Chapter 10
Further Topics

Abstract This epilogue chapter refers briefly to alternative methods and approaches in the analysis of contingency tables (latent class models, graphical models, and smoothing), not covered in the book. Furthermore, a bibliography on small sample inference, Bayesian inference, and the analysis of high-dimensional sparse contingency tables is discussed.

Keywords Association measures • Latent class models • Graphical models • Small sample inference • Bayesian inference

10.1 Overview

The focus of this book is on model-based approaches, so we did not refer to association measures other than the odds ratio which was in the kernel of the models developed. Measures of association however played an important role in the early development of categorical data analysis and continue to be of special interest in areas of social sciences and psychology. For this, an overview of the related literature is provided in Sect. 10.2 below.

The predominant modeling approach for contingency tables is log-linear models based. Alternative approaches of contingency table analysis with links to log-linear models are mentioned in Sect. 10.3.

Our approach was the classical frequentist approach, assuming large samples and non-sparse situations so that standard asymptotic theory applies. It is often the case that small samples occur. Sect. 10.4 refers to methods for analysis of small samples. Furthermore the Bayesian analysis of contingency tables is an attractive alternative in situations the asymptotic assumptions are not met. Beyond small sample setups, the Bayesian method is interesting for giving the option of incorporating prior information upon availability. Section 10.5 is devoted to Bayesian methods and applications for contingency tables.

M. Kateri, *Contingency Table Analysis: Methods and Implementation Using R*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4811-4_10,

Finally, clustering of categorical data has gained the last years renewed interest due to huge data sets and the need to organize their presentation and detect association structures. Huge data sets are normally extremely high dimensional. Bibliography on analyzing extreme high-dimensional categorical data sets and clustering techniques is to be found in Sect. 10.6.

10.2 On Measures of Association

Upon rejection of independence, information on the significant underlying relationship is traditionally summarized through *measures of association*. These measures differentiate for nominal and ordinal data. Measures for nominal variables refer only to the strength of the association while for ordinal variables, they incorporate information about the direction (positive or negative) of the association as well. Their values range preferably in the [0, 1] and [−1, 1] intervals, for the nominal and ordinal measures, respectively, with their absolute value being increasing in the strength of the association. Special interest has been attracted by measures for binary cross-classifications.

The interest in defining and measuring association in $I \times J$ contingency tables dates back to the 1840s and the related bibliography is enormously rich, motivated from diverse scientific fields, like social sciences, education, and meteorology. The history of their development, interesting and often generating controversies, is reviewed exhaustively in the book by Goodman and Kruskal (1979), which is the most classical reference on association measures. This book republishes their four *JASA* papers of Goodman and Kruskal in 1954, 1959, 1963, and 1972. In the first two papers, Goodman and Kruskal organized existing measures, presenting them unified. Furthermore, they focused on the general $I \times J$ table and developed new measures, for nominal and ordinal variables. They suggested measures taking into account the existence of a response variable in the table, introducing symmetric and asymmetric versions of their measures. In their last two papers they proved the asymptotic normality of their measures, making thus asymptotic inference feasible.

Association measures are organized in classes accordingly to their basis of formulation. They can be based on the X^2 statistic for independence, on the assumption of an underlying joint normal distribution, or they can be based on a probabilistic model (like measures based on pairs of observations or on scores assigned to the categories of the classification variables). The measures based on the X^2 statistic, as the contingency coefficient ϕ, the association coefficient C, and Cramer's V, are indicative only about the strength of the underlying association.

Focusing on 2×2 contingency tables, the odds ratio θ, $\theta \in [0, +\infty)$, introduced by Yule (1900, 1903, 1912), is undoubtedly the "gold standard" of measures. Yule proposed transformations of θ, the Q and Y, that range in [−1, 1]. A competitor of Yule's odds ratio for measuring association in 2×2 tables is the *tetrachoric correlation coefficient* of Pearson (1900b, 1904, 1913), a product–moment correlation between two unobserved quantitative variables that have been dichotomized.

The tetrachoric correlation opposes the odds ratio by this underlying continuum assumption. Pearson preferred to view contingency tables as discretizations of underlying multivariate normal distributions (Pearson and Heron 1913) while for Yule contingency tables were formed by discrete variables of fixed categories. The dispute between Yule and Pearson was long and strong. Over the years both approaches had their supporters and opposers. Edwards (1963) argued that the odds ratio should be the basis of an association measure for 2×2 tables. On the other hand, when the classification variables are of continuous nature and dichotomized at a certain cut point, as it is often the case in psychometrics, then the tetrachoric correlation is preferable. The tetrachoric correlation coefficient and related inference problems are discussed analytically by Bonett and Price (2005). Transformations of the odds ratio that approximate the tetrachoric correlation have been proposed by Digby (1983) and Becker and Clogg (1988). Bonett and Price (2007) proposed a generalized Yule coefficient that is similar in value to the contingency coefficient ϕ and has, among others, Yule's Q and the coefficient of Digby (1983) as special cases. A review of association coefficients for 2×2 tables with focus on their general properties is provided by Warrens (2008). Overall, the odds ratio predominated, also due to its connection to log-linear models. The odds ratio is the precursor of log-linear models and Yule can be considered as their "father." For an overview of the odds ratios and their role in contingency table analysis, see Rudas (1998).

For $I \times J$ contingency tables, the tetrachoric correlation was extended to the *polychoric correlation* by Lancaster and Hamdan (1964) while measures, generalizations of the odds ratio, were proposed by Altham (1970) for nominal association and by Agresti (1980) and Edwardes and Baltzan (2000) for ordinal. Cumulative odds ratios for ordinal variables have been considered earlier by Clayton (1974), who developed statistics based on them to summarize the difference in location between two distributions of an ordered categorical variable and for describing association between two such variables. The approach was generalized also in case some observations were subject to censorship (Clayton 1976).

For ordinal association, famous are Goodman and Kruskal's λ, λ_a, λ_b and γ. Kendall (1938, 1948) introduced a measure of rank correlation, known as Kendall's τ, and its asymmetric version τ_b. Stuart (1953) proposed a measure of association for contingency tables, the τ_c, based on Kendall's τ, which is compared to Goodman and Kruskal's γ in Hamdan (1977). Another popular measure of ordinal association is Somers' d (Somers 1962). The measures of *raters' agreement* consist a special category of association measures, presented in Sect. 9.5.2.

Different measures refer to different types of association; thus they should not be used blindly and routinely. The choice of measure should take into account the nature of the contingency table under consideration and be compatible with its distribution. Efficacies of the measures of association for ordinal contingency tables are discussed in Simon (1978). Contingency tables are better treated through model-based approaches, which are flexible regarding the structure imposed on the association and meanwhile are more informative, providing cell estimates under the assumed model. Since our approach is model focused, we will not discuss any further measures of association. For ordinal measures of association we refer to Agresti (2010, Chap. 7).

10.3 Alternative Approaches in Contingency Table Analysis

10.3.1 Latent Class Models

The classical latent class model is defined as a finite mixture of unobserved (latent) multinomial distributions, each of which exhibits statistical independence. Latent class models play an important role in multivariate data analysis and receive special attention in psychology and social sciences. They were first treated by Lazarsfeld (1950). For their connection to log-linear models, see Goodman (1974), Haberman (1979), Heinen (1996), and Hagenaars (1998). Formann (1992) connected latent class models to polytomous logistic models and Gilula (1979, 1984) to correspondence analysis. Goodman (1987) provides a nice overview of the connection between the approaches of CA, latent class analysis, and log-linear and association models. The connection of association models to latent class models is also discussed in Anderson and Vermunt (2000). Their link to models for rater agreement is due to Uebersax (1993), Uebersax and Grove (1990, 1993), and Yang and Becker (1997). Becker and Yang (1998) discuss latent class models for modeling marginal associations in contingency tables and provide an extended literature review on analysis of contingency tables by latent class models.

The volume *Applied Latent Class Analysis*, edited by Hagenaars and McCutcheon (Cambridge University Press, 2002), provides an interesting collection of papers on traditional latent class analysis as well as in connection to special topics as clustering, logistic regression, longitudinal data, missing data, and nonresponse (see, e.g., Goodman 2002b). For an updated source on latent variable models and their applications, we refer to Bartholomew et al. (2011). An overview on latent variable models for categorical responses is provided by Agresti and Kateri (2014).

10.3.2 Graphical Models

The graphical log-linear models and their role in the analysis of high-dimensional contingency tables are discussed in Sects. 4.7.2, 4.9.3, and 4.9.4. Beyond log-linear models, conditional independence graphs of a multi-way contingency tables are connected to the RC model and to correspondence analysis by de Falguerolles et al. (1995). Bounds on the cell counts of contingency tables for decomposable log-linear models and their related graphs are proposed by Dobra and Fienberg (2000, 2003). Bidirected graphical models of marginal independencies for categorical data are discussed in Lupparelli et al. (2009). Quasi-symmetry models are presented as graphical models by Gottard et al. (2011).

10.3.3 Smoothing Categorical Data

Alternative to modeling, nonparametric approaches can be adopted for analyzing contingency tables. The oldest and most well-known is the partitioning of the X^2 statistic for testing independence (see Sect. 2.5.4). Furthermore, smoothing methods bridge the gap between the parametric and nonparametric approaches, from strict assumptions to no assumptions at all. Smoothing methods for ordinal data were studied and compared by Titterington and Bowman (1985), who organized them into three major groups, the kernel-based methods (Aitchison and Aitken 1976), Bayesian-based methods (Leonard 1975), and the penalized minimum distance methods, which relate to maximizing the penalized likelihood (Scott et al. 1980). Smoothing methods are appropriate for the analysis of large sparse contingency tables (Simonoff 1983). The application of smoothing to categorical data analysis is reviewed in Simonoff (1995, 1998). For a literature review on the smoothing methods for categorical data, see the references cited in Titterington and Bowman (1985) and Simonoff (1995).

Coull and Agresti (2003) introduced a generalized log-linear model with random effects, useful for smoothing large sparse contingency tables by maximizing a penalized likelihood. The structures considered for the random effects mimic Goodman's association models. Geenens and Simar (2010) propose two nonparametric tests for testing independence of two categorical cross-classified variables, conditional on a set of explanatory variables, based on kernel estimation of the conditional probabilities.

10.4 Small Sample Inference for Contingency Tables

Starting with Fisher's exact test (see Sect. 2.1.7) for testing independence for 2×2 contingency tables with small cell entries, the development of methods and algorithms for exact inference for a variety of models, applied on two- and multi-way contingency tables, has a long history. In the early years, Yule's correction was quite popular, while it was criticized later. Despite this, adding a constant to the cells of a contingency table to avoid problems from small or zero count cells is still common in practice. Greenland (2010) argued against this practice, showing in the Bayesian framework that it can lead to a form of Simpson's paradox, and proposed more sophisticated methods of smoothing. In case of high-dimensional contingency tables, sparseness problems occur even for moderate to large sample sizes, leading to inferential implications (see also Sect. 10.6).

Fundamental is the network algorithm of Mehta and Patel (1983) for sampling from an $I \times J$ contingency table with given marginals, which served in extending the Fisher's exact test to tables of higher size. Sequent results by them with coauthors established the exact analysis of contingency tables. For example, Agresti et al. (1990) extended this algorithm for the exact analysis of two-way contingency

tables with ordinal classification variables. Basic summarizing reference is Mehta and Patel (1995). Exact analysis of models for binary response is provided in Cox (1970a) and Cox and Snell (1989). Exact conditional tests for testing quasi-independence in incomplete tables are considered by McDonald and Smith (1995). A survey on exact inference for categorical data is provided in the discussion paper of Agresti (1992) and in Agresti (2001). A detailed treatment is provided in the book by Hirji (2006).

Small sample inference can be developed based on Markov chain, as in Forster et al. (1996), or bootstrap algorithms. For applications of bootstrap methods on categorical data, we refer to Jhun and Jeong (2000) and Amiri and von Rosen (2011). Model-based bootstrap tests for independence in two-way tables are considered in Pettersson (2002) and Jeong et al. (2005). Bootstrapping for log-linear models in large, sparse contingency tables has been considered by Sauermann (1989). Alternatively to log-linear models, Streitberg (1999) developed a bootstrap approach for analyzing interactions in high-dimensional tables, based on the additive approach (see Darroch and Speed (1983) and references therein).

The problem of studying the exact distribution in a contingency table under a model assumption can also be faced through algebraic statistics, based on the pioneer work by Diaconis and Sturmfels (1998). They proposed an algorithm for sampling from a set of tables with given marginals, based on Markov bases computation, which is achieved by finding a Gröbner basis. Aoki and Takemura (2005) and Rapallo (2003, 2006) derive Gröbner bases for some classical log-linear models taking structural zeros into account. Dobra (2003) applied graphical models to identify special settings which lead to reduction of the required computations for the identification of a Markov basis. Dobra et al. (2009) dealt with the maximum likelihood estimation for log-linear models and a related disclosure limitation problem, focusing on the disclosure of small cell counts to protect the confidentiality of individual responses. Hara et al. (2012) proposed a new class of models for the analysis of multi-way contingency tables, more parsimonious than the usual hierarchical log-linear models, by modeling the interaction terms in each maximal compact component of a hierarchical model. They proceed to exact tests via Markov bases while their approach considers also the presence of structural zeros.

Exact inference for the model of symmetry for square contingency tables based on Diaconis and Sturmfels's algorithm is provided by Rapallo (2003) and Krampe and Kuhnt (2007). In the context of rater agreement, Rapallo (2005) provides algebraic testing procedures for Cohen's kappa, the quasi symmetry, and quasi-independence model. Krampe et al. (2011) develop algebraic tests for the models of conditional, diagonal, and ordinal quasi symmetry.

Alternatively, inferential problems due to sparseness, sampling zeros, or small frequencies can be treated in the Bayesian analysis framework.

10.5 Bayesian Analysis of Contingency Tables

Bayesian analysis can be the solution in situations of small samples or sparse tables, where standard asymptotic inference does not apply. Furthermore, the incorporation of prior inference can be essential in some applications' areas. The model selection procedure is benefited in the Bayesian framework. MCMC methods enable an efficient search of the model space even if it is large. For the models visited, the associated algorithm provides the posterior model probabilities, a powerful tool for models' evaluation and estimation of their uncertainty. These issues are of great importance in high multidimensional contingency tables. Model uncertainty might be high in small samples or in existence of more than one models of similar performance. High model uncertainty can be incorporated in the Bayesian statistical inference.

Early attempts on Bayesian analysis of categorical data go back to the 1950s and were based on conjugate prior analysis. Good (1956) proposed smoothing proportions in contingency tables while his approach for hierarchical Bayesian inference (Good 1965) is related to the early work by Johnson in the 1920s on the Dirichlet priors for the multinomial distribution (see Fienberg (2006) for a detailed discussion on the early and key Bayesian developments). Lindley (1964) focused on the Bayesian analysis of contingency tables (two- and three-way) and developed the Bayesian inference for the odds ratio. Altham (1969, 1971) dealt with the Bayesian analysis of 2×2 tables for small samples based on conjugate priors. Since then, the development of the Bayesian approach was rapid, mainly due to the progress of computer-intensive numerical methods for the evaluation of posterior distributions, which made the Bayesian analysis and Bayesian model selection of multidimensional problems and complex models feasible. For an overview, see Congdon (2005) and the review paper by Agresti and Hitchcock (2005).

The Bayesian analysis of log-linear models with non-conjugate priors originates from Leonard (1975) and Laird (1978). They proposed univariate normal priors for the parameters of the saturated model. Knuiman and Speed (1988) and King and Brooks (2001) considered multivariate normal prior for the parameter vector and extended the approach to multi-way contingency tables. In contingency tables' framework, the model fit evaluation through the Bayes factor has been considered by Spiegelhalter and Smith (1982), Raftery (1986), and Albert (1997). Issues for the Bayesian analysis of the 2×2 table are discussed in Howard (1998). For Bayesian log-linear model selection we refer to Dellaportas and Forster (1999) and Ntzoufras et al. (2000). The Bayesian analysis of log-linear models is reviewed in Forster (2010). Consonni and Pistone (2007) considered the Bayesian analysis of contingency tables with structural zeros based on algebraic statistics. The basic reference on Bayesian logit models is Albert and Chib (1993).

The Bayesian analysis of the simple U association model has been considered by Agresti and Chuang (1989). They imposed a Dirichlet prior distribution on the probability table $\pi = (\pi_{ij})_{I \times J}$. The prior mean was assigned from the U model. Alternatively to the conjugate prior-type analysis they proposed the Bayesian

log-linear analysis by considering independent uniform priors for the main effect parameters λ_i^X and λ_j^Y and normal priors for the interaction parameters $\lambda_{ij}^{XY} \sim N(\varphi \propto_i \nu_j, \sigma^2)$.

The first attempt for fitting the RC association model in the Bayesian framework was due to Chuang (1982). He set independent uniform priors on the main effect parameters λ_i^X and λ_j^Y and normal priors on the parametric row and column scores $\propto_i \sim \mathcal{N}(0, \sigma_1^2)$, $\nu_j \sim \mathcal{N}(0, \sigma_2^2)$ and proceeded with empirical variance estimation. Evans et al. (1993) adopted a different approach for the Bayesian analysis of the RC model. They based their analysis on the Bayesian estimation of the saturated log-linear model with normal priors on all its parameters and then concluded to the posterior for the RC by Euclidean projection from the posterior of the saturated log-linear model. Further, they studied the posterior distribution of the Euclidean distance between the interaction matrices of the saturated and the RC models, (λ_{ij}^{XY}) and $(\varphi \propto_i \nu_j)$, respectively. Finally, Bayesian inference for the more general RC(M) association model has been developed by Kateri et al. (2005). This procedure can be also applied for fitting the RC model (for $M = 1$). Albert (1997) provided an interesting Bayesian approach for testing the fit of simple models such as independence, quasi-independence, and uniform association models, as well as for modeling outliers via mixture models.

With respect to merging categories and the associated role of the association models (as discussed in Sect. 7.5), in the Bayesian framework, Tarantola et al. (2008) used methodology adopted from product partition models to make inferences about the clustering of scores in the row effect model. For the two-group comparison of an ordinal scale, Kateri and Agresti (2013) discussed stochastic orderings, based on generalized odds ratios for ordinal responses for $2 \times J$ contingency tables, from the Bayesian point of view.

We referred in Sect. 6.8.2 to the order-restricted inference for association models, through large sample asymptotic methods. The Bayesian inference for association models with order-constrained parametric scores has been developed by Iliopoulos et al. (2007). Their approach for identifying possible score equalities was based on calculating the posterior probabilities of possible order violations for successive categories in the unrestricted model. These probabilities were used in an isotonic regression-type logic, indicating which scores should be merged. Furthermore, the deviance information criterion (DIC, Spiegelhalter et al. 2002) was applied to identify the most appropriate model in terms of goodness of fit. However, this approach forms not a formal Bayesian evaluation in favor or against merging specific scores, since it is not based on the posterior model odds and probabilities (for details, see Kass and Raftery 1995). Toward this direction, Iliopoulos et al. (2009) proposed an alternative approach for this problem, focusing on the estimation of posterior model probabilities of the RC order-constrained model, in a full Bayesian way, by allowing for ties in the prior distribution level. They constructed a *trans*-dimensional MCMC algorithm (reversible jump MCMC, Green 1995) for assessing the equality of successive row and column scores.

For Bayesian graphical models we refer to Madigan and Raftery (1994) and Madigan and York (1995). Massam et al. (2009) developed a family of conjugate prior for a class of discrete hierarchical log-linear models for multi-way tables, with graphical models being in this class. Webb and Forster (2008) dealt with Bayesian graphical model selection for multivariate ordinal data. Ng et al. (2008) provided a conjugate Bayesian analysis of incomplete contingency tables based on a new family of distributions, the grouped Dirichlet distributions, which includes the classical Dirichlet distribution as special case.

10.6 Extreme High-Dimensional Categorical Data

High-dimensional contingency tables often lead to sparseness and related inferential discrepancies. Approaches for high-dimensional data discussed in Hastie et al. (2009) and Bühlmann and van de Geer (2011) apply also on contingency tables. In particular, Dahinden et al. (2007) extended the lasso algorithm, to the group lasso, in order to fit log-linear models for high-dimensional and sparse data arising in computational biology. An alternative approach based on graphical models is given by Dahinden et al. (2010). The lasso penalty for high-dimensional GLMs is considered by van de Geer (2008).

In high-dimensional problems the clustering of the subjects or items under study becomes often an important issue. This way customers, patients, or genes, for example, can be assigned to groups of similar profile (with respect to some characteristics). Clustering methods are based on measuring the dissimilarity between items with respect to their characteristics, captured in variables. Most of the clustering algorithms refer to continuous variables (see, e.g., Everitt et al. 2011). Bock (1986) developed clustering methods for categorical data, based on a logistic or log-linear models probability distribution. For an overview on clustering methods that apply also on categorical data, see also in van Mechelen et al. (2004). On clustering of categorical data, refer to Agresti (2013, Sect. 15.3).