Review 1, chs 1-3

Dear John,

I've now had a chance to read the materials that you sent. I'll begin by answering your questions, make some general suggestions for the book, and then address some specific, smaller points. With a couple of exceptions, I've deliberately not recorded typos, etc. -- the manuscript is clearly a draft and it would be tedious to list typos.

My general assessment is that this is an excellent project and that it will likely result in a definitive treatment of the subject.

1. Do you have a course for which this book could be used as a required or recommended text? If so, for which course would it be suitable? What is the enrollment and what majors take the course?

No.

2. Which text is used (or has been used) in your course? Does it have any particularly strong or weak features? Which software programs are used?

No applicable.

3. Which chapters would be included in your course? Are any essential topics not included?

Not applicable.

4. Please explain why you do or do not regard the manuscript as technically correct, clearly written, and at an appropriate level of difficulty. Does it have any particular strengths or weaknesses?

Although there are some small problems (see my comments below), both the technical quality of the manuscript and the quality of the writing are excellent. I'd characterize the level of difficulty of the material as low to moderate, which seems to me appropriate for a book on this topic. In my opinion, some of the basic material on use of R is unnecessary and can be eliminated. The major strengths of the manuscript and proposal are the quality of the exposition and the coverage of the book, which, as far as I know, is unique. The book is represented as an update to the lead author's text on graphing categorical data with SAS, but the proposed project is much more extensive and ambitious. In my opinion, R is a more fertile environment than SAS for these ideas.

5. Please explain, based on this sample, why you would or would not adopt this book. How does this book compare with competing books? Does it provide sufficient reason to change from your favorite text? If it is to be published, what are the most important changes that should be made before publication?

I would almost surely adopt this book if I were teaching a course on categorical data analysis, but I'm unlikely to do so. I would use the book along with a text that concentrates more directly on the statistics of categorical data analysis -- as opposed to graphics and computing -- such as Agresti's Categorical Data Analysis. I'd see the two books as complementary.

General Comments and Suggestions

(1) There are three authors of the proposal, but the proposal often uses the first-person, singular. It's not clear to me what the roles of the authors are.

(2) Use of fonts should be more carefully thought out, and conventions for fonts should be applied consistently. For example, on p. 2, an italic typewriter font is used for variable names, even though some of the variable names aren't standard R names (e.g., "Marital status"). Another example is the use of an upright typewriter font for "weight" on p. 4. Still another:
The variables "Treatment" and "Improvement" are given in both italic typewriter font and roman font. And another: Class names are consistently quoted, but sometimes given in typewriter font and sometimes in sans-serif font (e.g., "structable" on p. 28).

(3) Although it's no doubt a matter of taste, I strongly dislike the default knitr input and output style. In particular, the use of "##" double-comment characters for each output line is irritating. (I do understand the utility of being able to copy and paste the output at the command line but think that this is irrelevant to a book, and of dubious value even in an ebook.) I'd use something like opts_chunk$set(comment=NA, prompt=TRUE).

(4) Some of the use of colour in graphs seems unnecessary to me. For example, I don't think that colour adds much to Fig. 1.2.

(5) The book presumably requires a basic knowledge of R, but much of the content of Ch. 2 (properties of vectors, matrices, data frames, etc. -- but perhaps not multidimensional arrays) will be familiar even to relatively unsophisticated R users. I'd get rid of this material -- or perhaps place it in an on-line appendix providing a quick introduction to R -- to concentrate on the proper subject-matter of the book, which is in itself extensive.

(6) Many of the comments provided in the R code are unnecessary. I'd use comments only when something new and unobvious is introduced.

Specific Points

p. 3: I think that it's useful to distinguish between categorical/quantitative and discrete/continuous, since there can be discrete quantitative data.

p. 3 and elsewhere: "data" is sometimes treated as plural (my preference) and sometimes as singular.

p. 6: I think that it's useful to distinguish errors from residuals.

p. 16: The argument "rep" in sample() is "replace" not "repeat" -- and it's better not to abbreviate arguments when describing functions.

p. 17: Actually, the header argument to read.table() doesn't default to
FALSE: "If missing, the value is determined from the file format: header is set to TRUE if and only if the first row contains one fewer field than the number of columns." (From ?read.table.)

p. 18: The exposition here confuses the distinction between unordered and ordered factors with the order of factor levels. The latter need not be alphabetic, even for an unordered factor. Also see Sec. 2.3.

p. 44: I'd use "X", not "k" for the random variable (as is done, e.g., on p.
52).

p. 45: There's an unnecessary second call to with() in the code.

p. 50: I'd characterise the distribution as "reverse J-shaped" rather than "J-shaped."

p. 51: For discrete distributions, I'd prefer to call "d" the "probability-mass function" rather than the "density function."

p. 52: The quantile function is "qbinom(P, n, p)" not "pbinom(P, n, p)."

pp. 67-68: I'd redesign the print() and summary() methods for "goodfit"
objects to conform to the usual R convention that the print() method prints a brief report, while the summary() method provides more detail.

p. 68: The use of "par=list(size=12)" in the call to goodfit() isn't entirely clear to me (nor in ?goodfit). I believe that the implicit assumption is that the levels in the right tail are combined. If that's the case, why not provide more flexibility?

Figs. 3.13 (and others): When graphing on the square-root count scale, I'd prefer to display the original counts at the tick marks (analogous to using a log axis -- or possibly use a second, count axis).

p. 74: Use of an underscore in "Ord_plot()" appears to introduce a new naming convention for functions. Why not "Ordplot()"?

Fig. 3.18: A comment on the outlier at the right would be desirable (there are less-discrepant points that are discussed in the subsequent examples).
By the way, why not use robust weighted regression to fit the line?

p. 84: The double-binomial distribution is introduced here (in Table 3.13) but not discussed in the section on discrete probability distributions.

Review 2, chs 1-3

This book is an updated version of Michael Friendly's earlier book on visualising categorial data using SAS. This time R is used. The plan for the book looks OK. Books on categorical data cover modelling thoroughly and not graphics. This book should have more graphics, there are not many in the chapters provided. The planned chapters have more on modelling than on graphics, so the book's title is not right. Graphics should look good, not trimmed (Fig 2.1) or with bad labels (p28). Graphics need explanation, good captions. The new book by Gerhard Tutz should be mentioned, it covers models in this book in more detail. Many references are old, will the authors add newer ones?

The first draft chapter stops in the middle. The second one provides some useful R information. It is technical and the examples are not so interesting. Sometimes the code needed to do something looks very complicated and does not produce a good result (e.g. the code for Table 3.3 or the code on p75). Some of the code is needed to work with datasets loaded as tables. Real datasets do not often come in tables unless they are from textbooks, so this is old-fashioned. The third chapter is on fitting distributions and says this is important for fitting models later on. Here I don't understand, how do you check the Poisson assumption in a GLM?

Both the incomplete first chapter and the third chapter are the same as in the old book, including the examples and the quotations. New and better examples would be good. Exercises are useful for teaching. The exercises in chapters two and three are mostly just technical. In exercise 6(a) of chapter 3, students are asked to read data into R which is already in R. The horsekicks data are in two different data sets, why not just use one? Real applications would be good.

There are some strange remarks, perhaps jokes? In the proposal it says that 3D mosaic plots are a new graphical method. Mosaic plots are difficult, 3D mosaic plots sound impossible. On p6 it says "Questions involving tests of such hypotheses are answered most easily using a large variety of specific statistical tests, often based on randomization arguments. These include the familiar Pearson chi-square test for two-way tables, the Cochran-Mantel-Haenszel test statistics, Fisher’s exact test, and a wide range of measures of strength of association." So that is most easy?

What will the different authors contribute? The proposal is written by only one author. He plans to test the book on a course for psychologists. There are few examples and none for psychologists in the draft chapters. The Geissler dataset is not in the package vcdExtra. It is good to use Lindsey's work, it is not so good to use his packages, e.g. rmutil, they are difficult to find.

There is a need for a book in this area. The authors are good and the book should sell enough copies. More graphics and better examples would be a help.

It would help to know how much Meyer and Zeileis are going to do, especially if they are going to improve the R code. It is probably unfair to look at the draft chapters too closely, as it looks as though they are currently just a quick and incomplete revision of the old book.

Review 3

The book under review is "Visualizing Categorical Data with R" by Michael Friendly, David Meyer, and Achim Zeileis.

This book is a follow-up or update of Michael Friendly's "Visualizing Categorical Data" (2000). David Meyer and Achim Zeileis are the main authors of the 'vcd' package for R, which started life as an R implementation of the techniques described in Michael Friendly's book.

The original book was unique in focusing on visualization techniques for *categorical* data. The new book updates the old by including more recent visualization techniques and by presenting an R implementation of the visualization techniques (where the original used SAS).

The new book remains unique in its focus on visualization for categorical data. This is both a strength and a weakness. On the plus side, it means that the book has no direct competitors, but the strict focus reduces its potential for adoption as a primary course text. Countering that weakness is the fact that the book contains substantial theoretical and data analysis content as well as the information on visualization.

There is no course that deals exclusively with categorical data analysis at my insitution, but even if such a course existed, it is unlikely that this book would serve as the primary resource. Consequently, it is most likely to feature in a "recommended reading" list. On the other hand, the book is quite likely to appeal to individual researchers and practitioners who deal with categorical data (a VERY broad group) because it provides such a comprehensive coverage and because the authors are recognised experts in visualization of categorical data and/or are primary authors of these state-of-the-art visualization tools.

My understanding is that the original book sold well and I would expect the market to have only expanded. The fact that the new book is based on R should also enhance the potential audience.

Regarding the draft manuscript itself, the writing is clear, engaging, and well-organised. For example, the Introduction in Chapter 1 does a very good job of expressing the purpose and focus of the book and in several other places there are conscious efforts to inform the reader about the organisation of material and foreshadow later content (e.g., Section 1.2.3). Another important plus is the mixture of information not just on *what* analysis to perform, but also *how* to perform the analysis (in R). The R code examples are also very practical and reveal the authors' experience with real analyses of real data projects. My one suggestion is that the reader could be given clear warning that Chapter 2 contains technical R details that could be skipped if the reader is already familiar with R, or even if the reader just wants to get to the categorical data content and come back to the information about R later. The content of Chapter 2 is valuable and necessary and its position makes sense, but some readers may find it a hard slog so early on in the book.

Overall, I think this is a very strong proposal from well-respected authors and I recommend publishing this book.

Review 4, chs. 1-3G

1. Spelling and grammatical errors

There are many small errors, I presume the publishers will do the copy-editing. I report some from the first two pages, the frequency of errors is similar for the rest of the document.

Page 1, first line of Chapter 1: Categorical data consists ["consist"] of variables whose values comprise ["taking values that comprise", or similar...]

There is a mix-up of "data is" and "data are" throughout the text, it should be plural. Sometimes data can be replaced with data set when referring to a specific set of data, e.g. on page 19, (Example 2.2) The Arthritis data set is available ... (all in regular font).

Page 1, 5th last line: "Categorical" means different things... [remove "data"]

Page 2, line 3: Pearsons chi-square [the Greek symbol could be added in parentheses, but I would rather introduce it later in the text]

In general there are many small errors, and quite a few spelling mistakes that any spell-checker would detect (e.g., frequencyes, devine)

Page 2, 4 lines before Table 1.1: distinguish

same line: [could add "(or dichotomous)" after binary, anticipating "polytomous" in the next line]

Furthermore:

Bottom of page 13: "I" is used, but should be "we".

3-way, 4-way etc... should be three-way, four-way (in my opinion)

I do not report all the errors I found in the text provided to me, because a good copy-editor will find them.

2. Format and notation:

Page 4, line 3: ID==57 : it is not necessary to use R expressions in the text, the first-time reader might wonder if this is a misprint, using two == signs. Why not simply say ID equal to 57 in normal font; similarly in middle of same page: " with a weight variable", why use typewriter font for weight? in fact this phrase can be omiited; again on page 5 about halfway down, Arthritis, Sex, Treatment, Improved in typewriter font (with variable names in italics as well), then in next line of R code, the variables names are not italicized, not clear what the reason is for italics before. In some cases variables used in R formulas are printed in regular font, so the authors are not consistent. I would not use the R typewriter font for variable names in the text, just the R function names. At bottom of page 29, some R objects are in regular font between quotes...

Will there be colour graphics? It is not clear to me if this is necessary. Journals mostly do not allow colour graphics, unless you pay a lot, so one should be able to use grayscale to convey the essential. Granularity can also be used (small dots, or hatcheted...).

Page 12: the spacing between the R code and results seems excessive. If anything, the results of a command should be closer to the command, then a bigger line space to the next command & result pair lf lines.

There are line spacing changes now and again, see last quarter of page 36, for example.

Footnote on page 35 should be in acknowledgements section.

In a book on visualization, the authors should have realized that Figures 3.4 and 3.5 are not satisfactory. There is no reason for one set of bars to be thicker than the other (I suspect that the plotting frames are the same size, so the one with fewer categories has been spread out to have thicker bars!). The bars should be thinner in general but of the same width, especially since the vertical scales are identical.

In Section 3.2.1 some abbreviations are introduced for the distributions, e.g. Bin and Pois for binomial and Poisson. Since R already has abbreviations for the distributions and the book is oriented towards R, these should rather be the same as the R ones: e.g., bin and pois (lowercase). SImilarly in the middle of page 52, use lowercase for mean, var and skew, to accord with the function names in R (similarly in subsequent sections)

3. Some substantive issues:

Page 8, after logit displayed formula: which is the logarithm of the odds... [not "may be interpreted as"]. Is the reader meant to understand this at this early stage of the book?

Page 9, second last paragraph of Section 1.3: this is confusing, to say the confidence interval around the decrease of 0.116 is (-0.208 and -0.0235). What would be easier to understand is that a 1 degree decrease in temperature increases the log odds of failure by 0.116, with a confidence interval of (0.0235, 0.238). Then give the following results in terms of multiplicative increases as well (of failure), which is the relevant way to present this example (seeing that there was a lowering of temperature, and increased failure probability).

Page 17, section before Example 2.1: An explanation should be included what happens when row names are in the data file and are to be read in as such. Also, the issue of duplicate row names.

Page 20, 5th line of R code: a case has been made to disnguish frequency and count, and here a frequency is called "count", it should be called "freq" (for example).

Page 25, R output: these are proportions, not percentages

Page 30, middle of page: Table 3.3 referred to later on, is this forward reference OK?

Page 39: I find this example too complicated for the reader at this stage, and not explained at all, on the previous page in the middle "loglinear model" is also mentioned without explanation. At least an explanation of Figure 2.2 should be given.

Page 43: what's an "unstructured sample"?

Page 45, first line of R code: the command spar() is not explained

Page 47, footnote: how is it possible that another random experiment gave exactly a frequency of 18 for the category 10+ ?

Sections 3.1 and 3.2: It is not clear why only the binomial and Poisson distributions are discussed in Section 3.1 and then the full set in Section 3.2. It would be preferable to give an introduction to all the distributions in Section 3.1, what their reason for existence is, their practical importance, etc..., and then the technical properties and R functions in Section 3.2.

As a general comment about the examples used, the historical data sets (horse kicks, Weldon's dice, Federalist papers, families in Saxony,...) are certainly of interest, but there is too much emphasis on them and very little presence of a "modern" present-day application. Even the women in queues data from the last century are out-of-date and not really so interesting.

Page 52, middle of page: "negatively (positively) skewed" should read "positively (negatively) skewed"

Page 55, first line: the accent on the Poisson distribution being mainly used for rare events is not the case, it is used for all sorts of counting problems. Of course, its definition relies on the limit of a binomial as the time or space units tend to an infinite number and the probability of an occurrence in any unit tends to zero.

4. Other comments:

Is it OK to use "doesn't", "haven't", "isn't" and so on? There are several examples of this conversational English, also top of page 29: "OK, you've read...". See also start of Section 2.8.

A small typo in the first line of Chapter 10's plan: GLMs should be GLIMs