Contingency Tables and Log-Linear Models: Basic Results and New Developments

Stephen E. FIENBERG

1. HISTORICAL REMARKS ON CONTINGENCY TABLE ANALYSIS

Contingency table analysis is rooted in the turn-of-the-century work of Karl Pearson and George Udny Yule, who introduced the cross-product, or odds ratio, as a formal statistical tool. The subsequent contributions by R. A. Fisher linked their methods to basic statistical methodology and theory, but it was not until 1935 that Maurice Bartlett, as a result of a suggestion by Fisher, utilized Yule's cross-product ratio to define the notion of second-order interaction in a $2 \times 2 \times 2$ table and to develop an appropriate test for the absence of such an interaction (Bartlett 1935). The multivariate generalizations of Bartlett's work, beginning with a 1956 article by Roy and Kastenbaum, form the basis of the log-linear model approach to contingency tables, which is largely the focus of this vignette. Key articles in the 1960s by M. W. Birch (1963), Yvonne Bishop (1975), John Darroch (1962), I. J. Good (1963), Leo Goodman (1963), and Robin Plackett (1974), plus the availability of high-speed computers, led to an integrated theory and methodology for the analysis of contingency tables based on log-linear models, culminating in a series of books published in the 1970s. (Historical references can be found in various sources including Bishop, Fienberg, and Holland 1975, Carriquiry and Fienberg 1998, Fienberg 1980, and Haberman 1974.)

The next section outlines some of the basic results on likelihood estimation for log-linear models used to describe interactions in contingency tables, as the theory emerged by the early 1970s. I then briefly describe some of the major advances of the next three decades related to log-linear models. There is now an extensive literature on other classes of models and other methods of estimation, especially Bayesian, but I treat these only in passing, not because they are unimportant, but rather because they draw on similar foundations. Finally, I outline some important open research problems.

Many statisticians view the theory and methods of log-linear models for contingency tables as a special case of either exponential family theory or generalized linear models (GLMs) (Christensen 1996; McCullagh and Nelder 1989). It is true that computer programs for GLM often provide

Stephen E. Fienberg is Maurice Falk University Professor of Statistics and Social Science, Department of Statistics and Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA 15213 (E-mail: *fienberg@stat.cmu.edu*). This work was supported by National Science Foundation grant no. REC-9720374.

convenient and relatively efficient ways of implementing basic estimation and goodness-of-fit assessment. But adopting such a GLM approach leads the researcher to ignore the special features of log-linear models relating to interpretation in terms of cross-product ratios and their generalizations, crucial aspects of estimability and existence associated with patterns of zero cells, and the many innovative representations that flow naturally from the basic results linking sampling schemes. One very important development that I do not cover (due mainly to a lack of space) is the role of general estimating equations and marginal models for longitudinal data within the GLM framework (see, e.g., Diggle, Liang, and Zeger 1996).

2. SAMPLING MODELS AND BASIC LOG-LINEAR MODEL THEORY

Let $\mathbf{x}' = (x_1, x_2, \ldots, x_t)$ be a vector of observed counts for t cells, structured in the form of a cross-classification. Now let $\mathbf{m}' = (m_1, m_2, \ldots, m_t)$ be the vector of expected values that are assumed to be functions of unknown parameters $\boldsymbol{\theta}' = (\theta_1, \theta_2, \ldots, \theta_s)$, where $s < t$.

There are three standard sampling models for the observed counts in contingency tables. In the *Poisson model*, the $\{x_i\}$ are observations from independent Poisson random variables with means $\{m_i\}$, whereas in the *multinomial model*, the total count $N = \sum_{i=1}^{t} x_i$ is a random sample from an infinite population where the underlying cell probabilities are $\{m_i/N\}$. Finally, in the *product-multinomial model*, the cells are partitioned into sets, and each set has an independent multinomial structure, as in the multinomial model.

The following results hold under the Poisson and multinomial sampling schemes:

1. Corresponding to each parameter in $\boldsymbol{\theta}$ is a minimal sufficient statistic (MSS) that is expressible as a linear combination of the $\{x_i\}$. More formally, if $\mathcal{M}$ is used to denote the log-linear model specified by $\mathbf{m} = \mathbf{m}(\boldsymbol{\theta})$, then the MSSs are given by the projection of $\mathbf{x}$ onto $\mathcal{M}$, $P_{\mathcal{M}}\mathbf{x}$.

2. The maximum likelihood estimator (MLE), $\hat{\mathbf{m}}$, of $\mathbf{m}$, if it exists, is unique and satisfies the likelihood equations

$$P_{\mathcal{M}}\hat{\mathbf{m}} = P_{\mathcal{M}}\mathbf{x}. \tag{1}$$

Necessary and sufficient conditions for the existence of a solution to the likelihood equations, (1), are relatively complex (see, e.g., Haberman 1974). A sufficient condition is that all cell counts be positive (i.e., $\mathbf{x} > \mathbf{0}$), but MLEs for log-linear models exist in many sparse situations where a large fraction of the cells have zero counts.

For product-multinomial sampling situations, the basic multinomial constraints (i.e., that the counts must add up to the multinomial sample sizes) must be taken into account. Typically, some of the parameters in $\boldsymbol{\theta}$ that specify the log-linear model $\mathcal{M}$ [i.e., $\mathbf{m} = \mathbf{m}(\boldsymbol{\theta})$], are fixed by these constraints.

More formally, let $\mathcal{M}$ be a log-linear model for $\mathbf{m}$ under product-multinomial sampling that corresponds to a log-linear model $\mathcal{M}$ under Poisson sampling such that the multinomial constraints "fix" a subset of the parameters, $\boldsymbol{\theta}$, used to specify $\mathcal{M}$. Then the following result holds:

3. The MLE of $\mathbf{m}$ under product-multinomial sampling for the model $\mathcal{M}$ is the same as the MLE of $\mathbf{m}$ under Poisson sampling for the model $\mathcal{M}$.

The final basic result relates to assessing the fit of log-linear models:

4. Let ϕ be a real-valued parameter in the interval $-\infty < \phi < \infty$. If $\hat{\mathbf{m}}$ is the MLE of $\mathbf{m}$ under a log-linear model, and if the model is correct, then for each value of ϕ, the goodness-of-fit statistic,

$$K(\mathbf{x}, \hat{\mathbf{m}}) = \frac{2}{\phi(\phi+1)} \sum_{i=1}^{t} x_i \left[\left(\frac{x_i}{\hat{m}_i} \right)^{\phi} - 1 \right], \tag{2}$$

has an asymptotic chi-squared distribution with $t-s$ degrees of freedom as sample sizes tend to infinity, where s is the total number of independent constraints implied by the log-linear model and the multinomial sampling constraints (if any). If the model is not correct, then the distribution is stochastically larger than χ^2_{t-s}.

The usual Pearson chi-squared and likelihood-ratio chi-squared statistics are special cases of the family of *power-divergence statistics* defined by $K(\mathbf{x}, \hat{\mathbf{m}})$ in (2). The Pearson statistic chi-squared statistics corresponds to $\phi = 1$, and the statistic G^2 corresponds to the limit as $\phi \to 0$. (For further details on the properties of the general family of power divergence statistics, see Read and Cressie 1988.)

In the late 1970s, several authors attempted to address the problem of large sparse asymptotics, for example; for a sequence of multinomially structured tables in increasing size, where the sample size n and the number of cells t or the number of parameters s go to infinity in some fixed ratio. Results of Haberman (1977) and Koehler and Larntz (1980) provide some guidance to statistical practice and suggest that the usual advice that expected cell counts should be ≥ 5 is far too conservative and wasteful of information in large sparse tables.

Bartlett's (1935) no–second-order interaction model for the expected values in a $2 \times 2 \times 2$ table, with entries m_{ijk}, is based on equating the values of the cross-product ratio, α, in each layer of the table; that is,

$$\frac{m_{111}m_{221}}{m_{121}m_{211}} = \frac{m_{112}m_{222}}{m_{122}m_{212}}. \tag{3}$$

Expression (3) can be represented in log-linear form as

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}, \tag{4}$$

with suitable linear side constraints on the sets of u terms to achieve identifiability.

All of the parameters in (4) can be written as functions of cross-product ratios (see Bishop et al. 1975). Applying the basic results for the basic sampling schemes applied to the $2 \times 2 \times 2$ table, we have that the MSSs are the two-dimensional marginal totals, $\{x_{ij+}\}$, $\{x_{i+k}\}$, and

$\{x_{+jk}\}$ (except for linearly redundant statistics included for purposes of symmetry), where a "+" indicates summation over the corresponding subscript. Further, the MLEs of the $\{m_{ijk}\}$ under model (4) must satisfy the likelihood equations

$$\hat{m}_{ij+} = x_{ij+}, \qquad i,j = 1,2,$$

$$\hat{m}_{i+k} = x_{i+k}, \qquad i,k = 1,2,$$

and

$$\hat{m}_{+jk} = x_{+jk}, \qquad j,k = 1,2, \tag{5}$$

usually solved by some form of iterative procedure. For the example actually considered by Bartlett, the third set of equations in (5) corresponds to the binomial sampling constraints.

The class of log-linear models just described for the three-way table generalizes in a direct fashion to $k \geq 4$ dimensions. As long as the models retain a hierarchical structure (e.g., setting $u_{12(ij)} = 0$ for all i,j implies setting $u_{123(ijk)} = 0$ for all i,j,k), the MSSs are sets of marginal totals of the full table. Further, all independence or conditional independence relationships are representable as log-linear models, and these models have estimated expected values that can be computed directly. A somewhat larger class of log-linear models with this direct, or *decomposable*, representation is described later. All log-linear models that are not decomposable require an iterative solution of likelihood equations.

In a multiway contingency table, the model that results from setting exactly one two-factor term (and all of its higher-order relatives) equal to 0 is called a *partial association* model. For example, in four dimensions, if $u_{12(ij)} = 0$ for all i,j, then the MSSs are $\{x_{i+kl}\}$ and $\{x_{+jkl}\}$, and the resulting partial association model corresponds to the conditional independence of variables 1 and 2 given 3 and 4. The corresponding MLEs for the expected cell frequencies are

$$\hat{m}_{ijkl} = \frac{x_{i+kl}x_{+jkl}}{x_{++kl}} \quad \forall\, i,j,k,l. \tag{6}$$

Bishop et al. (1975) and Whitaker (1990) provided more details on partial association models and their uses.

3. MAJOR SUBSEQUENT DEVELOPMENTS

3.1 The Graphical Subfamily of Log-Linear Models

A major innovation in log-linear model methods over the past 20 years has been the development of methods associated with a subfamily of log-linear models known as *graphical log-linear models*. Darroch, Lauritzen, and Speed (1980) first described these models, and the monographs by Lauritzen (1996) and Whitaker (1990) made accessible most of the subsequent results in the literature.

This approach uses the vertices of a graph to represent variables and the edges among them to represent relationships. Conditional independence relationships correspond to the absence of edges in such an undirected graph. Models defined solely in terms of such relationships are said to be *graphical*. For categorical random variables, all graphical models are log-linear. The subfamily of graphical log-linear models includes the class of decomposable models, but not all nondecomposable models are graphical. Various authors have used graphical log-linear models to simplify approaches to model search, and they are intimately related to an extensive literature on *collapsibility* and estimability of parameters via marginal tables.

3.2 p_1 Models for Social Networks

Holland and Leinhardt (1981) introduced a log-linear model for representing relationships among individuals in a social network. Their model has a graphical representation, but one that is different from that of the previous section, in that it links individuals instead of variables. Fienberg, Meyer, and Wasserman (1985) showed how to explicitly handle social network data and the Holland–Leinhart model and its extensions in contingency table form using basic log-linear model tools. Wasserman and Pattison (1969) provided related logistic representations.

3.3 Latent Trait and Rasch Models

In psychological tests or attitude studies, one often is interested in quantifying the value of an unobservable *latent trait*, such as mathematical ability or manual dexterity, on a sample of individuals. Although latent traits are not directly measurable, one can attempt to assess indirectly a person's value for the latent trait from his or her responses to a set of well-chosen items on a test. The simplest model for doing so was introduced by Rasch (1960). Given responses for n individuals on k binary random variables, let $\mathbf{X}$ denote the $n \times k$ matrix of responses for n individuals on k binary variables, and let $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ denote the vectors of item and individual parameters. Then the simple dichotomous Rasch model states that

$$\log[P\,(X_{ij} = 1|\theta_i,\alpha_j)/P\,(X_{ij} = 0|\theta_i,\alpha_j)] = \theta_i - \alpha_j. \tag{7}$$

This is a logit model for the log odds for $X_{ij} = 1$ versus $X_{ij} = 0$. We can recast the observed data x_{ij} in the form of a $n \times 2^k$ array, with exactly one observation for each level of the first variable.

In the 1980s, Duncan (1983) and Tjur (1982) recognized an important relationship between the Rasch model and log-linear models for the corresponding collapsed 2^k contingency table. Darroch (1986) and Fienberg and Meyer (1983) represented these models in terms of the log-linear models of quasi-symmetry, but ignored the moment constraints described by Cressie and Holland (1983). More recently, Agresti ('1993a, 1993b) and others have carried these ideas further for other categorical data problems.

3.4 Multiple-Recapture Models for Population Estimation

If the members of a population are sampled k different times, the resulting recapture history data can be displayed in the form of a 2^k table with one missing cell, corresponding to those never sampled. Such an array is amenable to

log-linear model analysis, the results of which can be used to project a value for the missing cell (as in Fienberg 1972). Major applications of capture–recapture methodology include estimating the undercount in the U.S. decennial census, where $k = 2$ (see, e.g., the articles in the special 1993 section of JASA), and the prevalence of various epidemiological conditions, where typically $k \geq 3$.

The use of standard log-linear models in this context presumes that capture probabilities are constant across the population. Agresti (1994) and Darroch, Fienberg, Glonek, and Junker (1993) used a variation of the Rasch model to allow for special multiplicative forms of heterogeneity. Fienberg, Johnson, and Junker (1999) integrated this form of heterogeneity into the log-linear framework and explicitly incorporated the moment constraints in a Bayesian implementation.

3.5 Association Models for Ordinal Variables

Log-linear models as described in this article ignore any structure linking the categories of variables, yet biostatistical problems often involve variables with ordered categories; for example, differing dosage levels for a drug or the severity of symptoms or side effects. Goodman (1979) provided a framework for extending standard log-linear models via multiplicative interaction terms of the form

$$u_{12(ij)} = u^*_{1(i)} u^*_{2(j)} \tag{8}$$

to represent a two-factor u-term. This extended class of models, known as *association models*, have close parallels with correspondence analysis models and both classes have been developed and extended by Clogg, Gilula, Goodman, and Haberman, among others. (For a detailed description of these and other methods for ordinal variables, see Agresti 1990 and Clogg and Shidadeh 1994.)

3.6 Gröbner Bases and Exact Distributions

Haberman (1974) actually gave the conditional distribution for a table under a log-linear model given the marginals which are the MSSs under the model. But actually calculating that conditional distribution is quite complex and most attempts to work with it have focused solely on the calculation of specific quantiles such as p values (see, e.g., Agresti 1992). Diaconis and Sturmfels (1998) provided an elegant solution to the computational problem of computing such "exact" distributions for multiway contingency tables, using the group theory structure of Gröbner bases and a Markov chain Monte Carlo algorithm. Applications of this technology for disclosure limitation can be found in the recent 1998 special issue of the *Journal of Official Statistics*, but realistic implementation for high-dimensional tables is still an open issue.

4. SOME CHALLENGING OPEN PROBLEMS

Although the basic theory of log-linear models and methods for the analysis of contingency tables was in place over 20 years ago, and there have been major advances in various related topics over the ensuing years, some problems have eluded satisfactory solution. First and foremost among these are diagnostics for model fit and graphical representations for model search. Typical GLM diagnostics are geared largely to the noncategorical data response situation and most of the other methods suggested to date are ad hoc at best. Similarly, although graphical model tools have helped to simplify model search, we have only limited graphical representations to link to model search methodologies.

Graphical log-linear models gave new impetus to the developments of log-linear model theory in the 1980s and 1990s, and there were related graphical representations for social network models linking individuals. But these two graphical representations remain unconnected. Elsewhere in multivariate analysis, researchers have exploited the duality between representations in spaces for individuals and for variables. Perhaps these ideas of duality of representations might allow us to link the two types of graphical structures into a new mathematical framework.

The problem of assessing bound for the entries of contingency tables given a set of marginals has a long statistical history going back to work done more than 50 years ago independently by Bonferroni, Fréchet, and Hoeffding on bounds for cumulative bivariate distribution functions given their univariate marginals (see Fienberg 1999 for a review of related literature). For the more general problem of a k-way contingency table given a set of possibly overlapping marginal totals, there are tantalizing links to the literature on log-linear models described in this article, including to the recent work on exact distributions and Gröbner bases described earlier. Implementation of bounds for large sparse tables is a special challenge.

More generally, as computer power and storage grows, researchers are attempting to work with larger and larger collections of categorical variables. We need new methods of model selection that scale up to situations where the dimensionality k of the table may exceed 100, and we need to revisit the asymptotics that are relevant for such situations.

REFERENCES

Agresti, A. (1990), *Categorical Data Analysis*, New York: Wiley.

——— (1992), "A Survey of Exact Inference for Contingency Tables" (with discussion), *Statistical Science*, 7, 131–177.

——— (1993a), "Computing Conditional Maximum Likelihood Estimates for Generalized Rasch Models Using Simple Loglinear Models with Diagonal Parameters, *Scandinavian Journal of Statistics*, 20, 63–72.

——— (1993b), "Distribution-free Fitting of Logit Models with Random Effects for Repeated Categorical Responses," *Statistics in Medicine*, 12, 1969–1987.

——— (1994), "Simple Capture-Recapture Models Permitting Unequal Catchability and Variable Sampling Effort," *Biometrics*, 50, 494–500.

Bartlett, M. S. (1935), "Contingency Table Interactions," *Journal of the Royal Statistical Society Supplement*, 2, 248–252.

Birch, M. W. (1963), "Maximum Likelihood in Three-Way Contingency Tables," *Journal of the Royal Statistical Society*, Ser. B, 25, 229–233.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press.

Carriquiry, A., and Fienberg, S. E. (1998), "Log-linear Models," in *Encyclopedia of Biostatistics*, Vol. 3, eds. P. Armitage and T. Colton, New York: Wiley, pp. 2333–2349.

Clogg, C. C., and Shidadeh, E. S. (1994), *Statistical Models for Ordinal Variables*, Thousand Oaks, CA: Sage.

Cressie, N. E., and Holland, P. W. (1983), "Characterizing the Manifest Probabilities of Latent Trait Models," *Psychometrika*, 48, 129–141.

Christensen, R. (1996), *Plane Answers to Complex Questions: The Theory of Linear Models* (2nd ed.), New York: Springer-Verlag.

Darroch, J. N. (1962), "Interaction in Multi-Factor Contingency Tables," *Journal of the Royal Statistical Society*, Ser. B, 24, 251–263.

——— (1986), "Quasi-Symmetry," in *Encyclopedia of Statistical Sciences*, Vol. 7, eds. S. Kotz and N. L. Johnson, New York: Wiley, pp. 469–473.

Darroch, J. N., Fienberg, S. E., Glonek, G., and Junker, B. (1993), "A Three-Sample Multiple-Recapture Approach to Census Population Estimation With Heterogeneous Catchability," *Journal of the American Statistical Association*, 88, 1137–1148.

Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980), "Markov Fields and Log-Linear Interaction Models for Contingency Tables," *The Annals of Statistics*, 8, 522–539.

Diaconis, P., and Sturmfels, B. (1998), "Algebraic Algorithms for Sampling From Conditional Distributions," *The Annals of Statistics*, 26, 363–397.

Diggle, P. J., Liang, K-Y., and Zeger, S. L. (1996), *Analysis of Longitudinal Data*, New York: Oxford University Press.

Duncan, O. D. (1983), "Rasch Measurement: Further Examples and Discussion," in *Survey Measurement of Subjective Phenomena*, Vol. 2, eds. C. F. Turner and E. Martin, New York: Russell Sage, pp. 367–403.

Fienberg, S. E. (1972), "The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables," *Biometrika*, 59, 591–603.

——— (1980), *The Analysis of Cross-Classified Categorical Data* (2nd ed.), Cambridge, MA: MIT Press.

——— (1999), "Fréchet and Bonferroni Bounds for Multi-Way Tables of Counts With Applications to Disclosure Limitation," in *Statistical Data Protection: Proceedings of the Conference*, Luxembourg: Eurostat, 115–129.

Fienberg, S. E., Johnson, M. A., and Junker, B. (1999), "Classical Multilevel and Bayesian Approaches to Population Size Estimation Using Data From Multiple Lists," *Journal of the Royal Statistical Society*, Ser. A, 162, 383–406.

Fienberg, S. E., and Meyer, M. M. (1983), "Loglinear Models and Categorical Data Analysis With Psychometric and Econometric Applications," *Journal of Econometrics*, 22, 191–214.

Fienberg, S. E., Meyer, M. M., and Wasserman, S. S. (1985), "Statistical Analysis of Multiple Sociometric Relations," *Journal of the American Statistical Association*, 80, 51–67.

Good, I. J. (1963), "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables," *Annals of Mathematical Statistics*, 34, 911–934.

Goodman, L. A. (1963), "On Methods for Comparing Contingency Tables," *Journal of the Royal Statistical Society*, Ser. A, 126, 94–108.

——— (1979), "Simple Models for the Analysis of Association in Cross-Classifications Having Ordered Categories," *Journal of the American Statistical Association*, 74, 537–552.

Haberman, S. J. (1974), *The Analysis of Frequency Data*, Chicago: University of Chicago Press.

——— (1977), "Log-Linear Models and Frequency Tables With Small Expected Cell Counts," *The Annals of Statistics*, 5, 1148–1169.

Holland, P. W., and Leinhardt, S. (1981), "An Exponential Family of Probability Distributions for Directed Graphs" (with discussion), *Journal of the American Statistical Association*, 76, 33–65.

Koehler, K. J., and Larntz, K. (1980), "An Empirical Investigation of Goodness-of-Fit Statistics for Sparse Multinomials," *Journal of the American Statistical Association*, 75, 336–344.

Lauritzen, S. L. (1996), *Graphical Models*, New York: Oxford University Press.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Plackett, R. L. (1974), *The Analysis of Categorical Data*, London: Charles Griffin.

Rasch, G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, The Danish Institute of Educational Research, expanded ed. (1980), Chicago: The University of Chicago Press.

Read, T. R. C., and Cressie, N. A. C. (1988), *Goodness-of-Fit Statistics for Discrete Multivariate Data*, New York: Springer-Verlag.

Roy, S. N., and Kastenbaum, M. A. (1956), "On the Hypothesis of No Interaction in a Multi-way Contingency Table," *Annals of Mathematical Statistics*, 27, 749–757.

Tjur, T. (1982), "A Connection Between Rasch's Item Analysis Model and a Multiplicative Poisson Model," *Scandinavian Journal of Statistics*, 9, 23–30.

Wasserman, S., and Pattison, P. (1996), "Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and p^*," *Psychometrika*, 61, 401–425.

Whitaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, New York: Wiley.