Review 1, chs 1-9

As you suggested, I didn't read the book carefully in its entirety, but I did scan all of it and ended up reading about half the content reasonably carefully. I've also commented more on the material with which I'm more familiar.

Please see my responses below interleaved with your questions:

1. Who would find this type of book useful? Can you describe a kind of book that is needed in this area? Will the book serve as a reference, textbook, or both?

This is an excellent book, nearly encyclopedic in its coverage. I personally find it very useful and expect that many other readers will as well. The book can certainly serve as a reference. It could also serve as a supplementary text in a course on categorical data analysis that uses R for computation, or -- because so much statistical detail is provided -- even as the main text for a course on the topic that emphasizes graphical methods. That said, I don't think that it would be possible to cover the entire content of this book coherently in a single course.

2. Would you recommend any changes in the contents that would make this book more useful?

Some general suggestions:

(1) The title of the book is misleadingly narrow. First, the book delves much more deeply into statistical background -- distributions, statistical models, etc. -- than is implied by the title. Second, the book deals extensively with count data as well as categorical data, as the term is normally employed. Perhaps using "discrete" rather than "categorical" would be better, or maybe "categorical and count data." A suggestion: "Visualizing and Analyzing Categorical and Count Data."

(2) A significant strength of the book is its comprehensiveness, but that also makes it difficult for the reader to navigate the text. I see that the yet-to-be-written preface is to include an overview of the book. I think that this is a very important feature, and that it should provide a reasonably detailed guide to the contents of the book. I also worry that readers often skip prefaces; this important orienting material would be more likely read were it to appear in the first chapter.

(3) It would help, where this can be done, to move common or general material into Ch. 1 and possibly Ch. 2. For example, when I was reading Ch. 1, I thought that it should contain a general discussion of the use of color (and later noticed an authors' remark to this effect on p. 168). The authors should also as far as possible avoid assuming that readers will read the book sequentially, so it would be useful to minimize inter-chapter dependencies in the text.

(4) Although this is certainly a matter of taste, I don't like the "knitr" style of R input and associated output in which the output it marked as comments. (Yes, I understand the argument for doing this, concerning the ability to copy and paste commands, but don't buy it -- it's an example of the tail wagging the dog.) If the authors don't want to include command-prompts in displayed commands, then commands could be set in italics to distinguish them from output (or vice-versa), or different colors could be used for input and output. I also find the syntax highlighting in commands more distracting than helpful. Finally here, a consistent style should be adopted for R commands, concerning such matters as use of spaces around operators like +, spaces around = in function arguments, spaces after commas, and use of = in assignments (I'd personally prefer to remove the latter).

Specific comments:

p. 1: Do the authors really believe that truth is "only as perceived by the beholder"?

p. 8: I don't think that the authors ever explain (either here or later) to what hypothesis the p-value printed on the mosaic display refers -- or maybe I just missed the explanation. It could be given here and around p. 163.

p. 29: I wouldn't use the term "graphical user interface" for R Studio -- I'd call it an interactive development environment (IDE) or programming editor.

p. 29: I wouldn't include date objects here with numbers, characters, and logicals, which are atomic modes in R.

p. 37: nrowX should be nrow(X).

p. 39: HairEyeColor is in datasets, not vcd.

p. 42: You might also mention the use of with() here.

p. 43: The tables in the example show proportions, not percentages; you might want to multiply by 100.

p. 51: It would be a good idea to set the random number generator seed explicitly before each example using random data.

p. 217: I'd give a reference for biplots, canonical correlation, and PC analysis here.

p. 218: The matrix P is of proportions (or estimated probabilities), not probabilities.

p. 233: The description of the vertical dimension is reversed (older are at the top).

p. 238: I puzzled over 0+outer(etc.) for a minute before realizing that this simply was used to convert the logical result to numeric. Avoid cute but opaque tricks like this. It would be much clearer to use as.numeric(). Also, Age is consistently plotted over a much wider range than in the data -- another instance is the figure on p. 270.

p. 257: Although the general point that the LPM can produce inadmissible fitted values is of course correct, this example doesn't convincingly illustrate the problem since the fitted values are within the [0, 1] interval over the range of observed ages. A more effective example might be substituted here.

pp. 261-262: The residual deviance for a binary logit model isn't distributed as chi-square with df = residual df because the usual asymptotics don't apply. That is, as n grows so does residual df and the complexity of the saturated model. One can informally compare the residual deviance (or the Pearson statistic) to the residual df as an index of lack of fit but shouldn't compute a p-value. When, as here, the explanatory variable (Age) is discrete, then it's possible to perform a proper LR test of lack of fit by treating the explanatory variable as a factor, which will capture any pattern of relationship. There probably aren't enough cases for this approach in the example. This problem -- treating the deviance from a binary logit model as chi-square -- occurs later as well.

p. 264: There's a problem with using the loess smoother with binary data -- one can get inadmissible fits just as in the LPM. One could instead use a kernel smoother, which will never compute (weighted) proportions outside the [0, 1] interval, or, better, use a logit-based smoother. You can see this problem later in Fig. 7.16 (right panel) on p. 280.

p. 276, Fig. 7.11: The y-axis tick marks could be expanded.

pp. 282-283: The first call to Anova() produces a subset of the tests in the second call, and thus is redundant.

p. 284: Though not mentioned, the AIC and BIC disagree here.

p. 288, Fig. 7.18 (and elsewhere): In my opinion, using confidence bands for a factor is potentially misleading. I see the argument for connecting the points with a line ("profile"), especially when there is more than one line, but the envelope suggests visually that something is going on in-between the levels of the factor. I think that error bars showing the confidence limits are the better choice here.

p. 290, fn. 14: ylim here must be expressed on the logit scale. E.g., defining logit <- function(p) log(p/(1 - p)), then ylim=logit(c(0.5, 0.99)).

p. 292 and elsewhere: It's better to use pchisq(value, df, lower.tail=FALSE) rather than 1 - pchisq(value, df).

p. 315, Fig. 7.35: The code for showing the regression coefficients on the graphs isn't given.

p. 335: I'm not sure what the point is of (redundantly, it's also defined on p. 292) introducing the LRtest() function here and then awkwardly constructing a table line by line after using the function. Why not write a more general function -- at least have it return the p-value -- and possibly put it in the vcd package?

p. 352: The point about 0 observed counts could be misread: Although the saturated loglinear model can't be fit (without doing something special) in the presence of 0 counts, (some) other models typically can. There's a similar possibly misleading statement on p. 367. This point is in fact made later, and a clarification or caveat here might help.

p. 412, Fig. 8.23: The right-side probability axis could use some more ticks at the high end.

p. 416, Fig. 8.28: The labels for the 1-19 and 0 lines aren't really distinguishable; it would help to move them

3. Please explain why you do or do not regard the manuscript as technically correct, clearly written, and at an appropriate level of difficulty. What are its strengths and weaknesses? You may comment on the manuscript; if you do so, please separate the marked pages.

Some detailed comments are recorded above.

Yes, with a few minor lapses that I detected, the book is technically correct. I should add that I'm not expert in all of the topics covered in the text. Again, with minor lapses (e.g., the occasional colloquial "hopefully," use of "I" in a multi-authored book, typos, slightly mangled text and figures, and obvious error messages in the examples, that are to be expected in a draft MS), the book is well and clearly -- indeed at times eloquently -- written. I expect that the authors and copy-editor will take care of the small glitches.

I'd characterize the level of difficulty of the book as moderately high. That is, most social and behavioral science graduate students would find the book demanding; statistics undergrads would likely find its data-analytic sophistication beyond their experience. On the other hand, the book should be easily accessible to statisticians and statistically sophisticated social and other scientists. In the hands of a good teacher, the book could serve as a text for social science and statistics grad students, and grad students in other disciplines using the statistical methods covered in the text.

The strengths of the book include the high quality of the exposition; the range and quality of the examples; the uniqueness of the material; and the breadth and depth of coverage. I frankly can think of no major weaknesses, apart perhaps from the possibility that the encyclopedic coverage will overwhelm the reader. That's why I think it's important to provide some guidance to the reader in the preface or early chapters.

4. What other books are available on this subject? Do they have any particularly strong or weak features? Does this book offer any significant advantages?

There are many books that overlap partly with this book, covering for example categorical data analysis without the emphasis on graphics (such as Agresti's two texts), or covering several of the topics in the book (e.g., many applied-regression texts cover logit and related models), but there is nothing of which I'm aware that competes with it directly. Closest is the first author's own Visualizing Categorical Data, but that considerably older book is much less extensive and uses (in my opinion) software less suitable to the topic (SAS/GRAPH).

5. Please explain why you would or would not recommend publication. If you would, what are the most important changes that should be made before publication?

I most emphatically recommend publication. For suggested changes, please see above.