Discrete Data Analysis with R: Solutions and Hints to Exercises

December 27, 2015

Chapter 1

Introduction

These questions are all conceptual. No solutions are provided.

Exercise 1.1 A web page, "The top ten worst graphs," `http://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/` by Karl Broman lists his picks for the worst graphs (and a table) that have appeared in the statistical and scientific literature. Each entry links to graph(s) and a brief discussion of what is wrong and how it could be improved.

(a) Examine a number of recent issues of a scientific or statistical journal in which you have some interest. Find one or more examples of a graph or table that is a particularly bad use of display material to summarize and communicate research findings. Write a few sentences indicating how or why the display fails and how it could be improved.
(b) Do the same task for some popular magazine or newspaper that uses data displays to supplement the text for some story. Again, write a few sentences describing why the display is bad and how it could be improved.

Exercise 1.2 As in the previous exercise, examine the literature in recent issues of some journal of interest to you. Find one or more examples of a graph or table that you feel does a *good* job of summarizing and communicating research findings.

(a) Write a few sentences describing why you chose these displays.
(b) Now take the role of a tough journal reviewer. Are there any features of the display that could be modified to make them more effective?

Exercise 1.3 Infographics are another form of visual displays, quite different from the data graphics featured in this book, but often based on some data or analysis. Do a Google image search for the topic "Global warming" to see a rich collection.

(a) Find and study one or two images that attempt some visual explanation of causes and/or effects of global warming. Describe the main message in a sentence or two.
(b) What visual and graphic features are used in these to convey the message?

Exercise 1.4 The Wikipedia web page `en.wikipedia.org/wiki/Portal:Global_warming` gives a few data-based graphics on the topic of global warming. Read the text and study the graphs.

(a) Write a short figure title for each that would announce the conclusion to be drawn in a presentation graphic.
(b) Write a figure caption for each that would explain what is shown and the important graphical details for a reader to understand.

Exercise 1.5 The R Graph Gallery, `http://rgraphgallery.blogspot.com/`, contains a large collection of examples of graphs in R, tagged by type or content, together with the R code to produce them. Explore this collection for the terms (a) association plot (b) bar chart (c) categorical data (d) fluctuation diagram (e) mosaic plot Find one or two you particularly like and write a few sentences saying why you do.

Chapter 2

Working with Categorical Data

Exercise 2.1 The packages vcd (Meyer et al., 2015) and vcdExtra (Friendly, 2015) contain many data sets with some examples of analysis and graphical display. The goal of this exercise is to familiarize yourself with these resources.

You can get a brief summary of these using the function `datasets()` from vcdExtra. Use the following to get a list of these with some characteristics and titles.

```
> ds <- datasets(package = c("vcd", "vcdExtra"))
> str(ds, vec.len = 2)

'data.frame': 75 obs. of  5 variables:
 $ Package: chr  "vcd" "vcd" ...
 $ Item   : chr  "Arthritis" "Baseball" ...
 $ class  : chr  "data.frame" "data.frame" ...
 $ dim    : chr  "84x5" "322x25" ...
 $ Title  : chr  "Arthritis Treatment Data" "Baseball Data" ...
```

(a) How many data sets are there altogether? How many are there in each package? TRUE
`nrow()` gives the number of rows in a data frame. `table()` for a single variable gives the frequencies for each level.

```
> ds <- datasets(package=c("vcd", "vcdExtra"))
> nrow(ds)

[1] 75

> table(ds$Package)

     vcd vcdExtra
      33       42
```

TRUE

(b) Make a tabular display of the frequencies by `Package` and `class`. TRUE
TRUE

(c) Choose one or two data sets from this list, and examine their help files (e.g., `help(Arthritis)` or `?Arthritis`). You can use, e.g., `example(Arthritis)` to run the R code for a given example.
TRUE
TRUE

Exercise 2.2 For each of the following data sets in the vcdExtra package, identify which are response variable(s) and which are explanatory. For factor variables, which are unordered (nominal) and which should be treated as ordered? Write a sentence or two describing substantitive questions of interest for analysis of the data. (*Hint*: use `data(foo, package="vcdExtra")` to load, and `str(foo)`, `help(foo)` to examine data set `foo`.)

(a) Abortion opinion data: *`Abortion`* TRUE
TRUE

(b) Caesarian Births: *`Caesar`* TRUE
TRUE

(c) Dayton Survey: *DaytonSurvey* TRUE
TRUE
(d) Minnesota High School Graduates: *Hoyt* TRUE
TRUE

Exercise 2.3 The data set *UCBAdmissions* is a 3-way table of frequencies classified by Admit, Gender, and Dept.

(a) Find the total number of cases contained in this table. TRUE
TRUE
(b) For each department, find the total number of applicants. TRUE
TRUE
(c) For each department, find the overall proportion of applicants who were admitted. TRUE
TRUE
(d) Construct a tabular display of department (rows) and gender (columns), showing the proportion of applicants in each cell who were admitted relative to the total applicants in that cell. TRUE
TRUE

TRUE

Answer:

(a) Use `sum(UCBAdmissions)`.
(b) Use `margin.table(UCBAdmissions, 3)` to find the marginal total for the third dimension (dept).
(c)

TRUE

Exercise 2.4 The data set *DanishWelfare* in vcd gives a 4-way, $3 \times 4 \times 3 \times 5$ table as a data frame in frequency form, containing the variable Freq and four factors, Alcohol, Income, Status, and Urban. The variable Alcohol can be considered as the response variable, and the others as possible predictors.

(a) Find the total number of cases represented in this table. TRUE
TRUE
(b) In this form, the variables Alcohol and Income should arguably be considered *ordered* factors. Change them to make them ordered. TRUE
TRUE
(c) Convert this data frame to table form, DanishWelfare.tab, a 4-way array containing the frequencies with appropriate variable names and level names. TRUE
TRUE
(d) The variable Urban has 5 categories. Find the total frequencies in each of these. How would you collapse the table to have only two categories, City, Non-city? TRUE
TRUE
(e) Use structable() or ftable() to produce a pleasing flattened display of the frequencies in the 4-way table. Choose the variables used as row and column variables to make it easier to compare levels of Alcohol across the other factors. TRUE
TRUE

Exercise 2.5 The data set *UKSoccer* in vcd gives the distributions of number of goals scored by the 20 teams in the 1995/96 season of the Premier League of the UK Football Association.

```
> data("UKSoccer", package = "vcd")
> ftable(UKSoccer)

     Away  0  1  2  3  4
Home
0          27 29 10  8  2
1          59 53 14 12  4
2          28 32 14 12  4
3          19 14  7  4  1
4           7  8 10  2  0
```

This two-way table classifies all $20 \times 19 = 380$ games by the joint outcome (Home, Away), the number of goals scored by the `Home` and `Away` teams. The value `4` in this table actually represents 4 or more goals.

(a) Verify that the total number of games represented in this table is 380. TRUE
TRUE

(b) Find the marginal total of the number of goals scored by each of the home and away teams. TRUE
TRUE

(c) Express each of the marginal totals as proportions. TRUE
TRUE

(d) Comment on the distribution of the numbers of home-team and away-team goals. Is there any evidence that home teams score more goals on average? TRUE
TRUE

Exercise 2.6 The one-way frequency table *Saxony* in vcd records the frequencies of families with 0, 1, 2, . . . 12 male children, among 6115 families with 12 children. This data set is used extensively in Chapter 3.

```
> data("Saxony", package = "vcd")
> Saxony

nMales
   0    1    2    3    4    5    6    7    8    9   10   11   12
   3   24  104  286  670 1033 1343 1112  829  478  181   45    7
```

Another data set, *Geissler*, in the vcdExtra package, gives the complete tabulation of all combinations of `boys` and `girls` in families with a given total number of children (`size`). The task here is to create an equivalent table, `Saxony12` from the *Geissler* data.

```
> data("Geissler", package = "vcdExtra")
> str(Geissler)

'data.frame': 90 obs. of  4 variables:
 $ boys : int  0 0 0 0 0 0 0 0 0 0 ...
 $ girls: num  1 2 3 4 5 6 7 8 9 10 ...
 $ size : num  1 2 3 4 5 6 7 8 9 10 ...
 $ Freq : int  108719 42860 17395 7004 2839 1096 436 161 66 30 ...
```

(a) Use `subset()` to create a data frame, `sax12` containing the *Geissler* observations in families with `size==12`. TRUE
TRUE

(b) Select the columns for `boys` and `Freq`. TRUE
TRUE

(c) Use `xtabs()` with a formula, `Freq ~ boys`, to create the one-way table. TRUE
TRUE

(d) Do the same steps again to create a one-way table, `Saxony11`, containing similar frequencies for families of `size==11`. TRUE
TRUE

Exercise 2.7 ⋆ *Interactive coding of table factors*: Some statistical and graphical methods for contingency tables are implemented only for two-way tables, but can be extended to 3+-way tables by recoding the factors to interactive combinations along the rows and/or columns, in a way similar to what `ftable()` and `structable()` do for printed displays.

For the *`UCBAdmissions`* data, produce a two-way table object, `UCB.tab2`, that has the combinations of `Admit` and `Gender` as the rows, and `Dept` as its columns, to look like the result below:

```
                 Dept
Admit:Gender        A   B   C   D   E   F
  Admitted:Female  89  17 202 131  94  24
  Admitted:Male   512 353 120 138  53  22
  Rejected:Female  19   8 391 244 299 317
  Rejected:Male   313 207 205 279 138 351
```

(a) Try this the long way: convert *`UCBAdmissions`* to a data frame (`as.data.frame()`), manipulate the factors (e.g., `interaction()`), then convert back to a table (`as.data.frame()`).
TRUE
TRUE

(b) Try this the short way: both `ftable()` and `structable()` have `as.matrix()` methods that convert their result to a matrix.TRUE
TRUE

Exercise 2.8 The data set *`VisualAcuity`* in vcd gives a $4 \times 4 \times 2$ table as a frequency data frame.

```
> data("VisualAcuity", package = "vcd")
> str(VisualAcuity)

'data.frame': 32 obs. of  4 variables:
 $ Freq  : num  1520 234 117 36 266 ...
 $ right : Factor w/ 4 levels "1","2","3","4": 1 2 3 4 1 2 3 4 1 2 ...
 $ left  : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 2 2 2 2 3 3 ...
 $ gender: Factor w/ 2 levels "male","female": 2 2 2 2 2 2 2 2 2 2 ...
```

(a) From this, use `xtabs()` to create two 4×4 frequency tables, one for each gender. TRUE
TRUE

(b) Use `structable()` to create a nicely organized tabular display. TRUE
TRUE

(c) Use `xtable()` to create a LaTeX or HTML table. TRUE
TRUE

Chapter 3

Fitting and Graphing Discrete Distributions

Exercise 3.1 The *Arbuthnot* data in HistData (Friendly, 2014a) (Example 3.1) also contains the variable `Ratio`, giving the ratio of male to female births.

(a) Make a plot of `Ratio` over `Year`, similar to Figure 3.1. What features stand out? Which plot do you prefer to display the tendency for more male births?
(b) Plot the total number of christenings, `Males + Females` or `Total` (in 000s) over time. What unusual features do you see?

Exercise 3.2 Use the graphical methods illustrated in Section 3.2 to plot a collection of geometric distributions for $p = 0.2, 0.4, 0.6, 0.8$, over a range of values of $k = 0, 1, \ldots 10$.

(a) With `xyplot()`, try the different plot formats using points connected with lines, as in Figure 3.9, or using points and lines down to the origin, as in the panels of Figure 3.10.
(b) Also with `xyplot()`, produce one version of a multi-line plot in a single panel that you think shows well how these distributions change with the probability p of success.
(c) Do the same in a multi-panel version, conditional on p.

Exercise 3.3 Use the data set *WomenQueue* to:

(a) Produce plots analogous to those shown in Section 3.1 (some sort of bar graph of frequencies).
(b) Check for goodness-of-fit to the binomial distribution using the `goodfit()` methods described in Section 3.3.2.
(c) Make a reasonable plot showing departure from the binomial distribution.
(d) Suggest some reasons why the number of women in queues of length 10 might depart from a binomial distribution, $\text{Bin}(n = 10, p = 1/2)$.

Exercise 3.4 Continue Example 3.13 on the distribution of male children in families in Saxony by fitting a binomial distribution, $\text{Bin}(n = 12, p = \frac{1}{2})$, specifying equal probability for boys and girls. [*Hint*: you need to specify both `size` and `prob` values for `goodfit()`.]

(a) Carry out the GOF test for this fixed binomial distribution. What is the ratio of χ^2/df? What do you conclude?
(b) Test the additional lack of fit for the model $\text{Bin}(n = 12, p = \frac{1}{2})$ compared to the model $\text{Bin}(n = 12, p = \hat{p})$ where $\hat{p}$ is estimated from the data.
(c) Use the `plot.gootfit()` method to visualize these two models.

Exercise 3.5 For the *Federalist* data, the examples in Section 3.3.1 and Section 3.3.2 showed the negative binomial to provide an acceptable fit. Compare this with the simpler special case of geometric distribution, corresponding to $n = 1$.

(a) Use `goodfit()` to fit the geometric distribution. [Hint: use `type="nbinomial"`, but specify `size=1` as a parameter.]
(b) Compare the negative binomial and the geometric models statistically, by a likelihood-ratio test of the difference between these two models.
(c) Compare the negative binomial and the geometric models visually by hanging rootograms or other methods.

Exercise 3.6 Mosteller and Wallace (1963, Table 2.4) give the frequencies, n_k, of counts $k = 0, 1, \ldots$ of other selected marker words in 247 blocks of text known to have been written by Alexander Hamilton. The data below show the occurrences of the word *upon*, that Hamilton used much more than did James Madison.

```
> count <- 0 : 5
> Freq <- c(129, 83, 20, 9, 5, 1)
```

(a) Read these data into R and construct a one-way table of frequencies of counts or a matrix or data frame with frequencies in the first column and the corresponding counts in the second column, suitable for use with `goodfit()`.
(b) Fit and plot the Poisson model for these frequencies.
(c) Fit and plot the negative binomial model for these frequencies.
(d) What do you conclude?

Exercise 3.7 The data frame *Geissler* in the vcdExtra package contains the complete data from Geissler's (1889) tabulation of family sex composition in Saxony. The table below gives the number of boys in families of size 11.

boys	0	1	2	3	4	5	6	7	8	9	10	11
Freq	8	72	275	837	1,540	2,161	2,310	1,801	1,077	492	93	24

(a) Read these data into R.
(b) Following Example 3.13, use `goodfit()` to fit the binomial model and plot the results. Is there an indication that the binomial does not fit these data?
(c) Diagnose the form of the distribution using the methods described in Section 3.4.
(d) Try fitting the negative binomial distribution, and use `distplot()` to diagnose whether the negative binomial is a reasonable fit.

Exercise 3.8 The data frame *Bundesliga* gives a similar data set to that for UK soccer scores (*UKSoccer*) examined in Example 3.9, but over a wide range of years. The following lines calculate a two-way table, `BL1995`, of home-team and away-team goals for the 306 games in the year 1995.

```
> data("Bundesliga", package = "vcd")
> BL1995 <- xtabs(~ HomeGoals + AwayGoals, data = Bundesliga,
+                 subset = (Year == 1995))
> BL1995

         AwayGoals
HomeGoals  0  1  2  3  4  5  6
        0 26 16 13  5  0  1  0
        1 19 58 20  5  4  0  1
        2 27 23 20  5  1  1  1
        3 14 11 10  4  2  0  0
        4  3  5  3  0  0  0  0
        5  4  1  0  1  0  0  0
        6  1  0  0  1  0  0  0
```

(a) As in Example 3.9, find the one-way distributions of `HomeGoals`, `AwayGoals`, and `TotalGoals = HomeGoals + AwayGoals`.
(b) Use `goodfit()` to fit and plot the Poisson distribution to each of these. Does the Poisson seem to provide a reasonable fit?

(c) Use `distplot()` to assess fit of the Poisson distribution.
(d) What circumstances of scoring goals in soccer might cause these distributions to deviate from Poisson distributions?

Exercise 3.9 ⋆ Repeat the exercise above, this time using the data for all years in which there was the standard number (306) of games, that is for `Year>1965`, tabulated as shown below.

```
> BL <- xtabs(~ HomeGoals + AwayGoals, data = Bundesliga,
+             subset = (Year > 1965))
```

Exercise 3.10 Using the data *CyclingDeaths* introduced in Example 3.6 and the one-way frequency table `CyclingDeaths.tab = table(CyclingDeaths$deaths)`,

(a) Make a sensible plot of the number of deaths over time. For extra credit, add a smoothed curve (e.g., using `lines(lowess(...))`).
(b) Test the goodness of fit of the table `CyclingDeaths.tab` to a Poisson distribution statistically using `goodfit()`.
(c) Continue this analysis using a `rootogram()` and `distplot()`.
(d) Write a one-paragraph summary of the results of these analyses and your conclusions.

Exercise 3.11 ⋆ The one-way table, *Depends*, in vcdExtra and shown below gives the frequency distribution of the number of dependencies declared in $4,983$ R packages maintained on the CRAN distribution network on January 17, 2014. That is, there were 986 packages that had no dependencies, $1,347$ packages that depended on one other package, ... up to 2 packages that depended on 14 other packages.

Depends	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
# Pkgs	986	1,347	993	685	375	298	155	65	32	19	9	4	9	4	2

(a) Make a bar plot of this distribution.
(b) Use `Ord_plot()` to see if this method can diagnose the form of the distribution.
(c) Try to fit a reasonable distribution to describe dependencies among R packages.

Exercise 3.12 ⋆ How many years does it take to get into the baseball Hall of Fame? The Lahman (Friendly, 2014b) package provides a complete record of historical baseball statistics from 1871 to the present. One table, *HallOfFame*, records the history of players nominated to the Baseball Hall of Fame, and those eventually inducted. The table below, calculated in `help(HallOfFame, package="Lahman")`, records the distribution of the number of years taken (from first nomination) for the 109 players in the Hall of Fame to be inducted (1936–present). Note that `years==0` does not, and cannot, occur in this table, so the distribution is restricted to positive counts. Such distributions are called ***zero-truncated distribution***s. Such distributions are like the ordinary ones, but with the probability of zero being zero. Thus the other probabilities are scaled up (i.e., divided by $1 - \Pr(Y = 0)$) so they sum to 1.

years	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
inducted	46	10	8	7	8	4	2	4	6	3	3	1	4	1	2

(a) For the Poisson distribution, show that the zero-truncated probability function can be expressed in the form

$$\Pr\{X = k \,|\, k > 0\}) = \frac{1}{1 - e^{-\lambda}} \times \frac{e^{-\lambda}\, \lambda^k}{k!} \qquad k = 1, 2, \ldots$$

(b) Show that the mean is $\lambda/(1 - \exp(-\lambda))$.

(c) Enter these data into R as a one-way table, and use `goodfit()` to fit the standard Poisson distribution, as if you hadn't encountered the problem of zero truncation.

Chapter 4

Two-Way Contingency Tables

Exercise 4.1 The data set `fat`, created below, gives a 2×2 table recording the level of cholesterol in diet and the presence of symptoms of heart disease for a sample of 23 people.

```
> fat <- matrix(c(6, 4, 2, 11), 2, 2)
> dimnames(fat) <- list(diet = c("LoChol", "HiChol"),
+                       disease = c("No", "Yes"))
```

(a) Use `chisq.test(fat)` to test for association between diet and disease. Is there any indication that this test may not be appropriate here?
(b) Use a fourfold display to test this association visually. Experiment with the different options for standardizing the margins, using the `margin` argument to `fourfold()`. What evidence is shown in different displays regarding whether the odds ratio differs significantly from 1?
(c) `oddsratio(fat, log = FALSE)` will give you a numerical answer. How does this compare to your visual impression from fourfold displays?
(d) With such a small sample, Fisher's exact test may be more reliable for statistical inference. Use `fisher.test(fat)`, and compare these results to what you have observed before.
(e) Write a one-paragraph summary of your findings and conclusions for this data set.

Exercise 4.2 The data set *Abortion* in vcdExtra gives a $2 \times 2 \times 2$ table of opinions regarding abortion in relation to sex and status of the respondent. This table has the following structure:

```
> data("Abortion", package = "vcdExtra")
> str(Abortion)

 table [1:2, 1:2, 1:2] 171 152 138 167 79 148 112 133
 - attr(*, "dimnames")=List of 3
  ..$ Sex             : chr [1:2] "Female" "Male"
  ..$ Status          : chr [1:2] "Lo" "Hi"
  ..$ Support_Abortion: chr [1:2] "Yes" "No"
```

(a) Taking support for abortion as the outcome variable, produce fourfold displays showing the association with sex, stratified by status.
(b) Do the same for the association of support for abortion with status, stratified by sex.
(c) For each of the problems above, use `oddsratio()` to calculate the numerical values of the odds ratio, as stratified in the question.
(d) Write a brief summary of how support for abortion depends on sex and status.

Exercise 4.3 The *JobSat* table on income and job satisfaction created in Example 2.5 is contained in the vcdExtra package.

(a) Carry out a standard χ^2 test for association between income and job satisfaction. Is there any indication that this test might not be appropriate? Repeat this test using `simulate.p.value = TRUE` to obtain a Monte Carlo test that does not depend on large sample size. Does this change your conclusion?
(b) Both variables are ordinal, so CMH tests may be more powerful here. Carry out that analysis. What do you conclude?

Exercise 4.4 The *Hospital* data in vcd gives a 3×3 table relating the length of stay (in years) of 132 long-term schizophrenic patients in two London mental hospitals with the frequency of visits by family and friends.

(a) Carry out a χ^2 test for association between the two variables.
(b) Use `assocstats()` to compute association statistics. How would you describe the strength of association here?
(c) Produce an association plot for these data, with visit frequency as the vertical variable. Describe the pattern of the relation you see here.
(d) Both variables can be considered ordinal, so `CMHtest()` may be useful here. Carry out that analysis. Do any of the tests lead to different conclusions?

Exercise 4.5 Continuing with the *Hospital* data:

(a) Try one or more of the following other functions for visualizing two-way contingency tables with this data: `plot()`, `tile()`, `mosaic()`, and `spineplot()`. [For all except `spineplot()`, it is useful to include the argument `shade=TRUE`].
(b) Comment on the differences among these displays for understanding the relation between visits and length of stay.

Exercise 4.6 The two-way table *Mammograms* in vcdExtra gives ratings on the severity of diagnosis of 110 mammograms by two raters.

(a) Assess the strength of agreement between the raters using Cohen's κ, both unweighted and weighted.
(b) Use `agreementplot()` for a graphical display of agreement here.
(c) Compare the Kappa measures with the results from `assocstats()`. What is a reasonable interpretation of each of these measures?

Exercise 4.7 Agresti and Winner (1997) gave the data in Table 4.1 on the ratings of 160 movies by the reviewers Gene Siskel and Roger Ebert for the period from April 1995 through September 1996. The rating categories were Con ("thumbs down"), Mixed, and Pro ("thumbs up").

Table 4.1: Movie ratings by Siskel & Ebert, April 1995–September 1996. *Source*: Agresti and Winner (1997)

		Ebert			
		Con	Mixed	Pro	Total
Siskel	Con	24	8	13	45
	Mixed	8	13	11	32
	Pro	10	9	64	83
	Total	42	30	88	160

(a) Assess the strength of agreement between the raters using Cohen's κ, both unweighted and weighted.
(b) Use `agreementplot()` for a graphical display of agreement here.

(c) Assess the hypothesis that the ratings are *symmetric* around the main diagonal, using an appropriate χ^2 test. *Hint*: Symmetry for a square table $\boldsymbol{T}$ means that $t_{ij} = t_{ji}$ for $i \neq j$. The expected frequencies under the hypothesis of symmetry are the average of the off-diagonal cells, $\boldsymbol{E} = (\boldsymbol{T} + \boldsymbol{T}^\mathsf{T})/2$.
(d) Compare the results with the output of `mcnemar.test()`.

Exercise 4.8 For the *VisualAcuity* data set:

(a) Use the code shown in the text to create the table form, `VA.tab`.
(b) Perform the CMH tests for this table.
(c) Use the `woolf_test()` described in Section 4.3.2 to test whether the association between left and right eye acuity can be considered the same for men and women.

Exercise 4.9 The graph in Figure 4.23 may be misleading, in that it doesn't take into account of the differing capacities of the 18 life boats on the *Titanic*, given in the variable `cap` in the *Lifeboats* data.

(a) Calculate a new variable, `pctloaded`, as the percentage loaded relative to the boat capacity.
(b) Produce a plot similar to Figure 4.23, showing the changes over time in this measure.

Chapter 5

Mosaic Displays for n-Way Tables

Exercise 5.1 The data set *criminal* in the package logmult (Bouchet-Valat, 2015) gives the 4×5 table below of the number of men aged 15–19 charged with a criminal case for whom charges were dropped in Denmark from 1955–1958.

```
> data("criminal", package = "logmult")
> criminal

      Age
Year    15  16  17  18  19
  1955 141 285 320 441 427
  1956 144 292 342 441 396
  1957 196 380 424 462 427
  1958 212 424 399 442 430
```

(a) Use `loglm()` to test whether there is an association between `Year` and `Age`. Is there evidence that dropping of charges in relation to age changed over the years recorded here?
(b) Use `mosaic()` with the option `shade=TRUE` to display the pattern of signs and magnitudes of the residuals. Compare this with the result of `mosaic()` using "Friendly shading," from the option `gp=shading_Friendly`. Describe verbally what you see in each regarding the pattern of association in this table.

Exercise 5.2 The data set *AirCrash* in vcdExtra gives a database of all crashes of commercial airplanes between 1993–2015, classified by `Phase` of the flight and `Cause` of the crash. How can you best show is the nature of the association between these variables in a mosaic plot? Start by making a frequency table, `aircrash.tab`:

```
> data("AirCrash", package = "vcdExtra")
> aircrash.tab <- xtabs(~ Phase + Cause, data = AirCrash)
```

(a) Make a default mosaic display of the data with `shade=TRUE` and interpret the pattern of the high-frequency cells.
(b) The default plot has overlapping labels due to the uneven marginal frequencies relative to the lengths of the category labels. Experiment with some of the `labeling_args` options (`abbreviate`, `rot_labels`, etc.) to see if you can make the plot more readable. *Hint*: a variety of these are illustrated in Section 4.1 of `vignette("strucplot")`
(c) The levels of `Phase` and `Cause` are ordered alphabetically (because they are factors). Experiment with other orderings of the rows/columns to make interpretation clearer, e.g., ordering `Phase` temporally or ordering both factors by their marginal frequency.

Exercise 5.3 The Lahman package contains comprehensive data on baseball statistics for Major League Baseball from 1871 through 2012. For all players, the *Master* table records the handedness of players, in terms of throwing (L, R) and batting (B, L, R), where B indicates "both." The table below was generated using the following code:

```
> library(Lahman)
> data("Master", package = "Lahman")
> basehands <- with(Master, table(throws, bats))
```

	Bats		
Throws	B	L	R
L	177	2640	527
R	924	1962	10442

- Use the code above, or else enter these data into a frequency table in R.
- Construct mosaic displays showing the relation of batting and throwing handedness, split first by batting and then by throwing.
- From these displays, what can be said about players who throw with their left or right hands in terms of their batting handedness?

Exercise 5.4 ⋆ A related analysis concerns differences in throwing handedness among baseball players according to the fielding position they play. The following code calculates such a frequency table.

```
> library(Lahman)
> MasterFielding <- data.frame(merge(Master, Fielding, by = "playerID"))
> throwPOS <- with(MasterFielding, table(POS, throws))
```

(a) Make a mosaic display of throwing hand vs. fielding position.
(b) Calculate the percentage of players throwing left-handed by position. Make a sensible graph of this data.
(c) Re-do the mosaic display with the positions sorted by percentage of left-handers.
(d) Is there anything you can say about positions that have very few left-handed players?

Exercise 5.5 For the *Bartlett* data described in Example 5.12, fit the model of no three-way association, H_4 in Table 5.2.

(a) Summarize the goodness of fit for this model, and compare to simpler models that omit one or more of the two-way terms.
(b) Use a mosaic-like display to show the lack of fit for this model.

Exercise 5.6 Red core disease, caused by a fungus, is not something you want if you are a strawberry. The data set *jansen.strawberry* from the agridat (Wright, 2015) package gives a frequency data frame of counts of damage from this fungus from a field experiment reported by Jansen (1990). See the help file for details. The following lines create a $3 \times 4 \times 3$ table of crossings of 3 male parents with 4 (different) female parents, recording the number of plants in four blocks of 9 or 10 plants each showing red core disease in three ordered categories, C1, C2, or C3.

```
> data("jansen.strawberry", package = "agridat")
>
> dat <- jansen.strawberry
> dat <- transform(dat, category = ordered(category,
+                                          levels = c('C1','C2','C3')))
> levels(dat$male) <- paste0("M", 1:3)
> levels(dat$female) <- paste0("F", 1:4)
>
> jansen.tab <- xtabs(count ~ male + female + category, data = dat)
> names(dimnames(jansen.tab)) <- c("Male parent", "Female parent",
+                                  "Disease category")
> ftable(jansen.tab)
```

(a) Use `pairs(jansen.tab, shade=TRUE)` to display the pairwise associations among the three variables. Describe how disease category appears to vary with male and female parent. Why is there no apparent association between male and female parent?
(b) As illustrated in Figure 5.6, use `mosaic()` to prepare a 3-way mosaic plot with the tiles colored in increasing shades of some color according to disease category. Describe the pattern of category C3 in relation to male and female parent. (Hint: the `highlighting` arguments are useful here.)
(c) With `category` as the response variable, the minimal model for association is [MF][C], or `~ 1*2 + 3`. Fit this model using `loglm()` and display the residuals from this model with `mosaic()`. Describe the pattern of lack of fit of this model.

Exercise 5.7 The data set *caith* in MASS (Ripley, 2015) gives another classic 4×5 table tabulating hair color and eye color, this for people in Caithness, Scotland, originally from Fisher (1940). The data is stored as a data frame of cell frequencies, whose rows are eye colors and whose columns are hair colors.

```
> data("caith", package = "MASS")
> caith

       fair red medium dark black
blue    326  38    241  110     3
light   688 116    584  188     4
medium  343  84    909  412    26
dark     98  48    403  681    85
```

(a) The `loglm()` and `mosaic()` functions don't understand data in this format, so use `Caith <- as.matrix(caith)` to convert to array form. Examine the result, and use `names(dimnames(Caith))<-c()` to assign appropriate names to the row and column dimensions.
(b) Fit the model of independence to the resulting matrix using `loglm()`.
(c) Calculate and display the residuals for this model.

(d) Create a mosaic display for this data.

Exercise 5.8 The *HairEyePlace* data in vcdExtra gives similar data on hair color and eye color, for both Caithness and Aberdeen as a $4 \times 5 \times 2$ table.

(a) Prepare separate mosaic displays, one for each of Caithness and Aberdeen. Comment on any difference in the pattern of residuals.

(b) Construct conditional mosaic plots, using the formula `~ Hair + Eye | Place` and both `mosaic()` and `cotabplot()`. It is probably more useful here to suppress the legend in these plots. Comment on the difference in what is shown in the two displays.

Exercise 5.9 Bertin (1983, pp. 30–31) used a 4-way table of frequencies of traffic accident victims in France in 1958 to illustrate his scheme for classifying data sets by numerous variables, each of which could have various types and could be assigned to various visual attributes. His data are contained in *Accident* in vcdExtra, a frequency data frame representing his $5 \times 2 \times 4 \times 2$ table of the variables `age`, `result` (died or injured), `mode` of transportation, and `gender`.

```
> data("Accident", package = "vcdExtra")
> str(Accident, vec.len=2)

'data.frame': 80 obs. of  5 variables:
 $ age   : Ord.factor w/ 5 levels "0-9"<"10-19"<..: 5 5 5 5 5 ...
 $ result: Factor w/ 2 levels "Died","Injured": 1 1 1 1 1 ...
 $ mode  : Factor w/ 4 levels "4-Wheeled","Bicycle",..: 4 4 2 2 3 ...
 $ gender: Factor w/ 2 levels "Female","Male": 2 1 2 1 2 ...
 $ Freq  : int  704 378 396 56 742 ...
```

(a) Use `loglm()` to fit the model of mutual independence, `Freq ~ age+mode+gender+result` to this data set.

(b) Use `mosaic()` to produce an interpretable mosaic plot of the associations among all variables under the model of mutual independence. Try different orders of the variables in the mosaic. (*Hint*: the `abbreviate` component of the `labeling_args` argument to `mosaic()` will be useful to avoid some overlap of the category labels.)

(c) Treat `result` (`"Died"` vs. `"Injured"`) as the response variable, and fit the model `Freq ~ age*mode*gender + result` that asserts independence of `result` from all others jointly.

(d) Construct a mosaic display for the residual associations in this model. Which combinations of the predictor factors are more likely to result in death?

Exercise 5.10 The data set *Vietnam* in vcdExtra gives a $2 \times 5 \times 4$ contingency table in frequency form reflecting a survey of student opinion on the Vietnam War at the University of North Carolina in May 1967. The table variables are sex, year in school, and response, which has categories: (A) Defeat North Vietnam by widespread bombing and land invasion; (B) Maintain the present policy; (C) De-escalate military activity, stop bombing and begin negotiations; (D) Withdraw military forces immediately. How does the chosen response vary with sex and year?

```
> data("Vietnam", package = "vcdExtra")
> str(Vietnam)

'data.frame': 40 obs. of  4 variables:
 $ sex     : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1 1 1 ...
 $ year    : int  1 1 1 1 2 2 2 2 3 3 ...
 $ response: Factor w/ 4 levels "A","B","C","D": 1 2 3 4 1 2 3 4 1 2 ...
 $ Freq    : int  13 19 40 5 5 9 33 3 22 29 ...
```

(a) With `response` (R) as the outcome variable and `year` (Y) and `sex` (S) as predictors, the minimal baseline loglinear model is the model of joint independence, [R][YS]. Fit this model, and display it in a mosaic plot.
(b) Construct conditional mosaic plots of the `response` versus `year` separately for males and females. Describe the associations seen here.
(c) Follow the methods shown in Example 5.10 to fit separate models of independence for the levels of `sex`, and the model of conditional independence, $R \perp Y \mid S$. Verify that the decomposition of G^2 in Eqn. (5.6) holds for these models.
(d) Construct a useful 3-way mosaic plot of the data for the model of conditional independence.

Exercise 5.11 Consider the models for 4-way tables shown in Table 5.3.

(a) For each model, give an independence interpretation. For example, the model of mutual independence corresponds to $A \perp B \perp C \perp D$.
(b) Use the functions shown in the table together with `loglin2formula()` to print the corresponding model formulas for each.

Exercise 5.12 The dataset *`Titanic`* classifies the 2,201 pasengers and crew of the *Titanic* by `Class` (1st, 2nd, 3rd, Crew), `Sex`, `Age`, and `Survived`. Treating `Survived` as the response variable,

(a) Fit and display a mosaic plot for the baseline model of joint independence, [CGA][S]. Describe the remaining pattern of associations.
(b) Do the same for a "main effects" model that allows two-way associations between each of C, G, and A with S.
(c) What three-way association term should be added to this model to allow for greater survival among women and children? Does this give an acceptable fit?
(d) Test and display models that allow additional three-way associations until you obtain a reasonable fit.

Chapter 6

Correspondence Analysis

Exercise 6.1 The *`JobSat`* data in vcdExtra gives a 4×4 table recording job satisfaction in relation to income.

(a) Carry out a simple correspondence analysis on this table. How much of the inertia is accounted for by a one-dimensional solution? How much by a two-dimensional solution?
(b) Plot the 2D CA solution. To what extent can you consider the association between job satisfaction and income "explained" by the ordinal nature of these variables?

Exercise 6.2 Refer to Exercise 5.1 in Chapter 5. Carry out a simple correspondence analysis on the 4×5 table *`criminal`* from the logmult package.

(a) What percentages of the Pearson χ^2 for association are explained by the various dimensions?
(b) Plot the 2D correspondence analysis solution. Describe the pattern of association between year and age.

Exercise 6.3 Refer to Exercise 5.2 for a description of the *AirCrash* data from the vcdExtra package. Carry out a simple correspondence analysis on the 5×5 table of `Phase` of the flight and `Cause` of the crash.

(a) What percentages of the Pearson χ^2 for association are explained by the various dimensions?
(b) Plot the 2D correspondence analysis solution. Describe the pattern of association between phase and cause. How would you interpret the dimensions?
(c) The default plot method uses `map="symmetric"` with points for both rows and columns. Try using `map="symbiplot"` with vectors (`arrows=`) for either rows or columns. (Read `help(plot.ca)` for a description of these options.)

Exercise 6.4 The data set *caith* in MASS gives a classic table tabulating hair color and eye color of people in Caithness, Scotland, originally from Fisher (1940).

(a) Carry out a simple correspondence analysis on this table. How many dimensions seem necessary to account for most of the association in the table?
(b) Plot the 2D solution. The interpretation of the first dimension should be obvious; is there any interpretation for the second dimension?

Exercise 6.5 The same data, plus a similar table for Aberdeen, are given as a three-way table as *HairEyePlace* in vcdExtra.

(a) Carry out a similar correspondence analysis to the last exercise for the data from Aberdeen. Comment on any differences in the placement of the category points.
(b) Analyze the three-way table, stacked to code hair color and place interactively, i.e., for the loglinear model [Hair Place][Eye]. What does this show?

Exercise 6.6 The data set *Gilby* in vcdExtra gives a classic (but now politically incorrect) 6×4 table of English schoolboys classified according to their clothing and their teacher's rating of "dullness" (lack of intelligence).

(a) Compute and plot a correspondence analysis for this data. Write a brief description and interpretation of these results.
(b) Make an analogous mosaic plot of this table. Interpret this in relation to the correspondence analysis plot.

Exercise 6.7 For the mental health data analyzed in Example 6.2, construct a shaded sieve diagram and mosaic plot. Compare these with the correspondence analysis plot shown in Figure 6.2. What features of the data and the association between SES and mental health status are shown in each?

Exercise 6.8 Simulated data are often useful to help understand the connections between data, analysis methods, and associated graphic displays. Section 6.3.1 illustrated interactive coding in R, using a simulated 4-way table of counts of pets, classified by age, color, and sex, but with no associations because the counts had a constant Poisson mean, $\lambda = 15$.

(a) Re-do this example, but in the call to `rpois()`, specify a non-negative vector of Poisson means to create some associations among the table factors.
(b) Use CA methods to determine if and how the structure you created in the data appears in the results.

Exercise 6.9 The *TV* data was analyzed using CA in Example 6.4, ignoring the variable `Time`. Carry out analyses of the 3-way table, reducing the number of levels of `Time` to three hourly intervals as shown below.

```
> data("TV", package="vcdExtra")
> # reduce number of levels of Time
> TV.df <- as.data.frame.table(TV)
> levels(TV.df$Time) <- rep(c("8", "9", "10"), c(4, 4, 3))
> TV3 <- xtabs(Freq ~ Day + Time + Network, TV.df)
> structable(Day ~ Network + Time, TV3)

             Day Monday Tuesday Wednesday Thursday Friday
Network Time
ABC     8            536     861       744      735   1119
        9           1401    1205      1022      682    907
        10           910    1044       668      349    711
CBS     8           1167     646       550      680    509
        9            967     959       409      385    544
        10           789     798       324      270    426
NBC     8            858    1090       512     1927    823
        9            946     890       831     1858    590
        10           825     588       869     2101    585
```

(a) Use the stacking approach (Section 6.3) to perform a CA of the table with `Network` and `Time` coded interactively. You can create this using the `as.matrix()` method for a "structable" object.

```
> TV3S <- as.matrix(structable(Day ~ Network + Time, TV3), sep=":")
```

(b) What loglinear model is analyzed by this approach?
(c) Plot the 2D solution. Compare this to the CA plot of the two-way table in Figure 6.4.
(d) Carry out an MCA analysis using `mjca()` of the three-way table `TV3`. Plot the 2D solution, and compare this with both the CA plot and the solution for the stacked three-way table.

Exercise 6.10 Refer to the MCA analysis of the *PreSex* data in Example 6.8. Use the stacking approach to analyze the stacked table with the combinations of premarital and extramarital sex in the rows and the combinations of gender and marital status in the columns. As suggested in the exercise above, you can use `as.matrix(structable())` to create the stacked table.

(a) What loglinear model is analyzed by this approach? Which associations are included and which are excluded in this analysis?
(b) Plot the 2D CA solution for this analysis. You might want to draw lines connecting some of the row points or column points to aid in interpretation.
(c) How does this analysis differ from the MCA analysis shown in Figure 6.10?

Exercise 6.11 Refer to Exercise 5.10 for a description of the *Vietnam* data set in vcdExtra.

(a) Using the stacking approach, carry out a correspondence analysis corresponding to the loglinear model [R][YS], which asserts that the response is independent of the combinations of year an sex.
(b) Construct an informative 2D plot of the solution, and interpret in terms of how the response varies with year for males and females.
(c) Use `mjca()` to carry out an MCA on the three-way table. Make a useful plot of the solution and interpret in terms of the relationship of the response to year and sex.

Exercise 6.12 Refer to Exercise 5.9 for a description of the *Accident* data set in vcdExtra. The data set is in the form of a frequency data frame, so first convert to table form.

```
> accident.tab <- xtabs(Freq ~ age + result + mode + gender, data=Accident)
```

(a) Use `mjca()` to carry out an MCA on the four-way table `accident.tab`.
(b) Construct an informative 2D plot of the solution, and interpret in terms of how the variable `result` varies in relation to the other factors.

Exercise 6.13 The *UCBAdmissions* data was featured in numerous examples in Chapter 4 (e.g., Example 4.11, Example 4.15) and Chapter 5 (e.g., Example 5.14, Example 5.18).

(a) Use `mjca()` to carry out an MCA on the three-way table `UCBAdmissions`.
(b) Plot the 2D MCA solution in a style similar to that shown in Figure 6.10 and Figure 6.11
(c) Interpret the plot. Is there some interpretation for the first dimension? What does the plot show about the relation of admission to the other factors?

Chapter 7

Logistic Regression Models

Exercise 7.1 Arbuthnot's data on the sex ratio of births in London was examined in Example 3.1. Use a binomial logistic regression model to assess whether the proportion of male births varied with the variables `Year`, `Plague`, and `Mortality` in the *Arbuthnot* data set. Produce effect plots for the terms in this model. What do you conclude?

Exercise 7.2 For the Donner Party data in *Donner*, examine Grayson's 1990 claim that survival in the Donner Party was also mediated by the size of the family unit. This takes some care, because the `family` variable in the *Donner* data is a simplified grouping based on the person's name and known alliances among families from the historical record. Use the following code to compute a `family.size variable` from each individual's last name:

```
> data("Donner", package="vcdExtra")
> Donner$survived <- factor(Donner$survived, labels=c("no", "yes"))
> # use last name for family
> lame <- strsplit(rownames(Donner), ",")
> lame <- sapply(lame, function(x) x[[1]])
> Donner$family.size <- as.vector(table(lname)[lname])
```

(a) Choose one of the models (`donner.mod4`, `donner.mod6`) from Example 7.9 that include the interaction of age and sex and nonlinear terms in age. Fit a new model that adds a main effect of `family.size`. What do you conclude about Grayson's claim?
(b) Produce an effect plot for this model.
(c) Continue, by examining whether the effect of family size can be taken as linear, or whether a nonlinear term should be added.

Exercise 7.3 Use component+residual plots (Section 7.5.3) to examine the additive model for the *ICU* data given by

```
> icu.glm2 <- glm(died ~ age + cancer  + admit + uncons,
+                 data=ICU, family=binomial)
```

(a) What do you conclude about the linearity of the (partial) relationship between age and death in this model?

(b) An alternative strategy is to allow some nonlinear relation for age in the model using a quadratic (or cubic) term like `poly(age, 2)` (or `poly(age, 3)`) in the model formula. Do these models provide evidence for a nonlinear effect of age on death in the ICU?

Exercise 7.4 Explore the use of other marginal and conditional plots to display the relationships among the variables predicting death in the ICU in the model `icu.glm2`. For example, you might begin with a marginal `gpairs()` plot showing all bivariate marginal relations, something like this:

```
> library(gpairs)
> gpairs(ICU[,c("died", "age", "cancer", "admit", "uncons")],
+   diag.pars=list(fontsize=16, hist.color="lightgray"),
+   mosaic.pars=list(gp=shading_Friendly,
+                    gp_args=list(interpolate=1:4)))
```

Exercise 7.5 The data set *Caesar* in vcdExtra gives a 3×2^3 frequency table classifying 251 women who gave birth by Caesarian section by `Infection` (three levels: none, Type 1, Type2) and `Risk`, whether `Antibiotics` were used, and whether the Caesarian section was `Planned` or not. `Infection` is a natural response variable. In this exercise, consider only the binary outcome of infection vs. no infection.

```
> data("Caesar", package="vcdExtra")
> Caesar.df <- as.data.frame(Caesar)
> Caesar.df$Infect <- as.numeric(Caesar.df$Infection %in%
+                                  c("Type 1", "Type 2"))
```

(a) Fit the main-effects logit model for the binary response `Infect`. Note that with the data in the form of a frequency data frame you will need to use `weights=Freq` in the call to `glm()`. (It might also be convenient to reorder the levels of the factors so that `"No"` is the baseline level for each.)

(b) Use `summary()` or car (Fox and Weisberg, 2015)::`Anova()` to test the terms in this model.

(c) Interpret the coefficients in the fitted model in terms of their effect on the odds of infection.

(d) Make one or more effects plots for this model, showing separate terms, or their combinations.

Exercise 7.6 The data set *birthwt* in the MASS package gives data on 189 babies born at Baystate Medical Center, Springfield, MA during 1986. The quantitative response is `bwt` (birth weight in grams), and this is also recorded as `low`, a binary variable corresponding to `bwt < 2500` (2.5 Kg). The goal is to study how this varies with the available predictor variables. The variables are all recorded as numeric, so in R it may be helpful to convert some of these into factors and possibly collapse some low frequency categories. The code below is just an example of how you might do this for some variables.

```
> data("birthwt", package="MASS")
> birthwt <- within(birthwt, {
+   race <- factor(race, labels = c("white", "black", "other"))
+         ptd <- factor(ptl > 0)   # premature labors
+         ftv <- factor(ftv)       # physician visits
+         levels(ftv)[-(1:2)] <- "2+"
+         smoke <- factor(smoke>0)
+         ht <- factor(ht>0)
+         ui <- factor(ui>0)
+   })
```

(a) Make some exploratory plots showing how low birth weight varies with each of the available predictors. In some cases, it will probably be helpful to add some sort of smoothed summary curves or lines.
(b) Fit several logistic regression models predicting low birth weight from these predictors, with the goal of explaining this phenomenon adequately, yet simply.
(c) Use some graphical displays to convey your findings.

Exercise 7.7 Refer to Exercise 5.9 for a description of the *Accident* data. The interest here is to model the probability that an accident resulted in death rather than injury from the predictors `age`, `mode`, and `gender`. With `glm()`, and the data in the form of a frequency table, you can use the argument `weight=Freq` to take cell frequency into account.

(a) Fit the main effects model, `result=="Died" ~ age + mode + gender`. Use car::`Anova()` to assess the model terms.
(b) Fit the model that allows all two-way interactions. Use `anova()` to test whether this model is significantly better than the main effects model.
(c) Fit the model that also allows the three-way interaction of all factors. Does this offer any improvement over the two-way model?
(d) Interpret the results of the analysis using effect plots for the two-way model, separately for each of the model terms. Describe verbally the nature of the `age*gender` effect. Which mode of transportation leads to greatest risk of death?

Chapter 8

Models for Polytomous Responses

Exercise 8.1 For the women's labor force participation data (*Womenlf*), the response variable, `partic`, can be treated as ordinal by using

```
> Womenlf$partic <- ordered(Womenlf$partic,
+                           levels=c('not.work', 'parttime', 'fulltime'))
```

Use the methods in Section 8.1 to test whether the proportional odds model holds for these data.

Exercise 8.2 The data set *housing* in the MASS package gives a $3 \times 3 \times 4 \times 2$ table in frequency form relating (a) satisfaction (`Sat`) of residents with their housing (High, Medium, Low), (b) perceived degree of influence (`Infl`) they have on the management of the property (High, Medium, Low), (c) `Type` of rental (Tower, Atrium, Apartment, Terrace), and (d) contact (`Cont`) residents have with other residents (Low, High). Consider satisfaction as the ordinal response variable.

(a) Fit the proportional odds model with additive (main) effects of housing type, influence in management, and contact with neighbors to this data. (Hint: Using `polr()`, with the data in frequency form, you need to use the `weights` argument to supply the `Freq` variable.)
(b) Investigate whether any of the two-factor interactions among `Infl`, `Type`, and `Cont` add substantially to goodness of fit of this model. (Hint: use `stepAIC()`, with the scope formula `~ .^2` and `direction="forward"`.)
(c) For your chosen model from the previous step, use the methods of Section 8.1.5 to plot the probabilities of the categories of satisfaction.

(d) Write a brief summary of these analyses, interpreting *how* satisfaction with housing depends on the predictor variables.

Exercise 8.3 The data *TV* on television viewing was analyzed using correspondence analysis in Example 6.4, ignoring the variable `Time`, and extended in Exercise 6.9. Treating `Network` as a three-level response variable, fit a generalized logit model (Section 8.3) to explain the variation in viewing in relation to `Day` and `Time`. The *TV* data is a three-way table, so you will need to convert it to a frequency data frame first.

```
> data("TV", package="vcdExtra")
> TV.df <- as.data.frame.table(TV)
```

(a) Fit the main-effects model, `Network ~ Day + Time`, with `multinom()`. Note that you will have to supply the `weights` argument because each row of `TV.df` represents the number of viewers in the `Freq` variable.
(b) Prepare an effects plot for the fitted probabilities in this model.
(c) Interpret these results in comparison to the correspondence analysis in Example 6.4.

Exercise 8.4 * Refer to Exercise 5.10 for a description of the *Vietnam* data set in vcdExtra. The goal here is to fit models for the polytomous `response` varialble in relation to `year` and `sex`.

(a) Fit the proportional odds model to these data, allowing an interaction of `year` and `sex`.
(b) Is there evidence that the proportional odds assumption does not hold for this data set? Use the methods described in Section 8.1 to assess this.
(c) Fit the multinomial logistic model, also allowing an interaction. Use car::`Anova()` to assess the model terms.
(d) Produce an effect plot for this model and describe the nature of the interaction.
(e) Fit the simpler multinomial model in which there is no effect of year for females and the effect of year is linear for males (on the logit scale). Test whether this model is significantly worse than the general multinomial model with interaction.

Chapter 9

Loglinear and Logit Models for Contingency Tables

Exercise 9.1 Consider the data set *DaytonSurvey* (described in Example 2.6), giving results of a survey of use of alcohol (A), cigarettes (C), and marijuana (M) among high school seniors. For this exercise, ignore the variables `sex` and `race`, by working with the marginal table `Dayton.ACM`, a $2 \times 2 \times 2$ table in frequency data frame form.

```
> Dayton.ACM <- aggregate(Freq ~ cigarette + alcohol + marijuana,
+                         data=DaytonSurvey, FUN=sum)
```

(a) Use `loglm()` to fit the model of mutual independence, [A][C][M].
(b) Prepare mosaic display(s) for associations among these variables. Give a verbal description of the association between cigarette and alcohol use.

(c) Use `fourfold()` to produce fourfold plots for each pair of variables, AC, AM, and CM, stratified by the remaining one. Describe these associations verbally.

Exercise 9.2 Continue the analysis of the *DaytonSurvey* data by fitting the following models:

(a) Joint independence, [AC][M]
(b) Conditional independence, [AM][CM]
(c) Homogeneous association, [AC][AM][CM]
(d) Prepare a table giving the goodness-of-fit tests for these models, as well as the model of mutual independence, [A][C][M], and the saturated model, [ACM]. *Hint*: `anova()` and `LRstats()` are useful here. Which model appears to give the most reasonable fit?

Exercise 9.3 The data set *Caesar* in vcdExtra gives a 3×2^3 frequency table classifying 251 women who gave birth by Caesarian section by `Infection` (three levels: none, Type 1, Type2) and `Risk`, whether `Antibiotics` were used, and whether the Caesarian section was `Planned` or not. `Infection` is a natural response variable, but the table has quite a few zeros.

(a) Use `structable()` and `mosaic()` to see the locations of the zero cells in this table.
(b) Use `loglm()` to fit the baseline model [I][RAP]. Is there any problem due to zero cells indicated in the output?
(c) For the purpose of this excercise, treat all the zero cells as *sampling zeros* by adding 0.5 to all cells, e.g., `Caesar1 <- Caesar + 0.5`. Refit the baseline model.
(d) Now fit a "main effects" model [IR][IA][IP][RAP] that allows associations of `Infection` with each of the predictors.

Exercise 9.4 The *Detergent* in vcdExtra gives a $2^3 \times 3$ table classifying a sample of 1,008 consumers according to their preference for (a) expressed `Preference` for Brand "X" or Brand "M" in a blind trial, (b) `Temperature` of laundry water used, (c) previous use (`M_user`) of detergent Brand "M," and (d) the softness (`Water_softness`) of the laundry water used.

(a) Make some mosaic displays to visualize the associations among the table variables. Try using different orderings of the table variables to make associations related to `Preference` more apparent.
(b) Use a `doubledecker()` plot to visualize how `Preference` relates to the other factors.
(c) Use `loglm()` to fit the baseline model [P][TMW] for `Preference` as the response variable. Use a mosaic display to visualize the lack of fit for this model.

Chapter 10

Extending Loglinear Models

Exercise 10.1 Example 10.5 presented an analysis of the data on visual acuity for the subset of women in the *VisualAcuity* data. Carry out a parallel analysis of the models fit there for the `men` in this data set, given by:

```
> data("VisualAcuity", package="vcd")
> men <- subset(VisualAcuity, gender=="male", select=-gender)
```

Exercise 10.2 Table 10.1 gives a 4×4 table of opinions about premarital sex and whether methods of birth control should be made available to teenagers aged 14–16, from the 1991 General Social Survey (Agresti, 2013, Table 10.3). Both variables are ordinal, and their grades are represented by the case of the row and column labels.

Table 10.1: Opinions about premarital sex and availability of teenage birth control. *Source*: Agresti (2013, Table 10.3).

Premarital sex	Birth control			
	DISAGREE	disagree	agree	AGREE
WRONG	81	68	60	38
Wrong	24	26	29	14
wrong	18	41	74	42
OK	36	57	161	157

(a) Fit the independence model to these data using `loglm()` or `glm()`.
(b) Make a mosaic display showing departure from independence and describe verbally the pattern of association.
(c) Treating the categories as equally spaced, fit the $L \times L$ model of uniform association, as in Section 10.1. Test the difference against the independence model with a likelihood-ratio test.
(d) Fit the RC(1) model with `gnm()`, and test the difference of this against the model of uniform association.
(e) Write a brief summary of these results, including plots useful for explaining the relationships in this data set.

Exercise 10.3 For the data on attitudes toward birth control in Table 10.1,

(a) Calculate and plot the observed local log odds ratios.
(b) Also fit the R, C, and R+C models.
(c) Use the method described in Section 10.1.2 to visualize the structure of fitted local log odds ratios implied by each of these models, together with the RC(1) model.

Exercise 10.4 The data set *gss8590* in logmult gives a $4 \times 5 \times 4$ table of education levels and occupational categories for the four combinations of gender and race from the General Social Surveys, 1985–1990, as reported by Wong (2001, Table 2). Wong (2010, Table 2.3B) later used the subset pertaining to women to illustrate RC(2) models. This data is created below as `Women.tab`, correcting an inconsistency to conform with the 2010 table.

```
> data("gss8590", package="logmult")
> Women.tab <- margin.table(gss8590[,,c("White Women", "Black Women")], 1:2)
> Women.tab[2,4] <- 49
> colnames(Women.tab)[5] <- "Farm"
```

(a) Fit the independence model, and also the RC(1) and RC(2) models using `rc()` with marginal weights, as illustrated in Example 10.4. Summarize these statistical tests in a table.
(b) Plot the solution for the RC(2) model with 68% confidence ellipses. What verbal labels would you use for the two dimensions?
(c) Is there any indication that a simpler model, using integer scores for the row (Education) or column (Occupation) categories, or both, might suffice? If so, fit the analogous column effects, row effects, or $L \times L$ model, and compare with the models fit in part (a).

Chapter 11

Generalized Linear Models for Count Data

Exercise 11.1 Poole (1989) studied the mating behavior of elephants over 8 years in Amboseli National Park, Kenya. A focal aspect of the study concerned the mating success of males in relation to age, since larger males tend to be more successful in mating. Her data were used by Ramsey and Schafer (2002, Chapter 22) as a case study, and are contained in the Sleuth2 (Ramsey et al., 2012) package (Ramsey et al., 2012) as *case2201*.

For convenience, rename this to `elephants`, and study the relation between `Age` (at the beginning of the study) and number of successful `Matings` for the 41 adult male elephants observed over the course of this study, ranging in age from 27–52.

```
> data("case2201", package="Sleuth2")
> elephants <- case2201
> str(elephants)

'data.frame': 41 obs. of  2 variables:
 $ Age    : num  27 28 28 28 28 29 29 29 29 29 ...
 $ Matings: num  0 1 1 1 3 0 0 0 2 2 ...
```

(a) Create some exploratory plots of `Matings` against `Age` in the styles illustrated in this chapter. To do this successfully, you will have to account for the fact that `Matings` has a range of only 0–9, and use some smoothing methods to show the trend.

(b) Repeat (a) above, but now plotting `log(Matings+1)` against `Age` to approximate a Poisson regression with a log link and avoid problems with the zero counts.

(c) Fit a linear Poisson regression model for `Matings` against `Age`. Interpret the fitted model *verbally* from a graph of predicted number of matings and/or from the model coefficients. (*Hint*: Using `Age-27` will make the intercept directly interpretable.)

(d) Check for nonlinearity in the relationship by using the term `poly(Age,2)` in a new model. What do you conclude?

(e) Assess whether there is any evidence of overdispersion in these data by fitting analogous quasi-Poisson and negative-binomial models.

Exercise 11.2 The data set *quine* in MASS gives data on absenteeism from schools in rural New South Wales, Australia. 146 children were classified by ethnic background (`Eth`), age (`Age`, a factor), `Sex`, and Learner status (`Lrn`), and the number of days absent (`Days`) from school in a particular school year was recorded.

(a) Fit the all main-effects model in the Poisson family and examine the tests of these effects using `summary()` and car::`Anova()`. Are there any terms that should be dropped according to these tests?

(b) Re-fit this model as a quasi-Poisson model. Is there evidence of overdispersion? Test for overdispersion formally, using `dispersiontest()` from AER (Kleiber and Zeileis, 2015).

(c) Carry out the same significance tests and explain why the results differ from those for the Poisson model.

Exercise 11.3 The data set *AirCrash* in vcdExtra was analyzed in Exercise 5.2 and Exercise 6.3 in relation to the Phase of the flight and Cause of the crash. Additional variables include the number of Fatalities and Year. How does Fatalities depend on the other variables?

(a) Use the methods of this chapter to make some exploratory plots relating fatalities to each of the predictors.
(b) Fit a main effects poisson regression model for Fatalities, and make effects plots to visualize the model. Which phases and causes result in the largest number of fatalities?
(c) A linear effect of Year might not be appropriate for these data. Try using a natural spline term, ns(Year, df) to achieve a better, more adequate model.
(d) Use a model-building tool like add1() or MASS::stepAIC() to investigate whether there are important two-way interactions among the factors and your chosen effect for Year.
(e) Visualize and interpret your final model and write a brief summary to answer the question posed.

Exercise 11.4 Male double-crested cormorants use advertising behavior to attract females for breeding. The *Cormorants* data set in vcdExtra gives some results from a study by Meagan Mc Rae (2015) on counts of advertising males observed two or three times a week at six stations in a tree-nesting colony for an entire breeding season. The number of advertising birds was counted and these observations were classified by characteristics of the trees and nests. The goal was to determine how this behavior varies temporally over the season and spatially over observation stations, as well as with characteristics of nesting sites. The response variable is count and other predictors are shown below. See help(Cormorants, package="vcdExtra") for further details.

```
> data("Cormorants", package = "vcdExtra")
> car::some(Cormorants)

      category week station nest height  density tree_health count
42         Pre    1      B1 full   high      few     healthy     1
55         Pre    2      C1   no   high      few     healthy    11
82         Pre    2      C4   no    mid      few        dead     1
91         Pre    2      B1   no    mid moderate        dead     2
116        Pre    3      C2   no    low      few        dead     3
117        Pre    3      C2   no    mid      few     healthy     4
188 Incubation    4      B1 full    mid      few     healthy     1
209 Incubation    5      C3   no    mid      few     healthy     3
280 Incubation    7      B1   no   high      few     healthy     1
296 Incubation    8      B1   no    mid      few        dead     1
```

(a) Using the methods illustrated in this chapter, make some exploratory plots of the number of advertising birds against week in the breeding season, perhaps stratified by another predictor, like tree height, nest condition, or observation station. To see anything reasonable, you should plot count on a log (or square root) scale, jitter the points, and add smoothed curves. The variable category breaks the weeks into portions of the breeding season, so adding vertical lines separating those will be helpful for interpretation.
(b) Fit a main-effects Poisson GLM to these data and test the terms using Anova() from the car package.
(c) Interpret this model using an effects plot.
(d) Investigate whether the effect of week should be treated as linear in the model. You could try using a polynomial term like poly(week, degree) or perhaps better, using a natural spline term like ns(week, df) from the splines package.
(e) Test this model for overdispersion, using either a quasipoisson family or dispersiontest() in AER.

Exercise 11.5 For the *CodParasites* data, recode the `area` variable as an ordered factor as suggested in footnote 13. Test the hypotheses that prevalence and intensity of cod parasites is linearly related to area.

Exercise 11.6 In Example 11.10, we ignored other potential predictors in the *CodParasites* data: `depth`, `weight`, `length`, `sex`, `stage`, and `age`. Use some of the graphical methods shown in this case study to assess whether any of these are related to prevalence and intensity.

Exercise 11.7 The analysis of the *PhdPubs* data in the examples in this chapter were purposely left incomplete, going only as far as the negative binomial model.

(a) Fit the zero-inflated and hurdle models to this data set, considering whether the count component should be Poisson or negative-binomial, and whether the zero model should use all predictors or only a subset. Describe your conclusions from this analysis in a few sentences.
(b) Using the methods illustrated in this chapter, create some graphs summarizing the predicted counts and probabilities of zero counts for one of these models.
(c) For your chosen model, use some of the diagnostic plots of residuals and other measures shown in Section 11.6 to determine if your model solves any of the problems noted in Example 11.17 and Example 11.18, and whether there are any problems that remain.

Exercise 11.8 In Example 11.19 we used a simple analysis of $\log(\boldsymbol{y} + 1)$ for the multivariate responses in the *NMES1988* data using a classical MLM (Eqn. (11.16)) as a rough approximation of a multivariate Poisson model. The HE plot in Figure 11.40 was given as a visual summary, but did not show the data. Examine why the MLM is not appropriate statistically for these data, as follows:

(a) Calculate residuals for the model `nmes.mlm` using

```
> resids <- residuals(nmes.mlm, type="deviance")
```

(b) Make univariate density plots of these residuals to show their univariate distributions. These should be approximately normal under the MLM. What do you conclude?
(c) Make some bivariate plots of these residuals. Under the MLM, each should be bivariate normal with elliptical contours and linear regressions. Add 2D density contours (`kde2d()`, or `geom_density2d()` in ggplot2 (Wickham and Chang, 2015)) and some smoothed curve. What do you conclude?

References

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

Agresti, A. and Winner, L. (1997). Evaluating agreement and disagreement among movie reviewers. *Chance*, 10(2), 10–14.

Bertin, J. (1983). *Semiology of Graphics*. Madison, WI: University of Wisconsin Press. (trans. W. Berg).

Bouchet-Valat, M. (2015). *logmult: Log-Multiplicative Models, Including Association Models*. R package version 0.6.1.

Fisher, R. A. (1940). The precision of discriminant functions. *Annals of Eugenics*, 10, 422–429.

Fox, J. and Weisberg, S. (2015). *car: Companion to Applied Regression*. R package version 2.0-25/r421.

Friendly, M. (2014a). *HistData: Data sets from the history of statistics and data visualization*. R package version 0.7-5.

Friendly, M. (2014b). *Lahman: Sean Lahman's Baseball Database*. R package version 3.0-1.

Friendly, M. (2015). *vcdExtra: vcd Extensions and Additions*. R package version 0.6-7.

Geissler, A. (1889). Beitrage zur frage des geschlechts verhaltnisses der geborenen. *Z. K. Sachsischen Statistischen Bureaus*, 35(1), n.p.

Grayson, D. K. (1990). Donner party deaths: A demographic assessment. *Journal of Anthropological Research*, 46(3), 223–242.

Jansen, J. (1990). On the statistical analysis of ordinal data when extravariation is present. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 39(1), 75–84.

Kleiber, C. and Zeileis, A. (2015). *AER: Applied Econometrics with R*. R package version 1.2-3.

Mc Rae, M. (2015). *Spatial, Habitat and Frequency Changes in Double-crested Cormorant Advertising Display in a Tree-nesting Colony*. Masters project, environmental studies, York University.

Meyer, D., Zeileis, A., and Hornik, K. (2015). *vcd: Visualizing Categorical Data*. R package version 1.3-3.

Mosteller, F. and Wallace, D. L. (1963). Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302), 275–309.

Poole, J. H. (1989). Mate guarding, reproductive success and female choice in African elephants. *Animal Behavior*, 37, 842–849.

Ramsey, F. L. and Schafer, D. W. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis*. Belmont, CA: Duxbury, 2nd edn.

Ramsey, F. L., Schafer, D. W., Sifneos, J., and Turlach, B. A. (2012). *Sleuth2: Data sets from Ramsey and Schafer's* Statistical Sleuth *(2nd ed)*. R package version 1.0-7.

Ripley, B. (2015). *MASS: Support Functions and Datasets for Venables and Ripley's MASS*. R package version 7.3-40.

Wickham, H. and Chang, W. (2015). *ggplot2: An Implementation of the Grammar of Graphics*. R package version 1.0.1.

Wong, R. S.-K. (2001). Multidimensional association models: A multilinear approach. *Sociological Methods and Research*, 30(2), 197–240.

Wong, R. S.-K. (2010). *Association Models*. Quantitative Applications in the Social Sciences. Los Angeles: SAGE Publications.

Wright, K. (2015). *agridat: Agricultural Datasets*. R package version 1.11.