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.