Book Review

Reviewed by Zhongxian Wang and James Yao*

E-mail: wangj@mail.montclair.edu
E-mail: yaoj@mail.montclair.edu
*Corresponding author

Categorical Data Analysis using SAS (Third Edition)
by: Maura E. Stokes, Charles S. Davis and Gary G. Koch
Published 2012
by SAS Institute Inc.
100 SAS Campus Drive, Cary, NC 27513, USA, 580pp
ISSN: 978-1-60764-664-8

Qualitative variables make it impossible to carry out usual inferences. More often than not, we need to use non-parametric methods in order to make direct inferences about two or more distributions, either by asking if a population distribution has some particular specifiable forms or if two or more population distributions are identical (David and Averbuch, 2012; Chen et al., 2012). There are many current books on categorical data analysis (Agresti, 2012). What we would like to have for categorical data analysis are techniques that give us useful information like what we get from correlation and regression analyses for continuous data (Ferrari et al., 2011).

This book is exceptionally unique since it contains both solid theoretical foundations and practical guidance. It discusses hypothesis testing strategies for the assessment of association in contingency tables and sets of contingency tables. It also discusses various modelling strategies available for describing the nature of the association between a categorical response measure and a set of explanatory variables.

Chapter 1 describes various scales and illustrates them with datasets used in later chapters. It introduces various analysis strategies discussed in this book and describes how they relate to one another. It also discusses target populations generally assumed for each type of analysis and what types of inferences one can make to them. After reviewing how SAS software handles contingency tables and other forms of categorical data, the authors provide a guidance to the material in the book for various types of readers.

Chapter 2 introduces a 2×2 contingency table, one of the most common ways to summarise categorical data. There are several ways of testing whether there is a statistical association between row variable and column variable; many of the ways are based on chi-square statistic. In addition to examining the difference in proportions, the authors provide McNemar's test, sensitivity specificity, and computing incidence density ratios. Instead of association in just one table, Chapter 3 investigates overall association in sets of tables with similar strategies involving chi-square statistics and measures of association such as odds ratios.

Chapter 4 focuses on tables and sets of tables having other dimensions that also occur frequently. For 2 × r tables, there is interest in investigating a response variable with multiple ordered outcomes for a single table or for a combined set of strata. For r × 2 tables, there is interest in the trend of proportions across ordered groups for a single table or for a combined set of strata.

Chapter 5 deals with s × r tables. What differentiates Chapter 5 from earlier chapters is that the scale of measurement is always a consideration; the statistics one chooses depends on whether the rows and columns of the table are nominally or ordinally scaled. This is true for investigating whether association exists and for summarising the degree of association. Also, both tests for association and measures of association are addressed. *Observer agreement* studies how closely different researchers' evaluations agree. The Jonckheere-Terpstra test can be used for checking ordered differences.

Chapter 6 explores the formulation of Mantel-Haenszel statistics in matrix terminology. It illustrates the use of Mantel-Haenszel strategy for several applications. Finally, the use of Mantel-Haenszel strategy for repeated measurement analysis, an advanced topic, is demonstrated. Many of the commonly used non-parametric tests, such as Krukal-Wallis, Spearman correlation, and Friedman tests, can be computed using the Mantel-Haenszel procedures. While previous chapters have shown how to use Mantel-Haenszel procedures to analyse two-way tables and sets of two-way tables, Chapter 7 shows how to use the same procedures to preform non-parametric analyses of continuous response variables.

Chapters 8 and 9 concentrate mostly on asymptotic methods that require adequate sample size in order for model fit and effect assessment tests to be valid. However, sometimes datasets are so spares or have such small cell counts that these methods are not valid. LOGISTIC procedure is designed primarily for logistic regression analysis, and it provides useful information such as odds ratio estimates and model diagnostics. Chapter 8 also discusses exact logistic regression, which is an alternative strategy of these situations.

Chapter 9 asserts that logistic regression most often involves modelling a dichotomous outcome, but it also applies to multilevel either ordinal or nominal responses. We can model functions called *cumulative logits* by performing ordered logistic regression using the proportional odds model for ordinal response outcomes. On the other hand, for nominal response outcomes, we form *generalised logits* and preform a logistic analysis similar to those described in the previous chapter, except that we model multiple logits per subpopulation.

The usual maximum likelihood approach to estimation in logistic regression is not always appropriate. The appropriate form of logistic regression for these types of data is called *conditional logistic regression*. The use of conditional logistic regression for matched studies in epidemiological work is discussed and illustrated with two examples in Chapter 10. Finally, the use of exact logistic regression for the stratified setting is discussed with several examples.

Quantal response data analysis deals with subject response to a stimulus that occurs with great and greater density. Chapter 11 is concerned with quantal responses, which are analysed with categorical data analysis strategies. This chapter provides examples of applying quantal response data analysis techniques to drug development and growth studies.

Analysis of generalised logits is a form of log-linear model, discussed in Chapter 12. Log-linear model methodology is most appropriate when there is no clear distinction

between response and explanatory variables. It treats all variables as response variables, and the focus is on statistical independence and dependence. Log-linear modelling of categorical data is analogous to correlation analysis for normally distributed response variables and is useful in assessing patterns of statistical dependence among subset variables.

Categorical data often are generated from studies that have time from treatment or exposure until some events as their outcome. Frequently, interest lies in the computing of survival rates. Chapter 13 shows life table methods for computing these results and compares survival rates for treatment groups which determines whether there is a treatment effect with Mantel-Cox test. In addition to hypotheses testing, one may be interested in describing the variation in survival rates.

Weighted least squares (WLS) estimation provides a methodology for modelling a wide range of categorical data outcomes. Chapter 14 covers the application of WLS for modelling of mean scores and proportions in the stratified simple random sampling framework, as well as for the modelling of estimates produced by more complex sampling mechanisms, such as those required for complex sample surveys. The methodology is explained in the context of a basic example.

The generalised estimating equation (GEE) approach is an extension of generalised linear models that provides a semi-parametric approach to longitudinal data analysis with univariate outcomes for which the quasi-likelihood formulation is sensible. This approach encompasses a broad range of data situations, including missing observations, continuous explanatory variables, and time-dependent explanatory variables. In Chapter 15, GEE approach for the analysis of repeated measurements is discussed and illustrated with a series of examples.

References

Agresti, A. (2012) *Categorical Data Analysis*, 3rd ed., Wiley & Sons Inc., Hoboken, New Jersey.

Chen, T., Zhang, L.N., Liu, T., Poon, M.K. and Wang, Y. (2012) 'Model-based multidimensional clustering of categorical data', *Artificial Intelligence*, Vol. 176, No. 1, pp.2246–2269.

David, G. and Averbuch, A. (2012) 'SpectralCAT: categorical spectral clustering of numerical and nominal data', *Pattern Recognition*, Vol. 45, No. 1, pp.416–433.

Ferrari, A.P., Annoni, P., Barbiero, A. and Manzi, G. (2011) 'An imputation method for categorical variables with application to nonlinear principal component analysis', *Computational Statistics & Data Analysis*, Vol. 55, No. 7, pp.2410–2420.