Visualizing Categorical Data with R

Book Plan and Outline

Michael Friendly
York University

David Meyer
Wirtschaftsuniversität Wien

with contributions,
Achim Zeilleis
Universität Insbruck

December 9, 2013

Overview and rationale for VCDR

Visualizing Categorical Data with R (VCDR) is the successor to my book *Visualizing Categorical Data* (2000) published by SAS Institute. In the interim, there has been much development in the analysis and visualization of categorical data, and the bulk of that has been implemented in R.

The 2000 edition of *Visualizing Categorical Data* (VCD) stemmed from the premise that, while graphical methods for quantitative data are well-developed in most statistical software and widely used in practice, corresponding graphical methods for categorical data—counts, frequencies and discrete variables—were still in relative infancy at that time.

VCD was designed to present an overview of the analysis of categorical data focused on the graphical methods designed for data exploration, model building, model diagnostics, etc., analogous to the graphical techniques that are now commonly used for quantitative data. That book, along with published journal articles (e.g., Friendly 1994, 1999; Emerson, Green, Schloerke, Crowley, Cook, Hofmann, and Wickham 2013) has been highly influential in statistical practice.

The special nature of discrete variables and frequency data vis-a-vis statistical graphics is now more widely accepted, and many of these methods (e.g., mosaic displays, fourfold plots, diagnostic plots for generalized linear models) have become, if not main stream, then at least more widely used in research and teaching. As well, VCD spurred the implementation of many of these methods in R (e.g., the **vcd** package), and there has been considerable growth in both statistical methods for the analysis of categorical data (e.g., generalized linear models, zero-inflation models, mixed models for hierarchical and longitudinal data with discrete outcomes), along with some new graphical methods for visualizing and interpreting the results (3D mosaic plots, effect plots, diagnostic plots, etc.)

Thus, the time is right for a thorough revision of the central organization and framework of VCD to a modern perspective, with a focus on R.

Features

- Provides an accessible introduction to the major methods of categorical data analysis for data exploration, statistical testing and statistical models.
- The emphasis throughout is on computing, visualizing, understanding and communicating the results of these analyses.
- As opposed to more theoretical books, the goal here is to help the reader to translate theory into practical application, by providing skills and software tools for carrying out these methods.
- Includes many examples using real data, often treated from several perspectives.
- Supported directly by R packages **vcd** and **vcdExtra**, along with numerous other R packages

- All materials will be available online; the design of the book will support simultaneous publication as an ebook, possibly with hyperlinks to references and related material in the text or elsewhere.
- As far as possible, each chapter will contain one or more lab exercises, which work through applications of some of the methods presented in that chapter. This will make the book more suitable for both self-study and classroom use.

Audience

This book assumes basic understanding of statistical concepts at least at an intermediate undergraduate level including regression and analysis of variance (for example, at the level of Neter, Wasserman, and Kutner (1990); Mendenhall and Sincich (2003)).

It is written to appeal to two audiences:

- Students and methodologists in the social and health sciences, epidemiology, economics, business and (bio)statistics
- Substantive researchers in various disciplines wanting to be able to apply these methods to their own data

It is also assumed that the reader has at least basic knowledge of the R language and environment, including interacting with the R console (RGui for Windows, R.app for Mac OS X) or other graphical user interface (e.g., RStudio), using R functions in packages, getting help for these from R, etc. One introductory chapter (Chapter 2) is devoted to covering those topics beyond such basic skills needed in the book.

The book will be written so that it can be used as a primary or secondary text in courses dealing with categorical data analysis at the upper undergraduate and graduate levels. For example, in Winter, 2015, I will teach a graduate course in the Quantitative Methods area in Psychology at York, where this book will serve as the main text. The most important markets would include most of the applied areas of statistics, but principally:

- Statistics: Biostatistics and Epidemiology
- Statistics: Statistics for the Social and Behavioral Sciences

Relation to other books

There are now quite a few modern texts covering categorical data analysis from varying perspectives. Among these, Agresti (2013), *Categorical Data Analysis* is probably among the most complete, but advanced treatments, and his earlier *An Introduction to Categorical Data Analysis* (Agresti 1996) remains an accessible introductory text at a lower level.

Powers & Xie 2008, *Statistical Methods for Categorical Data Analysis*, covers a wide variety of these topics and others (multilevel models for binary data, event history data) at an advanced, graduate level, with emphasis on social science data.

Simonoff (2003), *Analysing Categorical Data*, is somewhat similar, with a different range of topics and more geared to advanced students in statistics.

Christensen (1997), *Log-Linear Models and Logistic Regression*, is also a somewhat advanced-level book, with coverage largely restricted to the methods in the title.

Stokes, Davis, and Koch (2000), *Categorical Data Analysis with the SAS System*, covers most of the non-parametric and model building methods of analysis of categorical data, but the emphasis is almost entirely on presenting the underlying theory, using SAS software to perform the analysis, and on interpreting the results from the numerical output. There are only a handful of graphs in the entire book.

None of these books feature graphical methods for categorical data; in fact, most of these show very few data graphs. A few of these contain a brief appendix mentioning software, or have a related web site with some data sets and software examples. Moreover, none actually describe how to do these analyses and graphics with R.

Yet, for categorical data, just as for quantitative data, there are many aspects of the relationships among variables, the adequacy of a fitted model, and possibly unusual features of the data which can best (or in some cases, only) be seen and appreciated from an informative graphical display.

Thus, there is a clear need to present these modern methods for the graphical display and analysis of categorical data. The existing books present the opportunity to describe and illustrate the graphical approach without the necessity to discuss as much of the underlying theory as would be required otherwise, and VCDR will be written to complement, rather than compete with, them. So, where sufficient theoretical background already exists, I will present short summaries and point readers to the other sources. Material that is unique to VCDR (e.g., correspondence analysis, GLIMs, GEE, CART models, etc.) will be developed more fully. This strategy will also help keep the book more manageable in both size and writing time.

TEXnical issues and production details

The book will be written using the **knitr** package and other tools for writing, reporting and reproducible research in R. This allows the writing to mix LaTeX for the text, equations, index entries, etc. with R code that generates tables and graphs for examples, guaranteeing that all of the examples, tables and graphs in the book are reproducible and up-to-date. It also makes it relatively easy to selectively export the R code and data for examples and exercises to a form that readers can download and work with on their own, and produce alternative (e.g., ebook) versions.

Color: Another important consideration is the use of color in the book. By its nature, the book will include many graphs, and I plan to use color liberally, particularly where it is essential for communication of the ideas, methods, and understanding of the techniques described and illustrated. Ideally, an all-color book would be best, but cost/price considerations might lead to some compromise.

Pages: As a rough guess, I expect that the book would come to approximately 400–450 printed pages.

Preferred format: In terms of format, structure and integration of R content, one book (on a different topic) that is similar to what I have in mind is James, Witten, Hastie, and Tibshirani (2013), *An Introduction to Statistical Learning with Applications in R*. This book, in the Springer Texts in Statistics series, is published in **full color** (even in the text), and sells for $80 (hardcover), $60 (ebook), with $\sim$ 20% discounts on Amazon.

Another related book using R that I admire in terms of layout and rich use of color graphics is Gower, Lubbe, and Roux (2011), *Understanding Biplots*. Unfortunately, too much of the content of this book is devoted to documentation of the methods in their **UBbipl** package, that should have been relegated to the package itself

Outline

The provisional outline below is based on the structure of topics in VCD, updated with a new introductory chapter (Chapter 2) and three new substantive topics in Chapters, 10–12. For each chapter, I give an overview of the content and a list of sections; some of the details given here will certainly change as writing progresses.

1 Chapter 1: Introduction

"Categorical data" means different things in different contexts. I introduce the topic with some examples illustrating (a) types of categorical variables: binary, nominal, and ordinal, and (b) the main types of categorical data: counted data and frequency data.

Methods for the analysis of categorical data also fall into two quite different categories, described and illustrated next: the simple non-parametric, and randomization-based methods typified by the classical Pearson χ^2, Fisher's exact test, and Mantel-Haenszel tests, and the model-based methods represented by logistic regression and generalized linear models. Chapters 3–6 are mostly related to the non-parametric methods, Chapters 7–12 to the model-based methods.

Finally, I describe some important differences between categorical data and quantitative data, discuss the implications of these differences for visualization techniques, and outline a strategy of data analysis focussed on visualization.

1.1. Data visualization and categorical data
1.2. What is categorical data?
1.3. Strategies for categorical data analysis
1.4. Graphical methods for categorical data
1.5. Visualization = Graphing + Fitting + Graphing

2 Chapter 2: Working with categorical data

Categorical data can be represented in various forms: case form, frequency form, and table form. This chapter describes and illustrates the skills and techniques in R needed to input, create and manipulate R data objects to represent categorical data, and convert these from one form to another for the purposes of statistical analysis and visualization which are the subject of the remainder of the book.

2.1. Forms of categorical data: case form, frequency form and table form
2.2. Ordered factors and reordered tables
2.3. Generating tables with `table()` and `xtabs()`
2.4. Printing tables with `structable()` and `ftable()`
2.5. Collapsing over table factors: `aggregate()`, `margin.table()` and `apply()`
2.6. Converting among frequency tables and data frames
2.7. A complex example
2.8. Lab exercises

3 Chapter 3: Fitting and graphing discrete distributions

Discrete frequency distributions often involve counts of occurrences, such as accident fatalities, words in passages of text, or blood cells with some characteristic. Often interest is focussed on how closely such data follow a particular probabiliy distribution, such as the Poisson, binomial, or geometric distribution. Understanding and visualizing such distributions in the simplest case of an unstructured sample provides a building block for generalized linear models where they serve as one component.

This chapter describes the well-known discrete frequency distributions: the binomial, Poisson, negative binomial, geometric, and logarithmic series distributions in the simplest case of an unstructured sample. The chapter begins with simple graphical displays (line graphs and histograms) to view the distributions of empirical data and theoretical frequencies from a specified discrete distribution.

It then describes methods for fitting data to a distribution of a given form and simple, effective graphical methods than can be used used to visualize goodness of fit, to diagnose an appropriate model (e.g., does a given data set follow the Poisson or negative binomial?) and determine the impact of individual observations on estimated parameters.

3.1. Introduction to discrete distributions
3.2. Plotting discrete distributions
3.3. Fitting discrete distributions
3.4. Diagnosing discrete distributions: Ord plots
3.5. Poissonness plots and generalzed distribution plots
3.6. Chapter summary
3.7. Further reading
3.8. Lab exercises

4 Chapter 4: Two-way contingency tables

This chapter begins with an overview of statistical tests for association in two-way frequency tables and extensions of these tests for the case of multi-way tables, where two primary variables are stratified by one or more others.

Several schemes for representing contingency tables graphically are based on the fact that when the row and column variables are independent, the estimated expected frequencies, e_{ij}, are products of the row and column totals (divided by the grand total). Then, each cell can be represented by a rectangle whose area shows the cell frequency, f_{ij}, or deviation from independence.

This chapter describes a number of relatively simple visualization techniques based on this relation (Sieve diagram, Association plot), and several more specialized techniques for particular data structures.

4.1. Introduction
4.2. Tests of association for two-way tables
4.3. Stratified analysis
4.4. Fourfold display for 2 x 2 tables
4.5. Sieve diagrams
4.6. Association plots
4.7. Observer agreement
4.8. Trilinear plots
4.9. Chapter summary
4.10. Further reading
4.11. Lab exercises

5 Chapter 5: Mosaic displays for n-way tables

When there are more than two classification variables, the visualization of categorical data becomes increasingly difficult. This chapter extends the use of the fourfold display to a collection of 2×2 tables, and introduces the mosaic display.

The mosaic display, proposed by Hartigan & Kleiner 1981 and extended by Friendly (1994, 1999), represents the counts in a contingency table directly by tiles whose size is proportional to the cell frequency. One important design goal is that this display should apply extend naturally to three-way and higher-way tables. Another design feature is to serve both exploratory goals (by showing the pattern of observed frequencies in the full table), and model building goals (by displaying the residuals from a given log-linear model).

The use of the mosaic display in connection with loglinear models is introduced here and extended in Chapter 7.

5.1. Introduction
5.2. Two-way tables
5.3. Three-way tables
5.4. Mosaic matrices for categorical data

6 Chapter 6: Correspondence analysis

Correspondence analysis is an exploratory technique related to to principal components analysis which finds a multidimensional representation of the association between the row and column categories of a two-way contingency table.

This chapter illustrates the use of correspondence analysis in understanding the nature of association in two-way tables, and describes how informative plots can be produced from the results of the **ca** package. Extensions of correspondence analysis to multi-way tables and related biplot methods are then described and illustrated.

7 Chapter 7: Loglinear and logit models

Loglinear models provide a comprehensive scheme to describe and understand the associations among two or more categorical variables, particularly when no one variable is singled out as a response to be predicted from the remaining explanatory variables.

For larger tables (three or more variables), however, it becomes difficult to interpret the nature of these associations from tables of parameter estimates. I first illustrate how results from such models may be more easily understood from plots of predicted log odds and probabilities.

The chapter then shows how mosaic displays and correspondence analysis plots can be used to complement the description provided by loglinear models, and how to construct diagnostic plots to determine if a few cells are having undue influence on the overall model.

A collection of mosaic plots, stratified by one (or more) variable(s) is used to display the partial associations among the remaining variables. A mosaic scatterplot matrix is introduced, showing all pairwise associations among variables.

8 Chapter 8: Logistic regression

Logistic regression describes the relationship between a dichotomous response variable and a set of explanatory variables. The explanatory variables may be continuous or (with dummy variables) discrete.

This chapter describes the general logistic regression model and illustrates how the analysis of dichotomous (and polytomous) response data can be enhanced by graphical display.

For interpreting and understanding the results of a fitted model, I emphasize plotting predicted probabilities and predicted log odds. For model criticism and diagnosis, I introduce some discrete analogs of the influence and other plots useful in ordinary least squares regression.

9 Chapter 9: Generalized linear models

Generalized linear models extend the familiar linear models of regression and ANOVA to include counted data, frequencies, and other data for which the assumptions of independent, normal errors are not reasonable. I rely on the analogies between ordinary and generalized linear models (GLIMs) to develop visualization methods to display the fitted relations and check model assumptions.

10 Chapter 10: Regression models for count data

An important special case of GLMs occurs when the response variable is a frequency or count of some event. Some examples are: the number of physician office visits by medical patients, number of deaths from an infectious disease, number of insects found on plants in an agricultural experiment.

The simplest, classical case is that of Poisson regression, where the conditional distribution of the response given the explanatory variables is Poisson, but this model is often overly restrictive. Negative binomial regression can be used for over-dispersed count data, that is, when the conditional variance exceeds the conditional mean. Zero-inflated models attempt to account for situations in which there is an excess of zero counts, by positing an additional sub-model to deal with the excess zeros.

10.2. Negative binomial models
10.3. Zero-inflated models
10.4. Zero-truncated models
10.5. Chapter summary
10.6. Further reading
10.7. Lab exercises

11 Chapter 11: Repeated measures and Longitudinal data

This chapter develops several somewhat different forms of analysis and graphical display related to repeated measures and longitudinal data with categorical responses.

Model-based methods are most easily visualized by plotting predicted probabilities. Sequential analysis pertains to the sequences of behavioral events (e.g., statements, actions of children and parents) observed over time and classified into categories of a classification scheme. Generalized Estimating Equations (GEE) provide one approach to extending GLIMs to longitudinal observations.

11.1. Analysis of marginal probabilities
11.2. Multiple populations
11.3. Sequential analysis
11.4. GEE models
11.5. Chapter summary
11.6. Further reading
11.7. Lab exercises

12 Chapter 12: Classification and regression trees

Recursive partitioning methods provide an alternative to (generalized) linear models for categorical responses, particularly when there are numerous potential predictors and/or there are important interactions among predictors. These methods attempt to define a set of rules to classify observations into mutually exclusive subsets based on combinations of the explanatory variables, and tend to work well when there are important non-linearities or interactions in the data.

12.1. Introduction to recursive partitioning methods
12.2. Recursive partitioning trees
12.3. Conditional inference for classification trees
12.4. Chapter summary
12.5. Further reading
12.6. Lab exercises

Appendix A: Other material

My intention is that all of the datasets, and R functions from the book will be contained in publicly available R packages, where they will be fully documented, with some examples. Nevertheless, the reader of the book will find it useful to have some of this information and other material summarized or listed in print form, in ways that will enhance reading and use of the book, and allow easy reference from the chapters where they are used. Some of this material might better appear in specialized indexes, e.g., a dataset index, an R index, etc.

References

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York: Wiley, 2nd edn.

Agresti, A. (2013). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. New York: Wiley-Interscience [John Wiley & Sons], 3rd edn.

Christensen, R. (1997). *Log-Linear Models and Logistic Regression*. New York, NY: Springer, 2nd edn.

Emerson, J. W., Green, W. A., Schloerke, B., Crowley, J., Cook, D., Hofmann, H., and Wickham, H. (2013). The generalized pairs plot. *Journal of Computational and Graphical Statistics*, 22(1), 79–91.

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89, 190–200.

Friendly, M. (1999). Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, 8(3), 373–395.

Gower, J., Lubbe, S., and Roux, N. (2011). *Understanding Biplots*. Wiley.

Hartigan, J. A. and Kleiner, B. (1981). Mosaics for contingency tables. In W. F. Eddy, ed., *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, (pp. 268–273). New York, NY: Springer-Verlag.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

Mendenhall, W. and Sincich, T. (2003). *A Second Course in Statistics: Regression Analysis*. Prentice Hall / Pearson Education.

Neter, J., Wasserman, W., and Kutner, M. H. (1990). *Applied Linear Statistical Models : Regression, Analysis of Variance, and Experimental Designs*. Homewood, IL: R. D. Irwin, Inc., 3rd edn.

Powers, D. A. and Xie, Y. (2008). *Statistical Methods for Categorical Data Analysis*. Bingley, UK: Emerald, 2nd edn.

Simonoff, J. S. (2003). *Analyzing Categorical Data*. Springer Texts in Statistics. New York: Springer.

Stokes, M. E., Davis, C. S., and Koch, G. G. (2000). *Categorical Data Analysis Using the SAS System*. Cary, NC: SAS Institute, 2nd edn.