyads including subject i, i = 1,2,...,k; note that $\sum_{i=1}^{k} N_i = 2N$), we can write

$$\log \mathrm{Lik}(p_1,\ldots,p_k) \propto \sum_{i=1}^{k} N_i \log p_i. \tag{2.5}$$

Note that the $N_i$ are sufficient statistics for the $p_i$.

We shall maximize (2.5) subject to the constraint $\sum\sum_{i<j} p_i p_j = 1$. Using Lagrange multipliers, we seek to maximize

$$L = \sum_{i=1}^{k} N_i \log p_i + \lambda\left(\sum\sum_{i<j} p_i p_j - 1\right) \tag{2.6}$$

Differentiating,

$$\begin{aligned} \frac{\partial L}{\partial p_i} &= \frac{N_i}{p_i} + \lambda \sum_{j \neq i} p_j = 0 \qquad (i = 1,2,\ldots,k) \\ \frac{\partial L}{\partial \lambda} &= \sum\sum_{i<j} p_i p_j - 1 = 0 \end{aligned} \tag{2.7}$$

Solution of the above equations will give maximum likelihood estimates of $p_1,\ldots p_k$ and hence of the $m_{ij}$.

## 2.2 Interative Proportional Fitting

An alternative and equivalent method of computing the maximum likelihood estimates of the $m_{ij}$, say $\hat{m}_{ij}$, uses iterative proportional fitting. A general method for iterative fitting of log-linear models is given by Goodman ([7], Section 3). (Goodman uses different notation). In the iterative procedure, $\hat{m}_{ij}$ are chosen to satisfy a set of linear constraints. If we let $M_i = \sum_{j>i} \hat{m}_{ij} + \sum_{j<i} \hat{m}_{ji}$ (i.e., $M_i$ is the total number of dyads fit for the i-th subject), the iterative proportional fit for the random pairing model chooses $\hat{m}_{ij}$ subject to the constraints that $M_i = N_i$, so that we constrain the fitted number of dyads for the i-th subject to equal the observed number of dyads, $i = 1,2,\ldots,k$.

Let $\hat{m}_{ij}(v)$ be the estimates of $m_{ij}$ at the v-th step in an iteration. The number of steps required to complete a single cycle of the iterative process, say t, is equal to the number of linear restrictions imposed, which, for the random pairing model is equal to k, the number of subjects. We first define

$$\hat{m}_{ij}(0) = \begin{cases} 1 \text{ if } i < j \\ 0 \text{ if } i \geq j \end{cases} \tag{2.8}$$

and let

$$\hat{M}_i(v) = \sum_{j>i} \hat{m}_{ij}(v) + \sum_{j<i} \hat{m}_{ji}(v). \tag{2.9}$$

Then, for v = 1,2,...,t, we find recursively

$$\hat{m}_{ij}(v) = \begin{cases} \hat{m}_{ij}(v-1)\ (N_i/\hat{M}_i(v-1)) & i=v \text{ or } j=v \\ \hat{m}_{ij}(v-1) & \text{otherwise} \end{cases} \tag{2.10}$$

At the end of the t-th step we set $\hat{m}_{ij}(0) = \hat{m}_{ij}(t)$ and repeat the iterative process until the $\hat{M}_i(t)$ converge to $N_i$; we use the convergence criterion

$$\max_i |M_i(t) - N_i| < \epsilon. \tag{2.11}$$

$\epsilon = .01$ is a good practical value. The maximum likelihood estimates, $\hat{m}_{ij}$, are the $\hat{m}_{ij}(t)$ from the last cycle in the iteration.

Convergence of the iterative fit, and the equivalence of the resulting $\hat{m}_{ij}$ to the maximum likelihood estimates follows from Haberman [8]. Similar iterative fitting schemes are discussed by Fienberg [4], Bishop and Fienberg [1], and Goodman [6]. A complete discussion of this and other aspects of counted data problems can be found in Bishop, Fienberg, and Holland [2].

## 2.3 Test Statistics

The usual chi-square goodness-of-fit statistics can be computed to test the adequacy of the random pairing (or a more complicated) model. These statistics are defined by the following equations:

$$\text{Pearson:} \quad X^2 = \sum_{i<j}\sum (\hat{m}_{ij}-n_{ij})^2/\hat{m}_{ij} \qquad (2.12)$$

$$\text{Likelihood Ratio:} \quad G^2 = 2\sum_{i<j}\sum n_{ij} \log (n_{ij}/\hat{m}_{ij}) \qquad (2.13)$$

$$\text{Freeman-Tukey:} \quad T^2 = \sum_{i<j}\sum ( \sqrt{n_{ij}} + \sqrt{n_{ij}+1} - \sqrt{4\hat{m}_{ij}+1})^2 \qquad (2.14)$$

In each case, we compare the statistic to the central chi-square distribution with degrees of freedom equal to the number of cells (k(k-1)/2) minus the number of linearly independent restrictions imposed in the iterative proportional fit, which, for random pairing, is equal to k. Hence for the random pairing model, we have $k(k-1)/2 - k = k(k-3)/2$ degrees of freedom.

The usefulness of each of the above statistics, and the accuracy of their chi-square approximations depends on the $N_i$ and is discussed in detail in Larntz [9].

## 2.4 Example 1: Recruit Interaction

For the data given in Table 2, the $\hat{m}_{ij}$ and the Freeman-Tukey deviates (defined by $f_{ij} = \sqrt{n_{ij}} + \sqrt{n_{ij}+1} - \sqrt{4\hat{m}_{ij}+1})$ are given in Table 3 for the random pairing model. Nine cycles of the iterative process were required to achieve convergence. Not unexpectedly, the large values of the three goodness-of-fit statistics indicate that random pairing is inadequate to describe this data. Further analysis is presented in Sections 3 and 4.

Table 3 goes about here

## 2.5 Example 2: Genetic Code

Good [5] gives an example due to Dr. R. V. Eck [3]. The data, given here as Table 4, consists of the observed frequencies of amino acid allele pairs in a protein. This data was originally analyzed before the genetic or RNA code table was completely known; if the random pairing model could be fit to the data, evidence would be given in favor of a non-overlapping code (which is now known to be the case); if random pairing failed, evidence against a non-overlapping code (in favor, perhaps,of an overlapping code) would be given. Eck originally presented the data to support the case for an overlapping code.

The data is taken on twenty commonly occurring alleles; in this protein, however, only 17 alleles actually occur. Eliminating the three alleles with observed totals of zero, we are left with k = 17 alleles. After fitting the maximum likelihood estimates, the goodness-of-fit statistics, each with 119 degrees of freedom are given by

$$X^2 = 137.0$$

$$G^2 = 103.4$$

$$T^2 = 66.4.$$

Table 4 goes about here

Although the three statistics are highly discrepant, due to the small counts, we have no evidence for lack of fit of the random pairing model, i.e. no evidence against the non-overlapping genetic code.

Good's procedure for analyzing this data was to test the largest observed value in the table, the 8 in the first row, to see if this was an extreme value. Our procedure allows analysis of the entire table simultaneously without ignoring any information.

## 3. A SINGLE MULTIPLICATIVE FACTOR

Suppose that $S_{q1}$ indexes a subset of dyads such that all $(i,j) \in S_{q1}$ have some characteristic in common that may influence the frequency of dyad formation (for q fixed). Letting $S_{q2} = \{(i,j) | (i,j) \notin S_{q1}\}$, the characteristic common to all dyads in $S_{q1}$ occurs in none of the dyads in $S_{q2}$. For example, in the data in Table 2, subjects 1 and 2 were black, while subjects numbered 3, 4, 5, and 6 were white. Thus, $S_{q1}$ may be taken to be the set of all intra-racial dyads: $S_{11} = \{(1,2), (3,4),(3,5),(3,6),(4,5),(4,6),(5,6)\}$, and $S_{12}$ is the set of all inter-racial dyads: $S_{12} = \{(1,3),(1,4),(1,5),(1,6),(2,3),(2,4),(2,5),(2,6)\}$. We consider for now only one dichotomous factor of this form.

The effect of this factor can be visualized through the mechanism of the urn model. As before, two balls are chosen at random with replacement with result $(i,j)$, $i < j$ (if $i=j$, we return the balls and draw again). For $(i,j) \in S_{qr}$ $r = 1,2$ and q fixed, we flip a coin with probability of heads proportional to some fixed constant, say $\delta_{qr}$, i.e. Prob{Heads for $(i,j) \in S_{qr}$} $= \delta_{qr}/\sum_r \delta_{qr}$ for fixed q. If heads occurs, we increase $n_{ij}$ by one; otherwise no change is noted. Hence, the probability of a randomly chosen and recorded dyad being $(i,j)$ is for fixed q given by

$$\Pr(\,(i,j)\,) = \delta_{qr}\pi_i\pi_j / \sum_{r=1}^{2} \delta_{qr} \sum_{S_{qr}} \pi_i\pi_j \qquad (i,j) \in S_{qr} \quad r = 1,2 \qquad (3.1)$$

Where $\sum_S \pi_i\pi_j$ means summation over $(i,j)$ such that $(i,j) \in S$. In analogy to Section 2, we can now proceed to write down and then solve the likelihood equations. However, we see that this model is of the multiplicative form discussed in [7] and hence maximum likelihood estimates of the $m_{ij}$ can be obtained by iterative proportional fitting.

Now suppose we define $N(S_{qr}) = \sum_{S_{qr}} n_{ij}$ and $M(S_{qr}) = \sum_{S_{qr}} \hat{m}_{ij}$. For this model, the $\hat{m}_{ij}$ will be chosen to satisfy the linear constraints for the random pairing model (i.e. $M_i = N_i$, $i = 1,\ldots,k$), plus the additional constraints $N(S_{qr}) = M(S_{qr})$, $r = 1,2$ for fixed q. Each cycle of the iterative procedure now has $t = k + 2$ steps, the first k identical to those given in Section 2, and the (k+r)-th, for $r = 1,2$, q fixed, given by

$$\hat{m}_{ij}(k+r) = \begin{cases} \hat{m}_{ij}(k+r-1)\dfrac{N(S_{qr})}{\hat{M}(k+r-1)} & \text{if } (i,j) \in S_{qr} \\ \\ \hat{m}_{ij}(k+r-1) & \text{if } (i,j) \notin S_{qr} \end{cases} \tag{3.2}$$

where

$$\hat{M}(k+r) = \sum_{S_{qr}} \hat{m}_{ij}(k+r), \quad r = 1,2. \tag{3.3}$$

The final estimates $\hat{m}_{ij}$ will be given by $\hat{m}_{ij}(t)$ after the procedure has converged.

### 3.1 Goodness-of-fit Statistics

As in Section 2, the usual goodness-of-fit statistics can be computed, with degrees of freedom equal to the number of possible dyads minus the number of linearly independent constraints, which will usually, though not always, be k+1 for fitting a single,dichotomous multiplicative factor. Hence, the statistics will usually have $k(k-3)/2 - 1$ degrees of freedom.

### 3.2 Example 1 (continued)

We now fit the racial effect (e.g., fit the random pairing model plus the constraints $M(S_{1r}) = N(S_{1r})$, $r = 1,2$) described in the first paragraph of this section to the data in Table 2; 12 complete cycles of the iterative procedure were required for convergence. The results ($\hat{m}_{ij}$ and Freeman-Tukey deviates) are given in Table 5. The three goodness-of-fit statistics (on 8 degrees of freedom) are $X^2 = 22.5$; $G^2 = 20.8$, and $T^2 = 19.7$ with significance probabilities of about .005, .01, and .02 respectively indicating probable lack of fit of this model to the data.

Table 5 goes about here

Because we have only two blacks in the data we are analyzing, fitting the racial effect results in $\hat{m}_{12} = n_{12}$; that is, we fit the only intra-black dyad exactly. This is due to the linear constraints put on the table. We can write out the linear constraints corresponding to Subject 1, Subject 2, and the set $S_{12}$ as

$$n_{12} + n_{13} + n_{14} + n_{15} + n_{16} = \hat{m}_{12} + \hat{m}_{13} + \hat{m}_{14} + \hat{m}_{15} + \hat{m}_{16}$$

$$n_{12} + n_{23} + n_{24} + n_{25} + n_{26} = \hat{m}_{12} + \hat{m}_{23} + \hat{m}_{24} + \hat{m}_{25} + \hat{m}_{26} \qquad (3.4)$$

$$n_{13}+n_{14}+n_{15}+n_{16}+n_{23}+n_{24}+n_{25}+n_{26} = \hat{m}_{13}+\hat{m}_{23}+\hat{m}_{24}+\hat{m}_{25}+\hat{m}_{26}+\hat{m}_{23}+\hat{m}_{34}+\hat{m}_{35}+\hat{m}_{36}$$

Subtracting the last equation from the sum of the first two in (3.4) leaves

$$2n_{12} = 2\hat{m}_{12}$$

or

$$n_{12} = \hat{m}_{12}.$$

When a fairly large number of linearly independent constraints are imposed on only a few cells, as is the case here, some cells will be fit exactly due to the constraints, and the effects of that cell cannot be adequately analyzed under those constraints. Whenever possible, experiments should be designed to avoid this problem; here, a third black subject would eliminate constrained exact fits.

A second factor of interest in this data is the effect of proximity of the subjects in the bunkhouse. As explained previously, the subjects occupied three adjacent two tier bunks. Hence, we can define $S_{21}$ to be the set of dyads formed by bunkmates or near neighbors: $S_{21} = \{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4),(3,5),(3,6),(4,5),(4,6),(5,6)\}$, while $S_{22}$ is the set of non-adjacent dyads: $S_{22} = \{(1,5),(1,6),(2,5),(2,6)\}$. If we fit this factor (which we will call "one degree of freedom proximity"), we get the expected values, Freeman-Tukey deviates, and goodness-of-fit statistics given in Table 6. Clearly, the one degree of freedom proximity factor provides an inadequate model.

Table 6 goes about here

## 4. SEVERAL MULTIPLICATIVE FACTORS

The generalization of the results of the last section to several multiplicative factors is straightforward. Suppose we have a doubly indexed sequence of subsets of the possible dyads $S_{qr}$, $q = 1,2,\ldots,Q$; $r = 1,2,\ldots,R_q$ such that, for *fixed* q, the $S_{qr}$ are a disjoint collection of subsets whose union is the set of all possible dyads. We now turn to the urn model in which two balls are chosen at random with replacement with result $(i,j)$ $i < j$. (If $i = j$, we draw again). We then flip Q coins each with probability proportional to $\delta_{qr}$ of heads, for $(i,j) \epsilon S_{qr}$, $r = 1,\ldots,R_q$. If *all* coins are heads, we increase $n_{ij}$ by one; otherwise no change is recorded. Thus, the probability of drawing and observing the $(i,j)$ dyad is proportional to $\pi_i\pi_j$ times the appropriate product of the $\delta_{qr}$.

The iterative procedure is as in Sections 2 and 3 except we now have $t = k + \sum_{q=1}^{Q} R_q$ linear constraints, the first k given by $M_i = N_i$, $i = 1,2,\ldots,k$, and the remainder given by $N(S_{qr}) = M(S_{qr})$, with $N(S_{qr}) = \sum_{S_{qr}} n_{ij}$ and $M(S_{qr}) = \sum_{S_{qr}} \hat{m}_{ij}$. For the iteration, the first k steps in each cycle are as given in Section 2, and the last $t\text{-}k = \sum_{q=1}^{Q} R_q$ are given recursively by

$$m_{ij}(k+\sum_{\ell=1}^{q-1} R_\ell + r) = \begin{cases} \hat{m}_{ij}(k+\sum_{\ell=1}^{q-1} R_\ell + r-1)\ \dfrac{N(S_{qr})}{\hat{M}(S_{qr})} & (i,j) \in S_{qr} \\ \\ \hat{m}_{ij}(k+\sum_{\ell=1}^{q-1} R_\ell + r-1) & (i,j) \notin S_{qr} \end{cases} \qquad (4.1)$$

where $\hat{M}(S_{qr}) = \sum_{S_{qr}} \hat{m}_{ij}(k+\sum_{\ell=1}^{q-1} R_\ell + r-1)$, and $\sum_{\ell=1}^{o} R_\ell \equiv 0$. The maximum likelihood estimates $\hat{m}_{ij}$ will be given by the $\hat{m}_{ij}(t)$ after the procedure has converged.

## 4.1 Example 1 (conclusion)

Fitting the racial and one degree of freedom factors described by $S_{11}$, $S_{12}$, $S_{21}$, and $S_{22}$ in Section 3 gives the values given in Table 7. All three goodness-of-fit statistics are beyond the .01 level suggesting that this model is also inadequate. If, however, we examine the Freeman-Tukey deviates, we see that two of the four largest deviates (in cells (3,4) and (5,6)) occur in cells that represent dyads between bunkmates (the third bunkmate dyad, (1,2), is fit exactly because of the racial effect constraint as previously noted). This suggests that the proximity effect should be divided into three sets: bunkmates ($S_{31} = \{(1,2),(3,4),(5,6)\}$), near neighbors ($S_{32} = \{(1,3),(1,4),(2,3),(2,4),(3,5),(3,6),(4,5),(4,6)\}$) and far dyads ($S_{33} = \{(1,5),(1,6),(2,5),(2,6)\}$). We will call this set of restrictions "two degrees of freedom proximity." The fit of two degree of freedom proximity alone (e.g., $S_{31}$, $S_{32}$, $S_{33}$) is given in Table 8; the fit of two degree of freedom proximity and the racial effect is given in Table 9.

Tables 7, 8, and 9 go about here

Note first in Table 8 that the (3,4) dyad is fit exactly due, again, to the constraints on the $\hat{m}_{ij}$. The values of the goodness-of-fit statistics are all near the seventy-fifth percentile of the chi-square distribution with 7 degrees of freedom, suggesting that this model is adequate for the data.

In Table 9, in which both race and two degree of freedom proximity are fit, all three bunkmate dyads are constrained to have fitted values equal to observed values. The goodness-of-fit statistics are all near the fifteenth percentile of the chi-square distribution with 6 degrees of freedom suggesting, perhaps, that inclusion of the racial effect "overfits" the data. However, the difference between the $G^2$ statistics in Table 8 and 9, 9.3 - 2.6 = 6.7 provides a one degree of freedom likelihood-ratio test for the effect of race after fitting two degrees of freedom for proximity. This test yields significance probability less than 0.01, suggesting the importance of racial effects. Differences between other $G^2$ statistics provide tests of other factors; for example, the second degree of freedom for proximity after the first, or proximity after race.

It is possible that different sets of constraints could lead to fitting the same expected values (e.g., several different sets of constraints can lead to the same set of linearly independent constraints). For example, suppose that we decide to delete the (5,6) dyad in Table 7, perhaps because of its large Freeman-Tukey deviate in Table 7. This can be accomplished by letting $S_{41} = \{(5,6)\}$ and $S_{42} = \{\text{all other cells}\}$.

Fitting the racial effect ($S_{11}$, $S_{12}$) and the (5,6) dyad ($S_{41}$, $S_{42}$) will lead to a new set of expected values, but fitting the racial effect, one degree of freedom proximity and the (5,6) dyad leads to the same linearly independent constraints as fitting racial effects and two degrees of freedom for proximity, and hence gives the same expected values as in Table 9.

Similarly, we could not divide the racial effect into three subsets for this data because the three linear constraints for fitting black-black, white-white and black-white would add only one linearly independent constraint after the six constraints for random pairing.

## 5. COMMENTS

While this article has been concerned with models for dyad formation, there are obvious extensions for other group sizes. Also, interactions between the various multiplicative factors can easily be incorporated into a general log-linear model. In this case the iterative proportional fit would be taken over two- or higher-way margins.

In the iteration schemes we have used in this paper, we have allowed one step in the iteration for each linear constraint imposed. Of course, we need only have as many steps as we have linearly independent constraints. However, there is no harm in the extra steps. Also, we have found that the extra steps can decrease the total number of cycles sufficiently to more than make up for the extra few steps per cycle.

## REFERENCES

[1] Bishop, Y. M. M. and Fienberg, S. E., "Incomplete Two-dimensional Contingency Tables," Biometrics 25 (March, 1969), 118-128.

[2] Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W., Discrete Multivariate Analysis: Theory and Practice, Cambridge: M.I.T. Press, 1974.

[3] Eck, R. V., "Non-randomness in Amino-acid 'Alleles'," Nature 19 (September 23, 1961), 1284-1285.

[4] Fienberg, S. E., "An Iterative Procedure for Estimation in Contingency Tables," Annals of Mathematical Statistics 41 (June, 1970) 907-917.

[5] Good, I. J., The Estimation of Probabilities, Cambridge: M.I.T. Press, 1965.

[6] Goodman, L. A., "The Analysis of Cross-classified Data: Independence, Quasi-independence and Interactions in Contingency Tables with or without Missing Entires," The Journal of the American Statistical Association 63 (December, 1968), 1091-1131.

[7] ____, "Some Multiplicative Models for the Analysis of Cross-classified Data," Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I (1972), 649-696.

[8] Haberman, S. J., The Analysis of Frequency Data, Chicago: University of Chicago Press, 1974.

[9] Larntz, K., "Small Sample Comparison of Likelihood-ratio and Pearson Chi-square Statistics for the Null Distribution," (in preparation).

[10] Mantel, N. and Crittenden, L. B., "Determination of Chi-square for Comparing the Number of Matched Pairs with its Expectation," Journal of the National Cancer Institute 27 (1967), 887-892.

[11] Sykes, R. E., Larntz, K., and Fox, J. C., "Proximity and Similarity Effects on Frequency of Interaction in a Class of Naval Recruits," (in preparation).

Table 1 Notation for observed dyad formations

| Subject no. | 1 | 2 | 3 | . . . . | k | Subject total |
|---|---|---|---|---|---|---|
| 1 | | $n_{12}$ | $n_{13}$ | . . . . | $n_{1k}$ | $N_1$ |
| 2 | | | $n_{23}$ | . . . . | $n_{2k}$ | $N_2$ |
| . | | | | . | . | . |
| . | | | | . | . | . |
| . | | | | . | . | . |
| k-1 | | | | . | $n_{k-1,k}$ | $N_{k-1}$ |
| | | | | | | $N_k$ |

Note that $N_i = \sum_{j<i} n_{ji} + \sum_{i>j} n_{ij}$ is the total number of dyads involving subject i, and that $\sum_i N_i = 2N$, where N is the sum of all counts in the table.

Table 2 Data from Sykes, Larntz, and Fox [11]

| Subject no. | 1 | 2 | 3 | 4 | 5 | 6 | $N_i$ |
|---|---|---|---|---|---|---|---|
| 1 | | 41 | 10 | 5 | 6 | 3 | 65 |
| 2 | | | 9 | 6 | 6 | 3 | 65 |
| 3 | | | | 42 | 13 | 5 | 79 |
| 4 | | | | | 15 | 7 | 75 |
| 5 | | | | | | 14 | 54 |
| 6 | | | | | | | 32 |

N = 185

Table 3 Expected values (upper entry) and Freeman-Tukey deviates (lower entry) for the data in Table 2, fitting random pairing model.

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | | 13.7<br>5.4 | 17.8<br>-2.0 | 16.5<br>-3.5 | 10.9<br>-1.6 | 6.0<br>-1.3 |
| 2 | | | 17.8<br>-2.3 | 16.5<br>-3.1 | 10.9<br>-1.6 | 6.0<br>-1.3 |
| 3 | | | | 21.4<br>3.7 | 14.2<br>-.25 | 7.8<br>-1.0 |
| 4 | | | | | 13.2<br>.54 | 7.3<br>-.02 |
| 5 | | | | | | 4.8<br>3.1 |
| 6 | | | | | | |

$X^2 = 122.9$

$G^2 = 102.1$ d.f. = 9

$T^2 = 94.0$

Table 4 Data for Example 2

| | | | | | | | | | | | | | | | | | | | $N_i$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 3 | 0 | 5 | 1 | 3 | 1 | 0 | 3 | 0 | 0 | 0 | 1 | 8 | 1 | 0 | 1 | 0 | 28 |
| | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 0 | 8 |
| | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | | 0 | 4 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | | | | 2 | 2 | 0 | 0 | 1 | 3 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 20 |
| | | | | | | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| | | | | | | | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 10 |
| | | | | | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| | | | | | | | | | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 3 | 0 | 5 |
| | | | | | | | | | | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 9 |
| | | | | | | | | | | | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 10 |
| | | | | | | | | | | | | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 |
| | | | | | | | | | | | | | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | | | | | | | | | | | | | 1 | 0 | 0 | 0 | 0 | 2 |
| | | | | | | | | | | | | | | | 1 | 0 | 1 | 0 | 20 |
| | | | | | | | | | | | | | | | | 0 | 1 | 0 | 4 |
| | | | | | | | | | | | | | | | | | 0 | 0 | 0 |
| | | | | | | | | | | | | | | | | | | 1 | 12 |
| | | | | | | | | | | | | | | | | | | | 1 |

N = 76

Table 5 Expected values (upper entry) and Freeman-Tukey deviates (lower entry) for data in Table 2 with racial effect fit.

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | | 41<br>.039 | 8.6<br>.54 | 7.8<br>-1.00 | 4.9<br>.54 | 2.7<br>.32 |
| 2 | | | 8.6<br>.23 | 7.8<br>-.59 | 4.9<br>.54 | 2.7<br>.32 |
| 3 | | | | 31.4<br>1.79 | 19.8<br>-1.61 | 10.7<br>-1.93 |
| 4 | | | | | 18.1<br>-.70 | 9.8<br>-.864 |
| 5 | | | | | | 6.2<br>2.54 |
| 6 | | | | | | |

$X^2 = 22.5$

$G^2 = 20.8$ d.f. = 8

$T^2 = 19.7$

Table 6 Expected values (upper entry) and Freeman-Tukey deviates (lower entry) for data in Table 2 fitting one degree of freedom for proximity.

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | | 20.1<br>3.9 | 18.5<br>-2.2 | 17.4<br>-3.7 | 5.9<br>.12 | 3.1<br>.10 |
| 2 | | | 18.5<br>-2.5 | 17.4<br>-3.3 | 5.9<br>.12 | 3.1<br>.10 |
| 3 | | | | 16.0<br>5.0 | 17.2<br>-1.0 | 8.8<br>-1.3 |
| 4 | | | | | 16.1<br>-.21 | 8.2<br>-.35 |
| 5 | | | | | | 8.9<br>1.6 |
| 6 | | | | | | |

$X^2 = 95.0$

$G^2 = 84.6$ d.f. = 8

$T^2 = 80.8$

Table 7 Expected values (upper entry) and Freeman-Tukey deviates (lower entry) for fitting racial effect and one degree of freedom proximity.

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | | 41.0<br>.038 | 7.8<br>.79 | 7.2<br>-.76 | 5.8<br>.17 | 3.2<br>.028 |
| 2 | | | 7.8<br>.47 | 7.2<br>-.35 | 5.8<br>.17 | 3.2<br>.028 |
| 3 | | | | 33.5<br>1.4 | 19.3<br>-1.5 | 10.5<br>-1.9 |
| 4 | | | | | 17.6<br>-.57 | 9.6<br>-.80 |
| 5 | | | | | | 5.5<br>2.8 |
| 6 | | | | | | |

$X^2 = 22.8$

$G^2 = 19.8$ d.f. = 7

$T^2 = 18.2$

Table 8 Expected values (upper entry) and Freeman-Tukey deviates (lower entry) for two degree of freedom proximity.

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | | 36.5<br>.76 | 10.3<br>-.017 | 9.2<br>-1.5 | 6.5<br>-.11 | 2.5<br>.43 |
| 2 | | | 10.3<br>-.33 | 9.2<br>-1.0 | 6.5<br>-.11 | 2.5<br>.43 |
| 3 | | | | 42.0<br>.038 | 11.9<br>.38 | 4.5<br>.32 |
| 4 | | | | | 10.6<br>1.3 | 4.0<br>1.3 |
| 5 | | | | | | 18.5<br>-1.0 |
| 6 | | | | | | |

$X^2 = 9.3$

$G^2 = 9.3$ d.f. = 7

$T^2 = 9.1$

Table 9 Expected values (upper entry) and Freeman-Tukey deviates (lower entry) for two degree of freedom proximity and racial effects.

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | | 41.0<br>.038 | 7.9<br>.76 | 7.1<br>-.73 | 6.2<br>.013 | 2.8<br>.24 |
| 2 | | | 7.9<br>.44 | 7.1<br>-.32 | 6.2<br>.013 | 2.8<br>.24 |
| 3 | | | | 42.0<br>.038 | 14.6<br>-.36 | 6.6<br>-.53 |
| 4 | | | | | 13.0<br>.59 | 5.9<br>.53 |
| 5 | | | | | | 14.0<br>.064 |
| 6 | | | | | | |

$X^2 = 2.6$

$G^2 = 2.6$ d.f. = 6

$T^2 = 2.6$