Journal of Statistical Software

January 2017, Volume 76, Book Review 2. doi: 10.18637/jss.v076.b02

Reviewer: James E. Helmreich
Marist College

Statistical Analysis and Data Display: An Intermediate Course with Examples in R (2nd Edition)

Richard M. Heiberger & Burt Holland
Springer-Verlag, New York, 2015.
ISBN 978-1-4939-2121-8. 898 pp. USD 79.99 (H), 59.99 (e).
http://www.springer.com/9781493921218

Overview

Statistical Analysis and Data Display is an intermediate level text, aimed at masters students in statistics as well as Ph.D students of various fields. It could serve as a text for advanced undergraduates in the mathematical sciences as well, though several of the later chapters are well beyond this level. The authors recommend that readers have a basic background in statistics; having familiarity with R as well would be a good idea, as the book does not really attempt to introduce readers to the language. The code for all examples and graphics is included with the accompanying package **HH**, but is quite dense from the perspective of someone new to R. But for those with a reasonable working knowledge of R careful dissection of their code will be quite informative.

The book's greatest strength is its emphasis on graphics. There are many innovative and informative graphics presented for a variety of techniques. Their graphics (even simple boxplots and histograms) are built on **lattice**; many include the ability to pipe them to **shiny** so that values can be changed dynamically with sliders, checkboxes etc. The **shiny** applications would make excellent pedagogical tools for the classroom or individual student use.

The various chapters cover basic inferential statistics, single and multivariate regression, analysis of variance both one and two-way and for more complex designs, analysis of categorical data, logistic regression, nonparametrics and time series. The level of presentation varies considerably. For the inferential statistics presentation, it is clear that while it starts with the absolute basics, a student really should not be seeing the ideas for the first time here. That is less the case with regression which starts out at a very low level and builds, but still, this is an intermediate level presentation. The one-way analysis of variance chapter could serve as a rigorous first introduction, but by the end of the several chapter sequence covering two-way analysis of variance and complex experimental designs the presentation is quite advanced.

The chapters on logistic regression and time series have brief presentations of the mathematics which would not be appropriate for a first encounter with the methods, but quickly go on to presentations of very nice graphical tools.

In the later chapters these graphics are clearly the main thing the authors wish to present, and they are quite worth studying. The same can be said for virtually every topic covered at whatever level – graphs are carefully constructed and presented prominently. On occasion a graph is presented without much in the way of discussion. Overall though, Heiberger and Holland have produced a very worthwhile and beautiful collection of statistical graphs. The paper it is printed on is very high quality and virtually all figures are in multiple colors. They provide a discussion of color choices for those with color impaired vision, as well as online sites to help the reader to construct color palettes.

Chapter Discussions

Heiberger and Holland begin with a brief introduction to R as well as downloading their package **HH**. The code for all analyses and graphics in their book is included with this package, a key feature of the book. The level of introduction seems a bit inconsistent. For instance, the first R code encountered is the creation of a dataframe and use of `melt` and `dcast` from the package **reshape2** to alter its structure. There is a fairly lengthy discussion of how R handles `NA`s – at a point when nothing else about R has been discussed. There is also a mention of packages to aid in data importation. Oddly, Heiberger and Holland also discuss tabular display and significant digits. The discussion is far too brief to learn R, yet redundant if you are already proficient in the language.

Heiberger and Holland provide a basic review of mathematical probability and statistics that is quite appropriate for a student with some background in the area. There are good illustrated discussions of general statistical tests, and a more extended treatment of errors and especially power calculations, also well illustrated. The presentation is succinct, but perfectly in keeping with an intermediate level textbook – a review of the important concepts that 'you need to know'. The graphs are excellent, and easily reproducible. This includes many excellent **shiny** apps. It is straightforward for those with reasonable competence in R to create their own presentation versions. The **shiny** apps would be good pedagogical tools, being interactive with sliders and buttons. Exercises are typical of any undergraduate mathematical probability/statistics text.

An extensive chapter on graphics follows. Using **lattice**, Heiberger and Holland start at the beginning: boxplots, histograms, scatterplots; even getting to the level of detail of defining such things such as x-axis tick marks, main titles, plotting character, legend, caption, and color. They quickly move to scatterplot matrices and conditioning panel plots. The discussion is quite detailed, outlining and providing examples of both good and poor layouts (they note that `pairs` and `splom` do not by default give the best layout). They do not really discuss code (which is buried deep in their package), so this again is not a primer on R. But for those with a knowledge of base graphics, studying the code for these **lattice** graphics is worthwhile. They have an extensive general discussion of transformations, providing a comprehensive list of reasons one might need to transform – stabilize variance, remove curvature or asymmetry, and respond to patterned residuals. Each is dealt with in different places in the text. There is a good discussion of the Box-Cox power transformations, yet few examples other than of logs. Indeed, it is somewhat disappointing that the examples in the rest of the text rarely if

ever need to be transformed.

The chapter is filled with useful EDA graphics, and discussions of how to interpret them. It is not a primer on how to create the graphics though, which can be quite complicated. For instance, a graph depicting a 'ladder of powers' transformation of two variables (Figure 4.17 p. 105) is breezily stated to use their `ladder.fstar` function. It does, but the actual code for the graphic is located within the one (many page) file for the entire chapter, is approximately 35 lines, contains two newly defined functions, and many presentation level tweaks to the finished graphic. It is relatively easy to get a version with your own data, but understanding what you did to get the graphic is opaque. It is possible to follow along and see what is being done to update the graph, what each step accomplishes with the labels and formatting et cetera – and thus learn by example. But in my estimation it would be unusual for a student at the level of the material being presented to have anywhere near the R skills necessary to parse the code adequately.

The next chapter on Introductory Inference is intended as a refresher for students who have already seen the material. R commands for the individual tests are reproduced in the text for most but not all cases. The presentation is too brief for one who is not familiar with the basic ideas, but that is not the intended audience. There is no real discussion of the subtleties of assumptions or appropriateness of tests (e.g. no cautions about tests on variances), or of non-parametric alternatives, which seems an odd omission in this level text. Maximum likelihood techniques are (very) briefly mentioned.

There is a separate chapter much later in the book that covers the usual nonparametric tests, and seems somewhat perfunctory. It strikes me that having alternative tests in the same location in the text would be a better presentation. Additionally, the authors do not mention bootstrap procedures or resampling techniques. This is a significant lack in any level text (consider that the Lock *et al.* 2012 text introduces resampling to freshman), and to my mind is absolutely necessary in one at this level.

They move on to one-way analysis of variance. The presentation and explanation of fixed and random effects models is quite good. There is an excellent brief presentation on contrasts. Unfortunately the **gmodels** function `fit.contrast` is never mentioned. Unusually, the chapter is very light on graphics. I would (humbly) suggest consideration of `granova.1w` or the more refined `granovagg.1w` in **granova** and **granovaGG** respectively.

Heiberger and Holland move on to a full chapter on multiple comparisons, discussing methods of Bonferroni, Tukey, Dunnett, Scheffé etc. They extend the interesting Mean-Mean scatterplot of Hsu and Peruggia to arbitrary contrasts with their function `mmc`, Figure 1 It takes some work to interpret at first, as is true of many of their graphics. It is appropriate for most (all?) experimental designs and is used repeatedly to good effect throughout the following chapters. The bulk of the chapter is spent on this extension with discussions on the construction and interpretation and examples of these plots. This is a very nice example of the strength of graphical analyses. They also provide very nice functions for panel displays of main effects and interactions, Figure 2.

As mentioned earlier, the text can be inconsistent in the level of the presentation and in the choice of material presented. This is an issue in the first regression chapter, which begins at a quite basic level. But then, the analysis leaps into a very detailed mathematical analysis of the `lm` object ANOVA table. There are odd inclusions: $\hat{\beta}_1$ is written as a weighted sum of the slopes from $(\bar{x}, \bar{y})$ to the individual points. This is accompanied by a graphic showing each

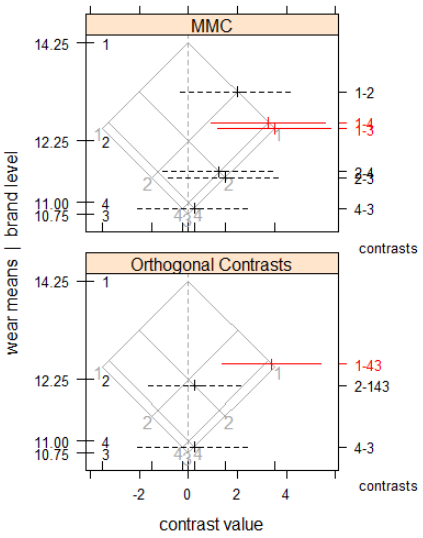

Figure 1: A mean-mean multiple comparisons plot showing Tukey 95% intervals in the top, and below the associated orthogonal contrasts from the Tukey procedure (from page 440 of Heiberger and Holland).

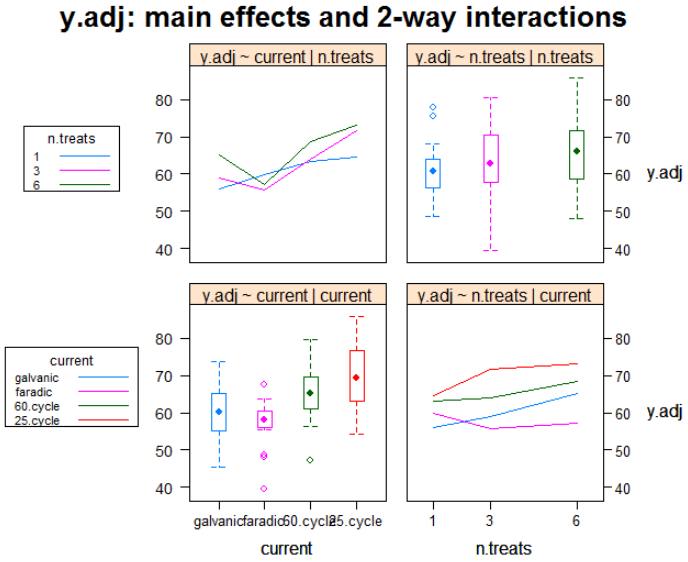

Figure 2: A plot for an ANCOVA analysis depicting two-way interactions of significant main effects (from page 430 of Heiberger and Holland).

line segment in an individual panel, yet no real discussion of it is provided. Many of the tables go well beyond anything that is commented on in the associated text. There are omissions as well: there is no mention of how to interpret coefficients in context, nor how to unravel those interpretations when data have been transformed in some way. To find the predicted values and associated statistics Heiberger and Holland do not choose to use `predict` but delve into

the hat matrix without explanation. Then a bit later, they do use `predict`, but oddly for a model of random response data regressed on three randomly generated covariates. The code is provided – a full page – yet later they find prediction and confidence bands with a nice graphic depicting them, but the code for this more salient analysis is buried deep in the book's R package.

The next chapter on multiple regression begins with a regression on two continuous and two dichotomous variables. This shows the conditioning panels of **lattice** to good effect: the `splom` is noninformative as far as the two categorical variables are concerned, but both can be easily displayed to good effect with **lattice**. However, having produced the graphic, the authors do not seem to use it to discuss adherence to assumptions, need for transformations etc. There is but one simple example with model selection, and they do not pursue the topic at length, which again, seems an odd choice. Discussion of factors and indicator variables is delayed to the next chapter. Finding confidence and prediction intervals are presented as the main goal, while modeling a relationship between a given explanatory variable and the response, and controlling for the other variates is not mentioned. There is a long discussion of the problem of collinearity, with thoughtful examples of manual stepwise covariate selection using VIFs and p-values, contrasted with automatic procedures. The chapter concludes with a very detailed and careful presentation of residual and partial residual plots. The following chapter continues with a presentation on indicator variables, contrasts and analysis of covariance. The discussion is succinct, but quite encompassing and worthwhile. However the choices of what R code to show and what to bury in **HH** are idiosyncratic at best. The third regression chapter covers case statistics graphics for regression diagnostics.

The next three chapters cover two-way ANOVA, factorial designs, and more complex experimental designs. These are arguably the best chapters of the text with excellent diagnostic graphics. The level here is quite a bit more advanced than earlier material on one-way ANOVA and least squares regression – the reader should already be fairly conversant with the ideas. But the back and forth model building and graphical analysis are excellent.

The chapter on logistic regression is a bit brief, though again graphics are put to good use. In several of the plots the predicted values are superimposed on the curves; this gives an idea of location and sample size. Perhaps it would have been too busy, but seeing the actual values as a color coded rug plot might have improved the display.

The final extensive chapter on time series definitely demands extensive prior knowledge of the subject. The mathematical development of the models is succinctly presented, though the real focus is on excellent informative panel plots for comparing different levels of ARIMA models.

There are fifteen sections of appendices totaling 173 pages(!). These provide brief background material on R, including the text's package **HH** and the package **shiny**, as well as data importation and **RExcel**. These are reasonable, though given the level of the text and assumed R knowledge of the reader they strike me as unnecessary, with the possible exception of the relatively new and useful **shiny**. More oddly, they include information on undergraduate level mathematics (algebra, parabolas, calculus, linear algebra) which strike me as quite unnecessary. A section on probability distributions would have been more useful if it had included more than simply graphs of exemplars. Similarly sections on editors (that does not mention **RStudio**) and LaTeX seem unnecessary. The references and the index are minimal.

Conclusion

Statistical Analysis and Data Display is a good choice for at least two different audiences. Some portions (regression, bivariate categorical comparisons, nonparametrics) would be reasonable for upper level undergraduates, as well as the stated audience of masters level statistics students and Ph.D. students in other fields. Other topics (time series, complex experimental designs) assume significant prior knowledge of the subject, but are quite worth studying here as much of the focus is on their informative graphical analyses. These techniques may well be new and of use to students with reasonable backgrounds in those subjects.

The approach is technique by technique, with, typically, data sets that have appeared elsewhere. The text would be improved by the addition of a few extensive case studies that put all of the ideas together and presented to students the overall flow of the data analysis. The most serious omission is a presentation of resampling and bootstrap techniques. The authors do not provide a particularly good set of references, nor discussions of some of the finer points on techniques and assumptions. The book is massive, weighing in at 4 or 5 pounds (2kg). Some editing control over the overextensive appendices would have slimmed it down a bit.

However, these are broadly quibbles. The aim is quite clearly on using data display to aid in statistical analyses, and here Heiberger and Holland have succeeded admirably. If students learn one overarching thing in my classes, it is that you have to look at your data first, and repeatedly. It is very exciting to find a text so strongly in accordance with that principle, and that employs it to such good effect. *Statistical Analysis and Data Display* is well worth consideration for adoption in your courses.

References

Lock RH, Frazer Lock P, Lock Morgan K, Lock EF, Lock DF (2012). *Statistics: Unlocking the Power of Data*. John Wiley & Sons, Chichester.

Reviewer:

James E. Helmreich
Marist College
Department of Mathematics
3399 North Road
Poughkeepsie, NY 12601, United States of America
E-mail: James.Helmreich@Marist.edu
URL: http://foxweb.marist.edu/users/james.helmreich/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
January 2017, Volume 76, Book Review 2 *Published:* 2017-01-14
doi:10.18637/jss.v076.b02