Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 137 (2007) 3430–3445
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation

Stephen E. Fienberg[a, b, *], Alessandro Rinaldo[a]

[a] *Department of Statistics, Carnegie Mellon University, USA*
[b] *Machine Learning Department, and Cylab, Carnegie Mellon University, USA*

Available online 31 March 2007

Abstract

The common view of the history of contingency tables is that it begins in 1900 with the work of Pearson and Yule, but in fact 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. Roy played a crucial role in this development with two papers co-authored with his students, 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.

Keywords: Algebraic statistics; Contingency tables; Chi-square tests; Log-linear models; Maximum likelihood; Multinomial sampling schemes; Sampling zeros

1. Introduction

Most papers and statistical textbooks on categorical data analysis trace the history back to the work of Pearson and 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 to the introduction by Galton (1892) of expected values of the form

$$\text{Expected count}(i, j) = \frac{(\text{Row marginal total } i) \times (\text{Column marginal total } j)}{\text{Grand total}}, \tag{1}$$

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 19th, the 20th, and the21st, 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,

* Corresponding author. Department of Statistics, Carnegie Mellon University, USA.
E-mail addresses: fienberg@stat.cmu.edu (S.E. Fienberg), arinaldo@stat.cmu.edu (A. Rinaldo).

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.03.022

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. Roy played a crucial role in this development with two papers co-authored with his students, Mitra and 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 very recently, namely maximum likelihood estimation in the presence of sampling zeros. In Sections 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) 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 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 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 d.f. 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.