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.

Review 2, chs 1-9

Some minor points while reading:

1. Why in the figures for the Donner Party do the data points look different in all three plots (Figures 1.3 and 1.4)?

2. Spelling mistake, line 9, page 12: perception

3. page 32, line 6, the authors use "I" when it should be "we"

4. page 36: in the example of the comma-delimited file, having the string "ID" at the start of the first line can be dangerous, as I found in a recent submission of a file as supplementary material to a journal, which could not be read properly: see

https://support.microsoft.com/kb/215591/en-us

To avoid this, call the ID something else, for example Record or RecNo.

5. Fig. 3.14 caption: I would say "curve" not a "smooth curve"

6. Section 6.2.1 first line: correspondence analysis (without C uppercase), also Section 6.2.2 first line.

7. Greenacre (1984) is referred to a few times, but is out of print for many years. However, a (legal) scanned version is now available at www.carme-n.org, this could be indicated in the bibliography (or the exact link to the PDF).

8. P.245, middle of page, line spacing seems to change, also middle of p.251 and possibly elsewhere too.

Chapter 6

Page 219: There is a confusion between the singular value, denoted here by lambda and the eigenvalue, also denoted by lambda. On the previous page the eigenvalue was correctly denoted by lambda squared, but later in the chapter the singular value is denoted by square root of lambda, which is the preferred notation (lambda is usually used for eigenvalues). At present (6.2) and (6.3) are incorrect (with present notation RHS´s should be the squares of lambda). But see later notation in (6.10), (6.11) and (6.12). Also, you do not ever say that the eigenvalue is the squared singular value, which is important because later there are outputs with the term eigenvalue.

Page 220, line 8: would read better as "rescaled by the inverse square roots of the column masses D_c^{-1/2}" (although the reader might be perplexed here by what rescaled by a matrix means...). Also in this sentence it is not true that the row coordinates are equal to these quantities, the row coordinates are equal to the projections of these quantities onto the principal axes.

Page 221, line 3: "supplementary points" mentioned here, but not defined yet, nor afterwards I think. Supplementary points are definitely worth explaining somewhere. Also, they are included in the ca package, whereas other packages generally do not have them.

Example 6.1: perhaps you should indicate that the coordinates given next to Dim1 and Dim2 are standard coordinates (although this could be changed in the package if you prefer... later you get the principal coordinates as products of the plot.ca function)

Page 222, after summary(haireye.ca), you could give the chi-square value and show that this value divided by n is equal to the "total" in the output shown. You do give the chi-square value in the next example, but also don't show the relation with "total", but in the third example (RepVict) on p. 226 you do make the actual computation. I think this should come earlier as the reader might wonder why you keep saying proportion of chi-square whereas they only see eigenvalues and total of eigenvalues (the total inertia).

p. 224, lines 9-10: to avoid confusion say "distances between row points and distances between column points"

Section 6.3: optimality of scores would be interesting to explain, also it is needed because it is referred to in first paragraph of Section 6.5. But simultaneous linear regressions is not so interesting, and I would omit that.

Section 6.4.1: could you either (1) give a real example, since this artificial one is not interesting, with a very dull table, or (preferably) (2) omit this 4-way example entirely, since the next example (suicide rates) gives code for making interactive coding.

(6.7), p. 241: consider removing the square brackets and using simpler notation N_{11}, N_{12}, etc..., and thus N_{ii} on next page.

p. 241, 2nd last para.: it is not true that the "greatest proportion...in all off-diagonal blocks" is being optimized, this is the JCA objective, which hasn't been explained in this chapter (and I'm not sure if it needs to be). The adjusted MCA solution is doing this, conditional on the MCA coordinates, but strictly speaking not even so, since finding the best two scaling factors to explain the "greatest proportion...etc." is not exactly what the adjustment is doing either (the adjustment is slightly sub-optimal). All this is complicated to explain, I know. My way of doing it is to explain that the MCA solution can be improved substantially by adjusting the scaling using that option (lambda="adjusted" in mjca), and leave it at that, maybe giving a reference to Greenacre (2007)'s chapter on JCA.

Figures 6.10 & 6.11?

P. 245, first three lines: this way of computing percentages of inertia is Benzecri's way, not Greenacre's. Also, this is not JCA. Benzecri's percentages are necessarily over-estimating the percentages, Greenacre's are lower bound estimates on the optimal percentages which JCA would obtain.

Fig. 6.14: calibrations are not interesting; they would be if in units of proportions (or percentages) in the column profiles (in this case).

Review 3 Michael Friendly and David Meyer: Visualizing Categorical Data with R

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?

As it stands, this book would I think find its main use as a reference. It might be used for a reading group for academic staff and senior students, or for researchers whose work makes heavy use of graphics. For use for a graduate course, a tutor would have to be very selective, even more selective for an undergraduate course.

It would be helpful to have more information, such as might appear in the preface, on what the authors had in mind in these respects when writing the book.

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

Guidance on some possible routes through the material, for different classes of users who'd not read it from end to end, would help widen its audience.

I'd like to see a couple of paragraphs on the history. Florence Nightingale's wedge plots, often called coxcomb plots, surely warrant a mention.

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.

The strengths of this book are its comprehensive coverage, and its focus on the interplay between analysis and graphical presentation. This latter (focus on the interplay) also makes it very demanding technically — it interconnects with a very wide range of areas of analysis. Even for the examples in the introduction, there are strong technical demands. That range is destined, with the addition of further chapters, to get even wider. Chapter 3 makes stronger mathematical demands than Chapter 4.

Some of the more technical mathematical detail might go into an appendix.

I have added a few specific technical comments, made separately.

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?

This book is to my knowledge unique in its comprehensive coverage. It devotes a whole book to what in most other books on graphics with R, occupies a chapter or two. Also the focus is different. There is very little, as the text stands, on the technicalities of constructing graphs. For example, a quick search turned up nothing on the choice of colour palettes, just lines of code that invoked one or other palette.

Winston Chang's "R Graphics Cookbook" and Paul Murrell's "R Graphics" have the technicalities as a major part of their focus. Users of this book will require direction to places where they can find this detail.

Contrast this book also with Tufte's books. Those books focus in elaborate detail on a small number of examples, with minimal mathematics.

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 recommend publication. I find it difficult to assess the extent of its appeal. It will certainly command an audience. Reviewers will I think find it interesting.

There is a much needed book, somewhere between this book and Tufte's book, that might command a wide audience of journalists, scientists and graphics artists. I do not expect that this book will to any extent fill that role.

The only changes that I can suggest, short of writing quite a different book (!), are those noted under items 2 and 4 above. My preference would be to first test the market with a much shorter and more focused book, leaning more towards the style of the Tufte books. The authors are likely too far down their chosen path to consider that, and perhaps it is not where their strengths lie.

6. Do you prefer the honorarium of $125 or $200 in books?

See my email message for details of the books that I have chosen.

Additional Comments

Section 1.2, page 2.
A note is needed somewhere on the categorisation of continuous data. The loss of information can be serious, leading even to misleading inferences.

page 6, line -3: "These tests make minimal assumptions . . ."

The assumptions, whether or not one cares to call them "minimal", are of strong consequence. Chi-squared tests are commonly and misleadingly used where counts do not enter independently into the cells of the table. This is just as important an issue (well, much the same issue) as the poison vs dispersed poison dichotomy.

page 12 “Goals and design principles for visual data display”

This needs to be front and centre. A less technical version might appear as the initial section, or perhaps following 1.1.

Chapter 3

For my taste, the focus is too much on the distributions, and too little on what motivates the choice of one or other such distribution. The distinction between models for independent counts and models that allow for dependence is surely crucial, and should be highlighted early on, in Table 3.7 or surrounding discussion if not earlier.

Surely the direct modelling of a distribution as dispersed Poisson, using quasi-likelihood, warrants a mention as an alternative to the negative binomial. Zero-inflated poisson or quasi-poisson and hurdle models have wide practical application. Note the wide variety of models handled by the *gamlss* package for R. NB also the *pscl* package.