Three Centuries of Categorical Data Analysis: Log-linear Models and Maximum Likelihood Estimation

Stephen E. Fienberg* Alessandro Rinaldo †

Abstract

The common view of the history of contingency tables is that it begins in 1900 with the work of Pearson and Yule, but it extends back at least into the 19th century. Moreover it remains an active area of research today. In this paper we give an overview of this history focussing on the development of log-linear models and their estimation via the method of maximum likelihood. S. N. Roy played a crucial role in this development with two papers co-authored with his students S. K. Mitra and Marvin Kastenbaum, at roughly the mid-point temporally in this development. Then we describe a problem that eluded Roy and his students, that of the implications of sampling zeros for the existence of maximum likelihood estimates for log-linear models. Understanding the problem of non-existence is crucial to the analysis of large sparse contingency tables. We introduce some relevant results from the application of algebraic geometry to the study of this statistical problem.

1 Introduction

Most papers and statistical textbooks on categorical data analysis trace the history back to the work of Karl Pearson and George Udny Yule at the turn of the last century. But as Stigler (2002) notes, there is an early history of contingency tables dating to at least the 19th century of Quetelet (1849) on measuring association and hypergeometric analysis for the 2×2 table by Bienaymé (see e.g., Heyde and Seneta, 1977), and the introduction by Francis Galton (1892) of expected values of the form

$$\text{Expected Count}(i,j) = \frac{(\text{Row Marginal Total } i) \times (\text{Column Marginal Total } j)}{\text{Grand Total}}, \qquad (1)$$

as a baseline for measuring association, a formula that would later play a crucial role in chi-square tests for independence. Categorical data analysis remains an active area of research today and thus our history covers activities that span three centuries, the nineteenth, the twentieth, and the twenty-first, as is suggested by the title of this article.

The literature on categorical data analysis is now vast and there are many different strands involving alternative models and methods. Our focus here is largely on the development of log-linear models, maximum likelihood estimation, and the use of related chi-square tests of goodness of fit. In the next section of this paper we give an overview of the main part of this history beginning with the work of Pearson and Yule and running up to the present. S. N. Roy played a crucial role in this development with two papers co-authored with his students S. K. Mitra and Marvin

*Center for Automated Learning and Discovery and Cylab, Carnegie Mellon University
†Department of Statistics, Carnegie Mellon University

Kastenbaum, at roughly the mid-point temporally in this development. We explain the importance of Roy's contributions and how they influenced the development of the modern theory of log-linear models. Then in Section 3, we turn our attention to a problem whose full solution eluded statisticians beginning with the work of Bartlett (1935) until this past year, namely maximum likelihood estimation in the presence of sampling zeros. In Section 4 and 5 we illustrate, largely through examples, the nature and implication of the sampling zeros problem and we introduce the results that have begun to emerge from new tools in algebraic geometry applied to statistics, an area recently dubbed as *algebraic statistics* by Pistone et al. (2000).

2 Historical Development

As we mentioned at the outset, the history of categorical data analysis extends back well into the 19th century. Here we pick up the history at the beginning of the 20th century, focusing largely on those contributions that frame the development of log-linear models and maximum likelihood estimation. We do this in five parts: (1) Pearson-Yule through Neyman (1900-1950), (2) S. N. Roy's contributions, (3) Emergence of log-linear models in the 1960s, (4) The modern log-linear model era (1970s through present), (5) Other noteworthy categorical data models and methods. Agresti (2002) gives a complementary historical overview.

2.1 Contingency Tables, Chi-Square, and Early Estimation Methods

The term *contingency*, used in connection with tables of cross-classified categorical data, seems to have originated with Karl Pearson (1900) who, for an $s \times t$ table, defined contingency to be any measure of the total deviation from "independent probability." The term is now used to refer to the table of counts itself. Pearson (1900) laid the groundwork for his approach to contingency tables when he developed his chi-square test for comparing observed and expected (theoretical) frequencies:

$$X^2 = \sum_{i,j} \frac{(\text{Observed Count}(i,j) - \text{Expected Count}(i,j))^2}{\text{Expected Count}(i,j)}. \tag{2}$$

Yet Pearson preferred to view contingency tables involving the cross-classification of two or more polytomies as arising from a partition of a set of multivariate, normal data, with an underlying continuum for each polytomy. This view led Pearson (1904) to develop his tetrachoric correlation coefficient for 2×2 tables, and this work in turn spawned an extensive literature. The most serious problems with Pearson's approach were (a) the complicated infinite series linking the tetrachoric correlation coefficient with the frequencies in a 2×2 table, and (b) his insistence that it always made sense to assume an underlying continuum, even when the dichotomy of interest was dead–alive or employed–unemployed, and that it was reasonable to assume that the probability distribution over such a continuum was normal. In contradistinction, Yule (1900) chose to view the categories of a cross-classification as fixed, and he set out to consider the structural relationship among the discrete variables represented by the cross-classification, via various functions of the cross-product ratios. Especially impressive in this, Yule's first paper on the topic, is his notational structure for n attributes or 2^n tables, and his attention to the concept of partial and joint association of dichotomous variables.

The debate between Pearson and Yule over whose approach was more appropriate for contingency-table analysis raged for many years (see e.g., Pearson and Heron, 1913), and the acrimony it engendered was exceeded only by that associated with Pearson's dispute with R. A. Fisher over the adjustment in the degrees of freedom (d.f.) for the *chi-square test* of independence associated with a $s \times t$ table. In this latter case, Pearson, who argued that there should be no adjustment, was simply incorrect. As Fisher (1922) first noted, d.f. $= (s-1)(t-1)$. In arguing for a correction or adjusted degrees of freedom to account for the estimation of the parameters associated with the row and column probabilities, Fisher built the basis for the asymptotic theory of goodness of fit and model selection as we came to know it decades later. In addition, he related the estimation procedure explicitly with the characterization of structural association among categorical variables in terms of functions of odds ratios proposed by Yule (1900) for the 2^n table.

Bartlett (1935) introduced the first instance of a methodology for computing the maximum likelihood estimates (MLEs) for contingency tables. The author considered the case of what at the time would be called "complex contingency tables," namely the $2 \times 2 \times 2$ table displayed in Figure 1 for testing second order interactions. Bartlett showed that, in order to obtain the MLE under the model with no-second-order interaction, one needs to solve a cubic equation to determine a constant which then must be added and subtracted to the table cells in an appropriate order. Norton (1945) extended Bartlett's results to the case of $2 \times 2 \times t$ tables.

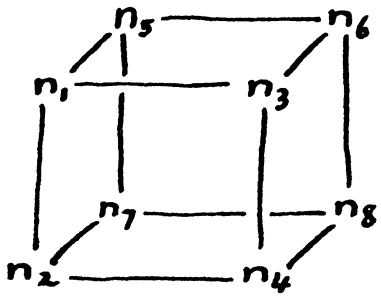

Figure 1: Bartlett's representation of a $2 \times 2 \times 2$ table (Bartlett, 1935, page 248).

Deming and Stephan (1940) proposed the method of iterative proportional fitting (IPF) for estimating the cell values in a contingency table subject to constraints coming from "known" marginal totals, e.g., from a population data set. The estimates were supposed to minimize a least squares criterion and were not related to statistical models of the sort usually associated with contingency tables. The methodology became known as "raking" and found widespread application in sampling, especially at the U.S. Census Bureau and other national statistical offices.

Another noteworthy development from the 1930s was the likelihood ratio test, proposed by Wilks (1935) as an alternative to Pearson's chi-square statistic:

$$G^2 = 2 \sum_{i,j} \text{Observed Count}(i,j) \log \left(\frac{\text{Observed Count}(i,j)}{\text{Expected Count}(i,j)} \right), \tag{3}$$

with the same asymptotic distribution under the null hypothesis of independence of row and column variables. Neyman (1949) added to the array of possible chi-square tests, setting the stage for the work of Roy and his students.

2.2 S. N. Roy's Contributions

S. N. Roy came to categorical data analysis from multivariate analysis more broadly and his principal contributions were in collaboration with two Ph.D. students at the University of North Carolina during the mid-1950s, S. K. Mitra and Marvin Kastenbaum. These results were also described as a chapter in Roy (1957).

In the first of these papers, Roy and Mitra (1956) began by making a clear distinction between response variables ('variates' in their terminology) and explanatory variables ('way of classification'), with the latter either being fixed by design or conditioned upon. They described the possible designs and models for two-way and three-way contingency tables and discussed how these are tied together through conditioning arguments, e.g. the model of homogeneity of proportions in a two-way table consisting of one response and one explanatory variable has a likelihood function that is derivable through conditioning from the model of independence for a pair of response variables. Then they derived asymptotic chi-square tests for these different situations, using the union-intersection principle Roy had developed in his earlier work on multivariate analysis, and they showed that the "equivalent" hypotheses/designs have the same maximum likelihood estimates and chi-square goodness-of-fit tests.

In the second paper, Roy and Kastenbaum (1956) filled in a major gap in the framework of Roy and Mitra (1956). They derived a formal mechanism for testing the hypothesis of no interaction in a 3-way table by offering new, "physically meaningful," multiplicative functional representations of cell probabilities which they then used for computing the MLE via Lagrange multipliers. Like Bartlett and Norton before them, Roy and Kastenbaum did not concern themselves with the possibility that some of the MLEs of the cell counts could be negative. Nonetheless, they implicitly noted the role played by "pivotal subscripts," taking values 0, $+1$ and -1, which are used to add or subtract certain quantities to the observed cell counts, in order to compute Pearson's chi-square statistic. As Birch (1963) would later point out, these results possess some undesirable properties, namely the estimates are not implicit functions of the frequencies and thus they are difficult to compute. Roy and Kastenbaum (1956) concluded by suggesting how their models generalize for higher-way tables.

Why did Roy and Kastenbaum not concern themselves with the existence of the MLE? While we can only speculate on the matter, the prevailing view at the time was that one needed relatively large cell counts for the validity of the asymptotic distribution of the chi-square tests. For example, Cochran (1952, 1954), in offering advice on the use of chi-quare tests, mentioned the need of a minimum expected value of 1 and relatively few expectations less than 5. If people were to follow this advice, then it is likely that they would encounter few if any sampling zeros.

This advice was driven by Cochran's attention to the comparison of the chi-square statistic with the exact distribution given the row and column totals, but the informal advice that resulted was that cell counts had to be at least 5 to use the chi-square methodology (see the related discussion of Cochran's work in Fienberg (1984). Thus for Roy and his students in the 1950s, sampling zeros did not pose an issue of statistical interest.

In the final section of his 1954 paper, Cochran presented a solution to the method for combining tests across a series of n 2×2 tables. This is in effect a way to test for conditional independence of the two binary variables given no second-order interaction in the $2 \times 2 \times n$ table. The link was not noted in Roy and Kastenbaums paper, and later Mantel and Haenszel (1959) independently proposed a similar test with two modifications.

Roy's influence ran deeper than the two papers with Mitra and with Kastenbaum. One of his

other Ph.D. students Vasant P. Bhapkar was to follow up on these ideas in a series of papers (e.g., see Bhapkar, 1961, 1966) and also in collaboration with Gary Koch (e.g., see Bhapkar and Koch, 1968). This work led to the paper by Grizzle et al. (1969) and a number of subsequent contributions by Koch and his students and colleagues.

2.3 The Emergence of Log-Linear Models and Methods

The 1960s saw a burgeoning literature on the analysis of contingency tables, largely but not solely focused on multiplicative or log-linear models. Key papers by Birch (1963), Darroch (1962), Good (1963), and Goodman (1963, 1964), plus the availability of high-speed computers, served to spur renewed interest in the problems of categorical data analysis and especially log-linear models and maximum likelihood estimation. Fienberg (1992), in the introduction to the reprinting of Birch (1963), details some of these developments.

Birch (1963) is, in many ways, the pivotal paper in the literature of the 1960s because it contains a succinct presentation of the basic results on maximum likelihood estimation for n-way contingency tables for $n \geq 3$, thereby generalizing and unifying the results derived by Roy and Mitra (1956) and Roy and Kastenbaum (1956). First, Birch introduced the use of the logarithmic expansion of the cell mean vector in terms of u-factors, thus allowing for the general log-linear representation and its connection with analysis of variance models. Then, under the assumption that the counts in the table are strictly positive, i.e., there are no sampling zero counts, he showed that the log-likelihood function admits a unique maximum, identical for a variety of sampling schemes. He also showed that the table marginal totals corresponding to the highest-order interaction terms in the model are the minimal sufficient statistics and that these marginal totals are equal to the maximum likelihood estimates of their expectations. This last result provided a justification for using the iterative proportional fitting algorithm to compute the MLEs of the expected cell values, which is in fact based on a sequence of cyclic adjustments using the marginal totals.

Bishop (1967, 1969) used Birch's results to derive connections between log-linear models and logit models, both from the theoretical and the computational point of view. She also proposed using a version of the iterative proportional fitting method developed by Deming and Stephan (1940) to perform computations for the MLE, as a practical way to implement the ideas of Birch to higher dimensional tables. The impetus for her work was the National Halothane Study, and Bishop applied the methodology to data from it. Because the tables of interest from this study exceeded the capacity of the largest available computers of the day, she was led to explore ways to simplify the IPF calculations by multiplicative adjustments to the estimates for marginal tables—an idea related to models with direct multiplicative estimates such as conditional independence, studied by Roy and Mitra. Moreover, despite the relatively large sample sizes, by the time the data were spread across the cells in the table, there were large numbers of zero counts, especially for the numerators in the rates of interest, i.e., surgical deaths in various categories. This practical application thus went well beyond the assumptions made by Birch and the literature that preceded him that the observed cell counts were all positive. Yet the methodology worked well, suggesting that the assumption could clearly be relaxed.

Fienberg, whose research had also been motivated by applications in the National Halothane Study, gave in Fienberg (1970a) a geometrical proof of the convergence of the IPF algorithm for tables with positive frequencies and showed that the rate of convergence is linear. He drew on the geometric representation of contingency tables described in Fienberg (1968) and Fienberg and Gilbert (1970), papers that anticipated some of the recent representations that have arisen in

algebraic statistics. Then, in Fienberg (1970b), he gave sufficient conditions for the existence of unique non-zero maximum likelihood estimates for the expected cell counts in incomplete tables for the model of quasi-independence, allowing for the presence of sampling zeros in the table. This allowed him to consider in an explicit manner the boundary points of the domain of the log-likelihood function under the mean value parametrization.

Building on ideas in Bishop (1967, 1969), Goodman (1970, 1971) presented methods for analyzing n-way tables using log-linear models and likelihood ratio statistics. In particular, he considered the class of hierarchical log-linear models in which the cell mean vector is expressible in closed form as a rational function of the sufficient statistics. For such models we can compute the MLE directly without resorting to any iterative numerical procedure. Goodman emphasized how these models are interpretable in terms of probability concepts such as independence, conditional independence and equiprobability. Haberman (1974) referred to these as *decomposable* models and studied them more thoroughly.

Generalizing the result by Birch to allow for sampling zeros, Haberman (1973) gave necessary and sufficient conditions for the existence of the MLE under Poisson and product-multinomial sampling schemes. One of his results provides a clear justification for the Bartlett construction of the MLE via a series of additions and subtractions of appropriate quantities to and from the observed frequencies. This observation led naturally to the "pathological" example of a $2 \times 2 \times 2$ table with positive margins and non-existent MLE shown in Figure 2. Haberman also gave an extended and rigorous proof, valid for general multi-way tables and log-linear models, of Birch (1963)'s result that the marginal totals of the MLE of the cell mean vector match the observed marginal totals and of the equivalence of the MLEs for Poisson and product-multinomial schemes. Furthermore, he provided a detailed study of IPF and Newton-Raphson's method for computing the MLE, introduced the general conditional Poisson sampling scheme for log-linear models, and gave an extensive derivation of the asymptotic properties of the MLE. He included these results, along with many others, in his 1974 monograph, Haberman (1974).

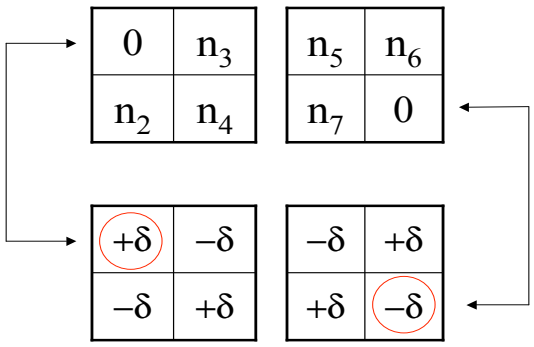

Figure 2: 2^3 table with only two sampling zeros and the model of no-second-order interaction, using Bartlett's notation from Figure 1. The two zero cells cause the MLE not to be defined because it is not possible to make one cell positive without making the other negative or changing the value of the margins.

Bishop et al. (1975) stressed the importance of being able to decide whether the MLE is defined

or not and, more generally, to characterize the configurations of sampling zeros associated with a non-existent MLE. Further, they raised concerns about the negative effect of a non-existent MLE on various inferential procedures and, in particular, on model selection.

2.4 The Modern Log-linear Model Era: Graphical Models Through to Algebraic Geometry

Darroch et al. (1980) introduced the formalism and language of graph theory and Markov properties for modeling interactions in the context of log-linear models for contingency tables. By representing conditional independence through the absence of an edge in the graph, they initiated what is now the theory of graphical statistical models. They also provided a novel graph-theoretic derivation of the properties of decomposable models and their MLEs. Later, Gloneck et al. (1988) proved, by means of counter-examples, that positivity of the margins is a necessary and sufficient conditions for existence of the MLE if and only if the model is decomposable. This work ultimately led to three major books, by Whittaker (1990), Edwards (1995) and Lauritzen (1996), which demonstrated the usefulness of the graphical representation, both for interpretation and for model search and related inferences, and which included additional theoretical insights.

Lauritzen (1996) offered many novel derivations of known results for decomposable and other models using the powerful machinery of graphical models. Although he was not directly concerned with the problem of existence of the MLE, he defined the parameter space to be the sequential closure of the vector subspace describing the log-linear model, which he termed *extended log-linear models*. By working with this enlarged parameter space, he was able to prove that the MLE is always defined, at least in an extended way, and furthermore, just as with the "ordinary" MLE, it satisfies the marginal equations.

Recent advances in the field of algebraic statistics, especially those following Diaconis and Sturmfels (1998), have suggested a more general approach to the study of log-linear models that takes advantage of the connections between algebraic and polyhedral geometry and the traditional theory of exponential families. Although Diaconis and Sturmfels (1998) originally introduced this formalism for studying the "exact" distribution over contingency tables under a log-linear model given its minimal sufficient statistics, i.e. the observed marginal totals , Fienberg (2000) suggested that these tools from algebraic geometry would be of potential value beyond this problem. Eriksson et al. (2006) and Rinaldo (2005) have begun to deliver on that promise. We illustrate some of their findings in the next sections.

2.5 Some Other Notable Contributions to the Analysis of Categorical Data

There have been many other important contributions to the analysis of categorical data over the past three decades which we do not have space to document in the present paper. But we would be remiss if we did not at least mention some of them even without extensive references:

- Model generation and testing using minimum modified chi-square estimation and the Wald statistic. This methodology emanated directly from the work of Roy and his students and often works for problems that do not have a convenient log-linear representation.

- Since the early 1970s much application involving log-linear models has been done in the context of generalized linear models and GLM computing methods. Unfortunately, as we

explain in Section 4, these methods are not good at handling the problems of zeros described in the next section.

- Goodman (1979, 1981) developed a special class of models which extend the log-linear family by replacing various interaction terms by multiplicative counterparts. Such models, which are often referred to as "association" models are especially valuable when we are dealing with ordinal as opposed to nominal categorical variables and give reduced numbers of parameters compared with traditional log-linear models.

- Correspondence analysis is yet another alternative to log-linear models that has found adherents and its structure is close to but different from that of "association" models, which resemble log-linear models with multiplicative interaction terms.

- Rasch models for item response theory and latent class models have emerged as a powerful extension or alternative to traditional log-linear model analysis.

- Bayesian methods for log-linear models have come into their own with the emergence of computation tools such as Monte Carlo Markov chain methods.

- A particular alternative to log-linear models that has emerged in the past two decades is the *Grade of Membership model* (GOM) which can be interpreted as a soft clustering methodology. The GoM turns out to be a special case of a class of Bayesian mixed membership models that work for text data and images as well.

- Causal modeling uses the directed graphical structures that entered the literature with Darroch et al. (1980). But the ideas from causal inference are more elaborate.

Goodman (1985, 1996) compares association and correspondence analysis models. Erosheva et al. (2002) give a succinct comparison of log-linear, Rasch, and GoM models. Pearl (2000) and Spirtes et al. (2001) give treatments of causal modeling including the role of latent variables and counterfactuals.

3 The Problem of Zeros and Existence of MLEs

Log-linear models are a powerful statistical tool for the analysis of categorical data and their use has increased greatly over the past two decades with the compilation and distribution of large, and very often sparse, data bases, in the social and medical sciences as well as in machine learning applications. In log-linear model analysis, the MLE of the expected value of the vector of observed counts plays a fundamental role for assessment of fit, model selection and interpretation.

The existence of the MLE is essential for the usual derivation of large-sample chi-square approximations to numerous measures of goodness of fit (Bishop et al., 1975; Agresti, 2002; Cressie and Read, 1988) which are utilized to perform hypothesis tests and, most importantly, are an integral part of model selection. If the distribution of the statistic measuring the goodness of fit is instead derived from the "exact distribution," or the conditional distribution given the minimal sufficient statistics (the margins), it is still necessary in most cases to have an MLE or some similar type of estimate in order to quantify the discrepancy of the the observed data from the fitted values. We also require the existence of the MLE to obtain a limiting distribution in the double-asymptotic

approximations for the likelihood ratio and Pearson chi-square statistic for tables in which both the sample size and the number of cells are allowed to grow unbounded, a setting studied, in particular, by Morris (1975), Haberman (1977) and Koehler (1986) (see Cressie and Read, 1988, for a relatively complete literature review). If the MLE is not defined, inferential procedures such as those involved with model search that use chi-square tools may not be applied meaningfully or, at a minimum, require alteration.

As we noted in our historical review in the preceding section, the non-existence of the MLE has long been known to relate to the presence of zero cell counts in the observed table (see, in particular, Bishop, 1967; Goodman, 1970; Haberman, 1974; Bishop et al., 1975). Indeed, a number of people informally observed that the only thing that appeared to matter was the numbers and locations of the the sampling zeros and not the values of the counts in the remaining cells (see, e.g., Fienberg, 1970b). Although Haberman (1974) gave necessary and sufficient conditions for the existence of the MLE, his characterization is nonconstructive in the sense that it does not directly lead to implementable numerical procedures and also fails to suggest alternative methods of inference for the case of an undefined MLE. Despite these deficiencies, Haberman (1974)'s results have remained all that exist in the published statistical literature. Furthermore, no one has presented yet a numerical procedure specifically designed to check for existence of the MLE, and the only indication of non-existence is lack of convergence of whatever algorithm is used to compute the MLE. As a result, the possibility of the non-existence of the MLE, even though well known, is rarely a concern for practitioners. Moreover, even for those cases in which the non-existence is easily detectable, e.g. when the observed table exhibits zero margins, there do not exist appropriate inferential procedures for dealing with such a situation.

Although zero counts can occur in small tables in which the expected values of some cells are significantly smaller than the others, they arise regularly in large tables in which the sample size is small relative to the number of cells. In particular, the non-existence of the MLE is potentially a very common problem in all applications in which the data are collected in the form of large and sparse databases, as in the social, biological and medical sciences. In many such cases, a common mispractice is to collapse the original table into smaller, less sparse, tables with fewer categories or variables. As Bishop et al. (1975) and Asmussen and Edwards (1983) have made clear, such collapsing can potentially lead to misleading and incorrect statistical inference about associations among the variables.

Identifying the cases in which the MLE is not defined has immediate practical implications and is crucial for applying appropriate modifications to traditional procedures of model selection based on both asymptotic and exact approximations of test statistics and for developing new inferential methodologies to deal with sparse tables.

Eriksson et al. (2006) offered a novel geometric interpetation of the necessary and sufficient conditions for the existence of the MLE as originally proved in Haberman (1974) for hierarchical log-linear models. Their findings were further generalized by Rinaldo (2005) to include non-hiearchial models and conditional Poisson sampling schemes. Overall, these results allowed for a full characterization for all possible patterns of sampling zeros in the table causing the MLE to be undefined and also led to efficient numerical procedures for checking the existence of the MLE.

As initially noted by Haberman (1974) and later formalized by Lauritzen (1996), under mean value parametrization the log-likelihood function always admits a unique maximizer, even if the MLE is not defined; for cases in which the MLE does not exist, Haberman (1974) heuristically called these maximizers *extended* MLEs. By combining results from the theory of linear exponential

families (see, e.g., Brown, 1986) with results from algebraic geometry, Rinaldo (2005) provided a rigorous definition of extended MLEs, derived some of their properties and proposed a two-step procedure for performing extended maximum likelihood estimation for log-linear models. In step one we identify the problematic sampling zeros that cause the MLE not to exist, and then in step 2 we condition on these and fit a a log-linear model to the remaining cells.

In the remainder of this article, we demonstrate, by means of examples, aspects of the patterns of zeros that lead to non-existence of the MLE and various practical considerations that follow from the non-existence. We believe that the examples below give a good indication of the difficulties and subtleties, both computational and theoretical, associated with this problem.

Remark Throughout this article, we use the wording "non-existence of the MLE" to signify lack of solutions for the maximum likelihood optimization problem, in accordance with a terminology long established in the log-linear model literature (see, for example, Birch, 1963; Fienberg and Gilbert, 1970; Haberman, 1974). Alternatively, we can say that the MLE of the cell mean vector does not exist whenever there is no strictly positive solution to the MLE defining equations (see, for example, Haberman, 1974, Equation 2.11).

4 Illustrations of Non-Existence Problems

Since, as we mentioned, Haberman (1974)'s necessary and sufficient conditions for the existence of the MLE provide only a non-constructive characterization, it is not surprising that virtually all implemented computational algorithms for fitting log-linear models are, by design, incapable of handling these cases. The following excerpts from the SAS online documentation[1] for the PROC FREQ, PROC CATMOD and PROC GENMOD procedures exemplifies this situation.

- *"By default,* PROC FREQ *does not process observations that have zero weights, because these observations do not contribute to the total frequency count, and because any resulting zero-weight row or column causes many of the tests and measures of association to be undefined."*

- *"For a log-linear model analysis using* WLS *or* ML=NR*,* PROC CATMOD *creates response profiles only for the observed profiles. Thus, for any log-linear model analysis with one population (the usual case), the contingency table will not contain zeros, which means that all zero frequencies are treated as structural zeros. If there is more than one population, then a zero in the body of the contingency table is treated as a sampling zero (as long as some population has a nonzero count for that profile). If you fit the log-linear model using* ML=IPF*, the contingency table is incomplete and the zeros are treated like structural zeros. If you want zero frequencies that* PROC CATMOD *would normally treat as structural zeros to be interpreted as sampling zeros, you may specify the* ZERO=SAMPLING *and* MISSING=SAMPLING *options in the* MODEL *statement. Alternatively, you can specify* ZERO=1E-20 *and* MISSING=1E-20*.[...] sampling zeros in the input data set should be specified with the* ZERO= *option to ensure that these sampling zeros are not treated as structural zeros. Alternatively, you can replace cell counts for sampling zeros by some positive number close to zero (such as 1E-20) in a* DATA *step."*

- *"*PROC GENMOD *treats each observation as if it appears n times, where n is the value of the* FREQ *variable for the observation. If it is not an integer, the frequency value is truncated to an integer. If it is less than 1 or if it is missing, the observation is not used."*

[1] available at `http://support.sas.com/onlinedoc/913/docMainpage.jsp`

Table 1: Example of a 2^3 table with non-existent MLE for the model of no-second-order interaction, $[12][13][23]$.

0	1
2	4

2	3
2	**0**

Table 2: Example of a 3^3 table with non-existent MLE for the model of no-second-order interaction, $[12][13][23]$.

0	4	**0**
0	1	2
5	1	2

4	2	2
5	5	2
1	**0**	**0**

5	3	4
2	1	3
2	**0**	**0**

In SAS, the presence of sampling zeros is dealt with by adding small positive quantities to the zero cells to facilitate the convergence of the underlying numerical procedure. This common practice can be very misleading, as the example in Table 5 below demonstrates.

We illustrate various issues and dangers associated to the usage of very common computational procedures for obtaining the MLE with artificially constructed tables and with a simple real-life example of a non-sparse contingency table. All the examples we present are still sufficiently small that, in principle, the user is able to detect lack of existence of the MLE by looking, heuristically, at the erratic behavior of the fitting algorithm or by applying directly some of the results from the theory. Both the theoretical results and the computational tools available to researchers, however, are of little help when we go to analyze high-dimensional complex and/or sparse databases.

4.1 Artificial Tables

Tables 1 and 2 show two examples of non-existent MLEs for the hierarchical log-linear model of no-second-order interaction, $[12][13][23]$. In both cases, the relevant margins are strictly positive, a condition which in practice is very frequently, but erroneously, taken to be necessary and sufficient for the MLE to exist and in the literature has been described as "pathological" (Bishop et al., 1975, page 115).

We have already illustrated the cause of non-existence for Table 1 in Figure 2. For Table 2 we can conveniently explain non-existence by using collapsing arguments (see Eriksson et al., 2006, Section 4.1). If we collapse rows 1 and 2 and columns 3 and 4, then the resulting pattern of zeros looks like that in Table 1. But unlike the situation in Table 1, not all the zero cells impact negatively on the existence of the MLE, i.e., the MLE would still not exist if the zero cell in position $(1, 3)$ in the first array were to be replaced by any positive count.

Despite the reduced dimension of both these tables, it is quite possible that, if we use standard statistical software for fitting log-linear models, we will not detect non-existence. We illustrate the problem with a short computational study of the behavior of two among the most commonly

used algorithms to calculate the MLE: the IPF algorithm and the Iterative Weighted Least Squares procedure, described by Charnes et al. (1976), which is an application of the Newton-Raphson optimization procedure.

For the IPF procedure we used the `loglin` routine in `R`, which implements a Fortran algorithm written by Haberman (1972, 1976). We used the default maximum number of iterations which is set at 20 and the default tolerance parameter, which measures the maximum deviation allowed between observed and fitted margins, and is set at 0.1. For the re-weighted least squares algorithm, we used the routine `glm` for fitting generalized linear models (McCullagh and Nelder, 1989; Hastie and Pregibon, 1991), with the parameter `family` set equal to `poisson`. The default maximum number of iterations is 25 and the default convergence tolerance is 10^{-8}, for the criterion

$$\frac{|\mathrm{d}_{\mathrm{new}} - d_{\mathrm{old}}|}{\mathrm{d}_{\mathrm{new}} + 0.1},$$

where d_{new} and d_{old} are the deviance at the current and previous iteration, respectively.

We summarize the computations performed using different values of the default parameters for both routines in Table 3 and offer the following observations:

1. The IPF algorithm in most cases fails to satisfy the convergence criterion, even after a large number of iterations, a surprising result given the low dimensionality of both tables. When the MLE is not defined, the IPF is guaranteed to converge to the extended MLE by design. This example shows that the rate of convergence to the extended MLE, which is linear when the MLE exists (Fienberg, 1970a), can be very slow. The behavior of IPF when the MLE does not exist has not been carefully studied to date.

2. The Newton-Raphson procedure in the `glm` routine uses a more stringent convergence criterion than the default ones for the `loglin` routine, because it converges at a much faster rate than the IPF algorithm when a solution exists. Precisely because of its numerical robustness, however, the Newton-Raphson method does not provide any indication, at least in small tables, that the MLE *does not* exists!

3. When they fail to convergence, both procedures will simply produce a warning message along with the output of fitted values and the usual number of degrees of freedom, correct only for the case in which the MLE is defined.

These simple examples illustrate the fact that detecting non-existence of the MLE for tables with positive margins using available software requires a careful monitoring of the convergence of the algorithm used to calculate the MLE. Although the examples seem to suggest that the Newton-Raphson method is computationally much more stable, in reality this is not the case. In fact, the estimated parameters of the `glm` routine tend to explode when the MLE does not exist because the maximum occurs on the boundary of the parameter space at minus infinity. Thus the numerical stability observed here is not imputable to the algorithm itself, but this is probably due to the small dimensionality of these examples. Rinaldo (2005) proposed modifications of the Newton-Raphson procedure that allow it to exploit its fast rate of convergence and, at the same time, to eliminate any exploding behavior.

Table 3: Summary of the computations performed on the Tables 1 and 2. Default values are marked with a (*).

Routine	Tolerance	Iterations	Convergence	Warnings
Table 1: 2^3 table and model [12][23][13]				
loglin	1e-0 (*)	12	Yes	
loglin	1e-05	20 (*)	No	stop("this should not happen")
loglin	1e-05	10,272	Yes	
glm	1e-08 (*)	21	Yes	
loglin	1e-08	500,000	No	stop("this should not happen")
Table 2: 3^3 table and model [12][23][13]				
loglin	1e-01 (*)	20 (*)	No	stop("this should not happen")
loglin	1e-01	27	Yes	
loglin	1e-05	251,678	Yes	
glm	1e-08 (*)	19	Yes	
glm	1e-12	28	Yes	
glm	1e-13	30	No	fitted rates numerically 0

4.2 Clinical Trial Example

Although non-existence of the MLE arises most frequently in sparse tables, it can very well occur also in tables with large counts and very few zero cells. Table 4 shows a $2\times2\times2\times3$ contingency table from Koch et al. (1983), which describes the results of a clinical trial to examine the effectiveness of an analgesic drug, for patients of two statuses and centers. The sample size is relatively large ($n = 193$) with respect to the number of cells ($p = 24$) and, except for two zero counts, the cell counts are quite big.

With the goal of illustrating statistical disclosure limitation techniques and discussing the risk of disclosure associated to various marginal releases, Fienberg and Slavkovic (2004) analyzed two nested models, both of which fit the data of Table 4:

1. [CST][CSR],
2. [CST][CSRT][TR].

The [CSR] margins has one zero entry, which causes the non-existence of the MLE for both models. Since the two models are decomposable (see, e.g., Lauritzen, 1996), the IPF algorithm

converged almost instantaneously (less than 4 iterations) in both cases. In fact, the remarkable efficiency achieved by the IPF algorithm for decomposable models, demonstrated by Haberman (1974), is not affected by the non-existence (see Rinaldo, 2005). As a result, because of this fast rate of convergence, no indications of non-existence is provided, except that some fitted values are zeros.

Table 4: Results of clinical trial for the effectiveness of an analgesic drug. Source: Koch et al. (1983).

Center	Status	Response / Treatment	Poor	Moderate	Excellent
1	1	Active	3	20	5
		Placebo	11	14	8
	2	Active	3	14	12
		Placebo	6	13	5
2	1	Active	12	12	**0**
		Placebo	11	10	**0**
	2	Active	9	3	4
		Placebo	6	9	3

The pathology of zeros in the margins leading to non-existence is easy to detect and deal with as we have done here. The same type of pathology also occurs in the $2 \times 2 \times 2$ and $3 \times 3 \times 3$ tables.

4.3 Other Artificial Examples

The problem of non-existence of the MLE, even in 3-way tables, is not simply reducible to collapsing or zeros in the margins, as is the case for the examples of Table 2 and 4, respectively. However, more complex examples have been difficult to construct until recently.

Using a polyhedral geometry representation and the software package `polymake` (Gawrilow and Joswig, 2000), Eriksson et al. (2006) have been able to compute the number of degeneracies caused by patterns of zeros for $p \times q \times r$ tables for $p = 2, 3$, and 4 and various choices of q and r, for the same no-second-order-interaction model. One of the first examples they discovered where the pattern of zeros is not simply characterized corresponds to the one illustrated in Table 5. After replacing the zero entries in this table with the small positive number 10^{-8}, as recommended for the `SAS` procedures mentioned in Section 4, we computed the MLE using the `R` routine `loglin` with a tolerance level of 10^{-8} (incidentally, we note that the algorithm failed to converge within $500,000$ iterations). The values of the X^2 statistic (2) and of the G^2 statistic (3) are 7.37021×10^{-6} and 1.472673×10^{-05}, respectively, which, using a χ^2 distribution with 8 d.f., results in p-values of nearly 1. In fact, the values of both goodness-of-fit statistics will always be almost zero, *no matter what the positive counts are*. Thus for this pattern of zero cells, we virtually never reject the fitted model.

Eriksson et al. (2006) studied many other examples and observed that the number of possible patterns of zero counts invalidating the MLE exhibits an exploding behavior as the number of classifying variables or categories grows, so much so that computations become quickly unfeasible.

Table 5: Example of a 3^3 table with non-existent MLE for the model of no-second-order interaction, $[12][13][23]$.

0	2	**0**
5	1	4
0	2	3

2	2	3
1	**0**	4
0	**0**	2

3	5	**0**
1	**0**	**0**
2	4	1

Table 6: Example of a 4^3 table with non-existent MLE for the model of no-second-order interaction, $[12][13][23]$.

0	**0**	**0**	4
0	**0**	1	2
0	1	2	3
5	1	2	3

4	**0**	**0**	2
5	**0**	5	2
5	6	5	2
1	**0**	**0**	**0**

1	5	**0**	2
5	3	4	2
0	2	**0**	**0**
1	2	**0**	**0**

1	5	3	2
0	**0**	2	**0**
0	2	4	**0**
1	2	3	**0**

Table 6 shows a more complex pattern for a larger 4^3 table. Note that, as was the case with Tables 1, 2 and 5, the 2-way margins are strictly positive here, but we cannot reduce the problem to a lower dimensional non-esistence case through collapsing.

We leave it to the interested reader to test the computation of the MLE for these tables in their favorite log-linear model computer program.

5 The 2^K Table and The No-$(K-1)$st-Order Interaction Model

In the $2\times2\times2$ table with the no-second order interaction model all problems of non-existence of the MLE come about if there are 2 zero counts. But not all pairs of zero counts cause problems. There are $\binom{8}{2} = 28$ ways to place 2 zeros in the table, and 12 of these pose no difficulties for maximum likelihood estimation. These correspond to taking a row, a column, or a layer (6 possible choices) and placing zeros in the corners of the corresponding 2×2 table (2 ways). The remaining 16 cases are degenerate. There can be a zero in a two-way marginal total and there are 12 possibilities for this to happen. And then finally we can place zeros into the "touching corners" of the full 2^3 table and there are 4 ways to do this.

Now we consider the 2^K contingency table and the model of no-$(K-1)$st-order interaction, for $K \geq 3$. Suppose that the table contains only two sampling zeros and has positive margins. Because the model is very constrained, it is very likely that the MLE does not exist. What we show is that the chance of this happening increases very fast with the number of factors K, a somewhat counter-intuitive fact.

Proposition 5.1. For a 2^K contingency table and the model of no-$(K-1)$st-order interaction, the probability that two randomly-placed sampling zeros cause the MLE not to exist without inducing

zero margins is

$$\frac{(2^{K-1}-K)}{(2^K-1)}. \tag{4}$$

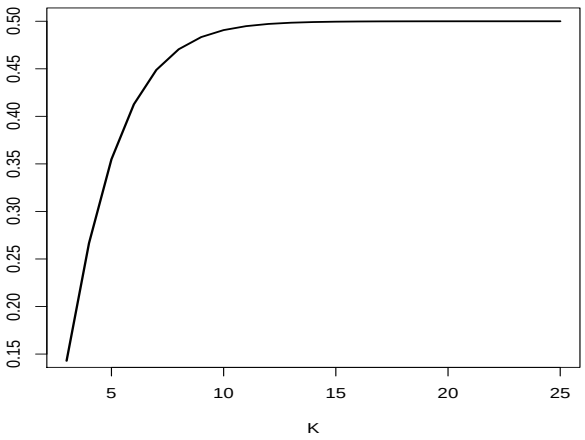

Figure 3: Probability (4) that two randomly sampling placed zeros cause the MLE to be undefined without inducing zero margins as a function of the number K of factors for the 2^K table and the model of no-$(K-1)$st-order interaction, $K \geq 3$.

Proof. We prove this result using a counting argument by identifying each pair of sampling zeros with one of the $\binom{2^K}{2}$ possible edges of the complete graph on K vertices. The orthogonal complement in $\mathbb{R}^{2^K}$ of the log-linear subspace for the model of no-$(K-1)$st-order interaction has dimension 1 and is spanned by a 2^K-dimensional vector $\boldsymbol{\delta}$, half of whose coordinates are $+1$ and the other half are -1. By Theorem 2.2 in Haberman (1974), the MLE is not defined whenever the two sampling zeros correspond to coordinates of $\boldsymbol{\delta}$ of opposite sign. Therefore, the number of edges leading to an existing MLE is

$$2\binom{2^{K-1}}{2} = 2^{K-1}(2^{K-1}-1).$$

Since the number of edges associated to zero margins is $K2^{K-1}$, the number of edges causing the MLE not to be defined, without generating null margins, is

$$\begin{aligned}
\tbinom{2^K}{2} - 2^{K-1}(2^{K-1}-1) - K2^{K-1} &= 2^{K-1}(2^K-1) - 2^{K-1}(2^{K-1}-1) - K2^{K-1} \\
&= 2^{K-1}\left(2^K - 1 - 2^{K-1} + 1 - K\right) \\
&= 2^{K-1}(2^K - 2^{K-1} - K) \\
&= 2^{K-1}(2^{K-1} - K).
\end{aligned}$$

The probability of this occurrence is then computed as

$$\frac{2^{K-1}(2^{K-1}-K)}{\binom{2^K}{2}} = \frac{2^{K-1}(2^{K-1}-K)}{2^{K-1}(2^K-1)} = \frac{(2^{K-1}-K)}{(2^K-1)}. \quad \square$$

Table 7: Example of a 3^3 table with many sampling zeros but for which the MLE for the model of no-second-order interaction, $[12][13][23]$, exists.

3	0	0	0	0	1	0	1	0
0	4	0	5	0	0	0	0	5
0	0	4	0	2	0	3	0	0

□

The probability (4) is a strictly increasing concave function of K and tends monotonically to $\frac{1}{2}$ as $K \uparrow \infty$. The limiting behavior of Equation (4) is illustrated in Figure 3. The convergence occurs very rapidly.

The previous result is rather peculiar as it is concerned with binary data and the most constrained hierarchical log-linear model. In the presence of less saturated models and/or more categories, the patterns of sampling zeros are in general much harder to descibe and intuition is of little help. As an example of this combinatorial complexity, Table 7 shows a 3^3 contingency table sparser than the ones presented in Table 2 and 5 but for which the MLE is well defined.

6 Conclusions

Categorical data analysis has a long and honorable tradition in the statistics literature, going back at lease to Quetelet and spanning the nineteenth, twentieth, and twenty-first centuries. In Section 2 of the paper we have provided an overview of this history. The major developments of log-linear models and their estimation using the method of maximum likelihood emerged during the period from 1900 to about 1975 and S.N. Roy and his students, S. K. Mitra and Marvin Kastenbaum, played a pivotal role in these developments. But because of the limits of computation at the time and the common wisdom that one should only analyze tables with positive entries, they did not entertain the prospect of analyzing large sparse contingency tables where the issue of the existence of the maximum likelihood estimates of the cell counts would be a serious issue.

In Sections 3, 4, and 5 of this paper we illustrated, through a series of examples, the nature of the problem of sampling zeros and we tried to explain its importance. Computational advances mean we can conceive of handling larger and larger tables in our statistical analyses, and this inevitably means we must face up to the problem of the existence of the MLE. Until the recent introduction of ideas from algebraic geometry into statistics, we had few tools to do so in a constructive fashion. The examples we have used to illustrate existence characterizations are drawn from Eriksson et al. (2006) and Rinaldo (2005). We hope to implement the characterizations in these sources in actual computer programs in the near future.

Silvapulle (1981) gave necessary and sufficient conditions for the existence of the MLE in binomial response models. These conditions are formulated in terms of convex cones spanned by column vectors of the design matrix and appear to be closely related to the polyhedral conditions for the existence of the MLE in log-linear models derived by Eriksson et al. (2006). The study of the connections between the non-existence of the MLE for log-linear models and for logistic models

involving categorical covariates deserves thorough investigation.

We conclude by discussing some issues relating the problem of non-existence of the MLE with conditional, or "exact", inference. Being caused by the presence of sampling zeros, non-existence of the MLE is more likely to occur in sparse tables with small counts, a setting in which the traditional χ^2 asymptotic approximation to various measures of goodness of fit is known to be unreliable (see, e.g., Cressie and Read, 1988). In these cases, inference can be conducted by using as a reference distribution the conditional distribution of the tables given the observed sufficient statistics (i.e. the margins), also known as the "exact" distribution. Starting with the seminal work of Diaconis and Sturmfels (1998), new results from algebraic statistics have provided powerful characterizations of the exact distribution and have emphasized the considerable difficulties associated with sampling from it. For example, recent work by De Loera and Onn (2006) demonstrated that the exact distribution can in fact be largely disconnected and possibly multi-modal and that the computational complexity of any sampling procedure cannot be bounded. With respect to the problem of the existence of the MLE, we note that the computation of the conditional distribution of measures of goodness of fit is closely related with the existence of the MLE. In fact, when the MLE does not exist, all the tables in the support of the exact distribution have zero entries precisely in the cells for which the expected counts cannot be estimated using maximum likelihood. Moreover, for log-linear models there are no general results characterizing the optimality of exact methodologies. For exponential families, conditioning on the sufficient statistics is a device for eliminating nuisance parameters and for obtaining optimal tests and confidence intervals; however, the validity of this approach has been shown only for restricted models and very small tables (see, e.g., Agresti, 1992; Lehmann and Romano, 2005) and it is unclear whether it can be extended to general log-linear models on tables of arbitrary dimension. Further, it is unclear whether the exact distribution provides tools for inference in sparse tables that can be considered optimal in some sense.

ACKNOWLEDGMENTS

The research reported here was supported in part by NSF grants EIA9876619 and IIS0131884 to the National Institute of Statistical Sciences and draw in part on material from earlier historical overviews by the first author, including the one in Fienberg (1980), as well as from the Ph.D. dissertation of the second author. We thank a referee for helpful comments and the reference to Silvapulle (1981).

References

Agresti, A. (1992). A survey of exact inference for contegency tables. *Statitical Science 7*, 131–153.

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). New York: Wiley.

Asmussen, S. and D. Edwards (1983). Collapsibility and response variables in contingency tables. *Biometrika 70*, 567–578.

Bartlett, M. S. (1935). Contingency table interactions. *Supplement to the Journal of the Royal Statistical Society 2*(2), 248–252.

Bhapkar, V. P. (1961). Some tests for categorical data. *Annals of Mathematical Statistics 29*, 302–306.

Bhapkar, V. P. (1966). A note on the equivalence of two test criteria for hypotheses in categorical data. *Journal of the American Statistical Association 61*, 228–235.

Bhapkar, V. P. and G. G. Koch (1968). On the hypotheses of "no interaction" in contingency tables. *Biometrics 24*, 567–594.

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society 25*, 220–233.

Bishop, Y. M. M. (1967). *Multi-dimensional Contingency Tables: Cell Estimates*. Ph. D. thesis, Department of Statistics, Harvard University.

Bishop, Y. M. M. (1969). Full contingency tables, logits, and split contingency tables. *Biometrics 25*(2), 119–128.

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

Brown, L. D. (1986). *Fundamental of Statistical Exponential Families*, Volume 9 of *IMS Lecture Notes-Monograph Series*. Hayward, CA: Institute of Mathematical Statistics.

Charnes, A., E. L. Frome, and P. L. Yu (1976). The equivalence of generalized least squares and maximum likelihood estimates in the exponential family. *Journal of the American Statistical Association 71*(353), 169–171.

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics 23*, 315–345.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics 10*, 417–451.

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

Darroch, J. N. (1962). Interaction in multi-factor contingency tables. *Journal of the Royal Statistical Society 24*, 251–263.

Darroch, J. N., S. L. Lauritzen, and T. P. Speed (1980). Markov fields and log-linear interaction models for contingency tables. *Annals of Statistics 8*(3), 522–539.

De Loera, J. and S. Onn (2006). Markov bases of three-way tables are arbitrarily complicated. *Journal of Symbolic Computation 41*, 173–181.

Deming, W. E. and F. F. Stephan (1940). On a least square adjustment of a sampled ffrequency table when the expected marginal totals are known. *Annals of Mathematical Statistics 11*, 427–444.

Diaconis, P. and B. Sturmfels (1998). Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics 26*(1), 363–397.

Edwards, D. (1995). *Introduction to Graphical Modelling*. New York: Springer-Verlag.

Eriksson, N., S. E. Fienberg, A. Rinaldo, and S. Sullivant (2006). Polyhedral conditions for the nonexistence of the mle for hierarchical log-linear models. *Journal of Symbolic Computation 41*, 222–233.

Erosheva, E. A., S. E. Fienberg, and B. W. Junker (2002). Alternative statistical models and representations for large sparse multi-dimensional contingency tables. *Annales de la Faculté des Sciences de Toulouse 11*(4), 485–505.

Fienberg, S. E. (1968). The geometry of an $r \times c$ contingency table. *Annals of Mathematical Statistics 39*(4), 1186–1190.

Fienberg, S. E. (1970a). An iterative procedure for estimation in contingency tables. *Annals of Mathematical Statistics 41*, 907–917.

Fienberg, S. E. (1970b). Quasi-independence and maximum likelihood estimation in incomplete contingency tables. *Journal of the American Statistical Association 65*(232), 1610–1616.

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

Fienberg, S. E. (1984). The contributions of William Cochran to categorical data analysis. In P. S. R. S. Rao (Ed.), *W. G. Cochran's Impact on Statistics*, pp. 103–118. New York: Wiley.

Fienberg, S. E. (1992). Introduction to paper by M. W. Birch. In S. Kotz and N. Johnson (Eds.), *Breakthroughs in Statistics: Volume II, Methodology and Distribution*, pp. 453–461. New York: Springer-Verlag.

Fienberg, S. E. (2000). Contingency tables and log-linear models: Basic results and new developments. *Journal of the American Statistical Association 95*, 643–647.

Fienberg, S. E. and J. P. Gilbert (1970). The geometry of a two by two contingency table. *Journal of the American Statistical Association 65*(330), 694–701.

Fienberg, S. E. and A. B. Slavkovic (2004). Making the release of confidential data from multi-way tables count. *Chance 17*(3), 5–10.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society 85*(1), 87–94.

Galton, F. (1892). *Finger Prints*. London: Macmillan.

Gawrilow, E. and M. Joswig (2000). Polymake: a framework for analyzing convex polytopes. In *Polytopes, Combinatorics and Computation*. Boston, MA: Birkhauser.

Gloneck, G., J. N. Darroch, and T. P. Speed (1988). On the existence of the maximum likelihood estimator for hierarchical log-linear models. *Scandinavian Journal of Statistics 15*, 187–193.

Good, I. J. (1963). Maximum entropy for hypotheses 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 126*, 94–108.

Goodman, L. A. (1964). Simultaneous confidence limits for cross-product ratios in contingency tables. *Journal of the Royal Statistical Society 26*, 86–102.

Goodman, L. A. (1970). The multivariate analysis of qualitative data: Interactions among multiple classifications. *Journal of the American Statistical Association 65*(329), 226–256.

Goodman, L. A. (1971). Partitioning of chi-square, analysis of marginal contingency tables, and estimation of expected frequencies in multidimensional contingency tables. *Journal of the American Statistical Association 66*(334), 339–344.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association 74*, 537–552.

Goodman, L. A. (1981). Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *Journal of the American Statistical Association 76*, 320–334.

Goodman, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *The Annals of Statistics 13*, 10–69.

Goodman, L. A. (1996). A single general method for the analysis of cross-classified data: Reconciliation and synthesis of some methods of Pearson, Yule, and Fisher, and also some methods of correspondence analysis and association analysis. *Journal of the American Statistical Association 91*, 408–428.

Grizzle, J. E., C. F. Starmer, and G. G. Koch (1969). Analysis of categorical data by linear models. *Biometrics 24*, 489–504.

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

Haberman, S. J. (1972). Log-linear fit for contingency tables - Algorithm AS51. *Applied Statistics 21*, 218–225.

Haberman, S. J. (1973). Log-linear models for frequency data: Sufficient statistics and likelihood equations. *Annals of Statistics 1*(4), 617–632.

Haberman, S. J. (1976). Correction to as 51: Log-linear fit for contingency tables. *Applied Statistics 25*(2), 193.

Haberman, S. J. (1977). Log-linear models and frequency tables with small expected counts. *Annals of Statistics 5*(6), 1148–1169.

Hastie, T. J. and D. Pregibon (1991). Generalized linear models. In J. M. Chambers and T. J. Hastie (Eds.), *Statistical Models in S*. Pacific Grove, CA: Wadsworth & Brooks/Cole.

Heyde, C. C. and E. Seneta (1977). *I. J. Bienaymé : Statistical Theory Anticipated*. New York: Springer-Verlag.

Koch, G., J. Amara, S. Atkinson, and W. Stanish (1983). Overview of categorical analysis methods. *SAS–SUGI 8*, 785–795.

Koehler, K. J. (1986). Goodness-of-fit tests for log-linear models in sparse contingency tables. *Journal of the Americal Statistical Association 81*(394), 483–493.

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

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypothesis* (3rd ed.). Springer-Verlag.

Mantel, N. and W. Haenszel (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute 22*, 719–748.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. New York: Chapman and Hall.

Morris, C. (1975). Central limit theorems for multinomial sums. *Annals of Statistics 3*(1), 165–188.

Neyman, J. (1949). Contributions to the theory of the χ^2 test. In *Proceedings of the Berkeley Symposium on Mathematical Statistics*, pp. 239–273. Berkeley: University of California Press.

Norton, H, W. (1945). Calculation of chi-square for complex contingency tables. *Journal of the American Statistical Association 40*(230), 251–258.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

Pearson, K. (1900). On the criterion that a given system of deviation from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philosophical Magazine 50*, 59–73.

Pearson, K. (1904). Mathematical contributions to the theory of evolution, XIII: On the theory of contingency and its relation to association and normal correlation. *Draper's Co. Research Memoirs in Biometry Ser. I*, 1–35.

Pearson, K. and D. Heron (1913). On theories of association. *Biometrika 9*, 159–315.

Pistone, G., E. Riccomagno, and W. P. Wynn (2000). *Algebraic Statistics: Computational Commutative Algebra in Statistics*. New York: Chapman & Hall/CRC.

Quetelet, M. (1849). *Letters Addressed to H. R. H. the Grand Duke of Saxe Coburg and Gotha on the Theory of Probabilities as Applied to the Moral and Political Sciences (translated from the French by Olinthus Gregory Downs)*. London: Charles and Edwin Layton.

Rinaldo, A. (2005). *Maximum Likelihood Estimates in Large Sparse Contingency Tables*. Ph. D. thesis, Department of Statistics, Carnegie Mellon University.

Roy, S. N. (1957). *Some Aspects of Multivariate Analysis*. New York: Wiley.

Roy, S. N. and M. A. Kastenbaum (1956). On the hypothesis of no "interaction" in a multi-way contingency table. *Annals of Mathematical Statistics 27*(3), 749–757.

Roy, S. N. and S. K. Mitra (1956). An introduction to some nonparametric generalizations of analysis of variance and multivariate analysis. *Biometrika 43*, 361–376.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society 43*(3), 310–313.

Spirtes, P., C. Glymour, and R. Scheines (2001). *Causation, Prediction, and Search* (2nd ed.). Cambridge, MA: MIT Press.

Stigler, S. (2002). The missing early history of contingency tables. *Annales de la Faculté des Sciences de Toulouse 11*(4), 563–573.

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

Wilks, S. S. (1935). The liklihood test of independence in contingency tables. *Annals of Mathematical Statistics 6*, 190–196.

Yule, G. U. (1900). On the association of attributes in statistics: with illustration from the material of the childhood society, &c. *Philosophical Transaction of the Royal Society 194*, 257–319.