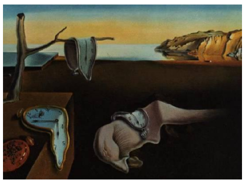

Journal of Statistical Software

November 2005, Volume 14, Book Review 7. *http://www.jstatsoft.org/*

Reviewer: Robert Gould
University of California at Los Angeles

Using R for Introductory Statistics

John Verzani
Chapman & Hall/CRC, Boca Raton, Florida, 2005.
ISBN 1-58488-4509. xvi + 414 pp. USD 44.95 (P).
http://www.math.csi.cuny.edu/UsingR/

In the 1996 R. A. Fisher Lecture, Efron (1998) described pre-Fisher Statistics as "an ingenious collection of ad hoc devices". Roughly 80 years after Fisher's first paper, we educators are still concerned that our courses are oriented more towards teaching disconnected procedures than towards providing students with a coherent approach to understanding the world.

Our ability to teach students the central concepts that organize the ingenious devices depends on many things, not the least of which is the software we use in our class. Educators have a wide variety of software choices that we can crudely dichotomize into the analytically powerful and the easy-to-use. Those of us who want our students to get their hands on the data as quickly and transparently as possible prefer easy-to-use software, such as Fathom. We want our students to study regression in class, and then go home and successfully fit a line, evaluate the fit, learn about the influence of outliers, and maybe explore the effects of external variables on the model; and we want out students to do this without our having to spend class time teaching software commands and without frustrated late-night e-mails. On the other hand, there is a very strong argument to be made for teaching students to use a package that will be useful throughout their careers and that will enable students to learn data management skills. Among these "analytically powerful" packages, R is a very strong choice. R is free, runs on any platform, and is powerful and flexible. A major disadvantage, though, is the steep learning curve. Few, if any, students will be able to use R without explicit instruction.

Using R for Introductory Statistics will help provide that instruction. Anyone who has struggled to produce his or her own notes to help students use R will appreciate this thorough, careful and complete guide aimed at beginning students. In addition to the usual topics found in an introductory statistics text, *Using R* includes slightly more advanced topics such as ANOVA, ANCOVA, and multiple regression as well a chapter on programming in R and a valuable chapter on using R to manipulate and access data.

Although the author claims that the book is meant as a text for beginning students, I suspect that the true strength of this book is as a supplement to a regular textbook. *Using R* does not offer the detail and richness of the best of the introductory textbooks. Explanations are often brief, and read more as an outline of notes rather than an explanation intended to help teach a difficult concept. This means that some topics that introductory students have trouble with, formulating a null hypothesis, for example, are barely touched upon. The major statistical tests are presented in some detail, but no instruction is provided to help students learn how to choose the appropriate test. Another problem is that the distinction between samples and populations is sometimes blurred. The section titled "Shape of a distribution" (Chapter 2.3) shows how to use histograms and frequency polygons to see the shape of the distribution of a set of data, but then introduces the `density()` function before both densities and populations have been introduced.

Cobb (1987) wrote "[j]udge a book by its exercises, and you cannot go far wrong." The exercises in *Using R* solidify my impression that the true strength of this book is in teaching R and not in teaching basic statistical concepts. With some exceptions, homework problems are fairly procedure-oriented, which is a great irony to those who hope that teaching data analysis will help focus students on concepts rather than procedures. For example, the subsection on confidence intervals for differences (Chapter 7.5) has four homework problems. All problems ask students to state their assumptions, but three of the four can be answered simply by stating assumptions and calculating the confidence interval. Although a fairly rich context (comparisons of dosage of AZT for AIDS patients, for example) is provided to motivate the calculation of the confidence intervals, students do not interact with this context and are not pushed to consider the meaning of the confidence interval in this context. The fourth problem is slightly more involved, but extends beyond the other three only by asking whether or not the interval includes zero. I would have preferred "Is this evidence that mothers' and fathers' mean ages are the same?" over "Does it [the confidence interval] contain 0?" (problem 7.28). I believe that routine, procedurally-oriented problems are needed in every textbook, but I prefer a generous helping of problems that require students to make their own decisions about methods to answer research questions and that require interpretations of results.

One strength of this text is the data provided. I suspect that many educators will want this book just for the great variety of interesting data sets and the guidance offered for when and how to use the data to teach. (The datasets are installed with the **UsingR** package.) For example, the chapter on simple linear regression includes data from engineering (predicting time to fluid breakdown as a function of voltage applied), social sciences (the association between illiteracy and graduation rates from the US Census data), economics (predicting the cost of a diamond based on its size), and health (predicting body fat from waist size). Other interesting datasets include cardiac data for domestic dogs, ovarian cancer survival, and the 2000 Florida presidential election data.

Integrating technology into an introductory statistics class remains a challenge to educators, and one useful guideline is that technology should be used to help students visualize and explore data and not "just to follow algorithms to predetermined ends" (Garfield 1995). *Using R* provides many tools for helping instructors help their students. For example, code is provided to access the full range of R's rich graphical tools. (Although this book does not cover more richly interactive graphics, such as those provided by the **iPlots** package.) *Using R* also offers code to run simulations, and relies on simulations to teach confidence intervals and sampling distributions. However, for these lessons to be effective the students must be capable of engaging in the construction of the code and understand the logic behind the simulations. Otherwise the students are merely following an algorithm.

Using R provides instructors with a useful collection of resources for teaching introductory statistics with R. Most will find that it will not stand on its own as a textbook, but that will depend very much on the level and needs of the students as well as the instructor's own talents for teaching the central concepts of statistics and providing coherence to the many ingenious devices covered in this book.

References

Cobb GW (1987). "Introductory Textbook: A Framework for Evaluation." *Journal of the American Statistical Association*, **82**, 321–339.

Efron B (1998). "R. A. Fisher in the 21st Century: Invited Paper Presented at the 1996 R. A. Fisher Lecture." *Statistical Science*, **13**, 95–122.

Garfield J (1995). "How Students Learn Statistics." *International Statistical Review*, **1**.

Reviewer:

Robert Gould
University of California at Los Angeles
Department of Statistics, MC 155404
Los Angeles, CA 90095-1554
United States of America
E-mail: rgould@stat.ucla.edu
URL: http://www.stat.ucla.edu/~rgould/

Journal of Statistical Software *Published:* 2005-11-08
November 2005, Volume 14, Book Review 7.
http://www.jstatsoft.org/