Graphs 'R Us: A Discussion of Antony Unwins Graphical Data Analysis With R

Article *in* Journal of Educational and Behavioral Statistics · December 2015

DOI: 10.3102/1076998615606114

CITATIONS
0

READS
366

3 authors:

Howard Wainer
Independent Statistician and Author
388 PUBLICATIONS **10,740** CITATIONS

SEE PROFILE

Michael Friendly
York University
132 PUBLICATIONS **5,086** CITATIONS

SEE PROFILE

Pere Millan
Universitat Politècnica de Catalunya
14 PUBLICATIONS **1** CITATION

SEE PROFILE

Journal of Educational and Behavioral Statistics
2015, Vol. 40, No. 6, pp. 665–670
DOI: 10.3102/1076998615606114

Graphs *′R* Us: A Discussion of Antony Unwin's *Graphical Data Analysis With R*

Howard Wainer
National Board of Medical Examiners

Michael Friendly
York University

Pere Millán-Martínez
Universitat de Valencia

Antony Unwin. *Graphical Data Analysis With R*. Boca Raton, FL: Taylor & Francis, 2015; 310 pages, $69.95, ISBN 9781498715232.

Thirty-one years ago, a treasure appeared, Andrews and Herzberg's (1984) *Data*. It was composed of 71 separate chapters, each contained only a single, modestly sized, data set; but as a group, they spanned an impossibly broad set of topics, including the number of deaths by horse kicks in the Prussian army, the Stanford heart transplant data, and the motion of stars. Each chapter began with the original reference to the data and a short description of them. All of the data sets had obviously been screened by the authors and had, hidden within them, secrets to be mined and useful lessons to be learned. Since the publication of this trove, unnumbered statistics' students have benefited from their instructors' using some of the content as examples to illustrate and enrich their courses. The days of courses dominated by unreal toy examples consisting of coin flips and dice throws were at an end.

Data, as wonderful as it was, suffered from a shortcoming of its preweb time—there was no easy way to transform the various data sets into digital form, leaving the user to enter the data by hand. Some data sets, like Darwin's data on growth rates of plants (with only 15 rows and three columns), were easy to enter, and doing so helped familiarize you with the data. Others, like Cudworth's data on the motions of stars, which filled six book pages with four digit entries, were sufficiently tedious to enter so as to deter all but the most motivated.

Correspondence should be directed to Howard Wainer, 3750 Market Street, Philadelphia, PA 19104; Email: HWainer@NBME.org

We do not wish to seem ungrateful, but *Data* was only the first third of the book that was needed. The second part would show how to analyze each data set, including both methods and their software implementation. And the last third, would show how to transform the many rough-and-ready plots, aimed at clearly communicating only to the analyst, into presentational graphics (again including the requisite software and how to use it) suitable for communication to a broader audience. Such a document would offer a guided tour from the beginning of a data analysis to a logical conclusion. The instructive value of having a wise mentor, leading us over a broad sample of examples, would be immense. *Data* took us part of the way, but there was still a long way to go.

To a large extent, the second part has been provided in Antony Unwin's very clever new book, *Graphical Data Analysis with R* (hereafter GDA-R). It analyzes 107 data sets, all of which are available online in R packages. The easy availability of all of the data sets not only relieves the reader of having to enter them but also frees Unwin to use data sets of any size. But reading GDA-R is not a passive activity, any more than was reading Tukey's (1977) iconic *Exploratory Data Analysis* (EDA). Forty years ago, when EDA first appeared, a reader needed a pad of graph paper, a set of colored pencils, and a straight edge. A reader of GDA-R now needs a computer and an Internet connection.

But Unwin's goals in GDA-R are more ambitious than merely redoing *Data* with easy-to-access data sets. He wanted to provide the second part of the tripartite ideal book. So, in addition to providing sample data sets, he uses those data to illustrate various approaches to their analysis. He does this in the hopes of instilling in the reader both increased knowledge and good taste. And finally, Unwin shows how these graphical analyses can be done, at least in a rudimentary way, using the powerful high-level programming language R.[1] Be warned, however, GDA-R is meant for readers who have a fair amount of experience with R.

The author's focus in GDA-R is on exploratory data analysis and readers who are looking for advice on making presentational graphics must await the third part of the ideal book. Instead, Unwin shows how to use R for data screening, identifying patterns, structures, relations, trends, and anomalies. The 14 brief chapters cover the ground in a well-organized way, starting from distributions for univariate data and then multivariate relations, and going on to a collection of more specialized topics that lend greater depth to what might otherwise have only limited scope.

It is well written, clearly by a practitioner with wide experience, gives generally good (though sometimes opinionated) advice, and includes R code for nearly all examples, as well as nice collections of additional exercises for each chapter. As such, it is surely of interest to students and researchers in many disciplines who want to learn to use R for the kinds of data problems and questions addressed

here. It would not be directly usable as the principal text for most courses in applied statistics because it gives no statistical details, but it would make an attractive supplementary text for undergraduate courses that use R as the computational/statistical package.

By design, the author has restricted himself to the descriptive/exploratory aspects of GDA, with only brief sections in some chapters addressing "modeling and testing features" of the type of data and questions considered. This keeps the book relatively short but may limit its potential appeal.

Another possible limitation is that the book essentially focuses on the **why** of GDA, rather than the details of the **how**, which is only explicitly stated in the final chapter. Although R code is shown throughout, the code is kept as short as possible (often omitting key aspects like legends or axis labels that would make a better graphic), and the details are usually not explained. Again this keeps the book shorter but with some loss for the reader.[2]

There are now, of course, many books on aspects of graphics for data analysis. Where does Unwin's GDA-R fit in this landscape and in relation to the tripartite view of a complete topic? To find out, we surveyed a representative collection of 16 books on the topic that have appeared over the last 80 years and inferred some salient dimensions that might provide an organization of the topic.

The books ranged in time from Brinton's (1939) *Graphical Presentation* to Katy Börner's (2015) *Atlas of Knowledge* and included such iconic texts as Bertin's (1967) *Sémiologie Graphique*, Tukey's (1977) *Exploratory Data Analysis*, Cleveland's (1985) *The Elements of Graphing Data*, and Wilkinson's (2005) *The Grammar of Graphics*.

As one dimension, we chose the degree to which the main emphasis is on how to **encode** quantitative and categorical information in graphs versus how a graph reader can **decode** that information from a given graphical display. A second dimension would classify books on the degree to which they have a **theoretical** perspective (theory of graphics or graphical perception) or a **practical** point of view (how to). We rated each book on a 9-point scale, from -4 to $+4$. Many books, of course, try to mix some of each, so a completely "balanced" book would appear at (0, 0). Finally, we classified each book by its principal purpose—data analysis, graph construction, or communication.

The final ratings of the books were subjectively determined by consensus among the three authors of this review. Our result, shown in Figure 1, makes this discussion concrete. Readers familiar with these books are invited to ask themselves how they would position them.

In this figure, Unwin (2015) appears in the bottom right quadrant, that is, as a book more on the side of encoding than decoding and more on the practical than the theoretical side. The only book more extreme on both dimensions is Paul Murrell's (2005) *R Graphics*, which is entirely devoted to the low-level details of producing graphic displays with R. Diagonally opposite, in the decoding-theoretical

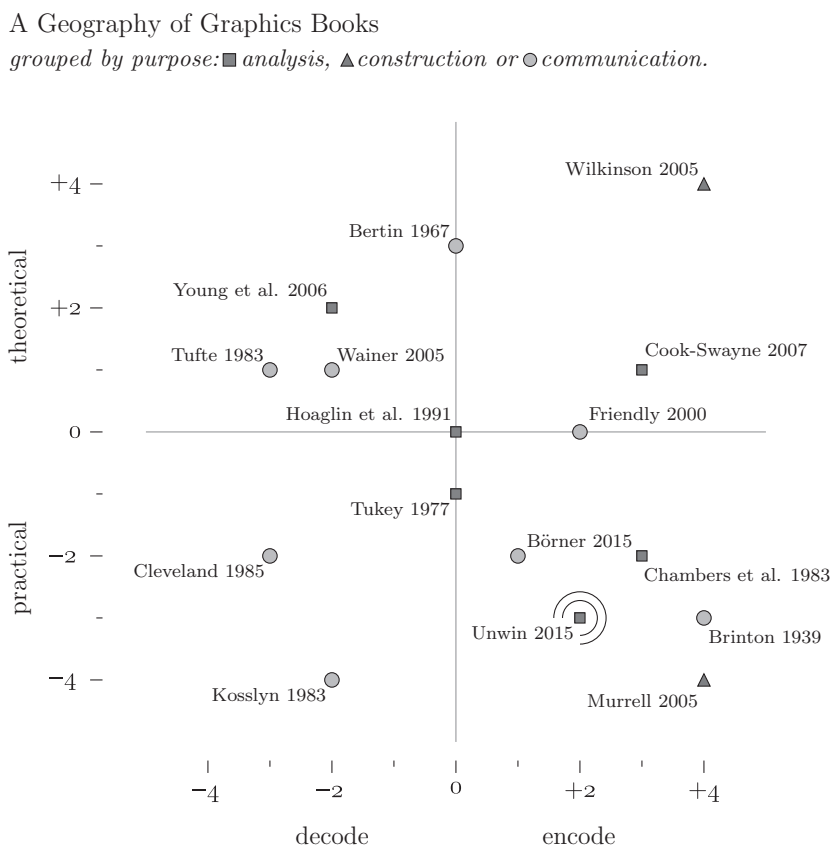

FIGURE 1. *A classification of 16 graphics books in three dimensions. One dimension is the degree to which the main emphasis is on how to encode information versus how this information is decoded. A second dimension is the degree to which they are written from a theoretical perspective (theory of graphics or graphical perception) or a practical point of view (how to). The third dimension is the principal purpose; data analysis, graph construction, or communication.*

quadrant, we see Tufte's (1983) *The Visual Display of Quantitative Information*, Young, Valero-Mora, and Friendly's (2006) *Visual Statistics*, and Wainer's (2005) *Graphic Discovery*. In the opposite diagonal quadrants, Wilkinson's *Grammar of Graphics* stands alone as most theoretical and on the encoding side, and Cleveland's *The Elements of Graphing Data* and Kosslyn's (1994) *Elements of Graph Design* appear as more practically oriented and with more emphasis on how graphs are decoded. The middle ground, mostly balanced on both dimensions, is occupied by Tukey's *EDA*, and its extension, Hoaglin, Mosteller, and Tukey's (1991), *Fundamentals of Exploratory Analysis of Variance*.

What should we make of this in the context of the present review? Andrews and Herzberg's *Data* doesn't appear because it is a book only about data. Unwin's GDA-R is well positioned along the encoding (what to show) and practical (how to show it) axes. What remains missing is a single, comprehensive account of data graphics (perhaps in a restricted domain) that also incorporates aspects of communication goals (what and how to show data for a given audience or to make a particular point).

Beyond the content, Unwin also does an admirable job of conveying enthusiasm for data graphics, and he uses frequent quotations to highlight the poetry and aspirational purpose of statistical graphs. We close this review with one more, from the preface by Henry D. Hubbard to Brinton's (1939) *Graphic Presentation:*

> There is a magic in graphs. The profile of a curve reveals in a flash a whole situation—the life history of an epidemic, a panic, or an era of prosperity. The curve informs the mind, awakens the imagination, convinces.

Notes

1. R grew out of the Bell Labs system S originated by John Chambers and was created by two statisticians at the University of Aukland, Ross Ihaka, and Robert Gentleman. It is freely available under the GNU General Public License.
2. Not that big a problem for most of the examples that use ggplot2, but a number of examples using ggparcoord() and the extracat and dplyr packages use features that may be bewildering to less experienced, readers.

References

Andrews, D. F., & Herzberg, A. M. (1984). *Data: A collection of problems from many fields for the student and research worker*. New York, NY: Springer-Verlag.

Bertin, J. (1967). *Sémiologie Graphique: Les Diagrammes, les Réseaux, les Cartes*. Paris, France: Gauthier-Villars. [English translation: Semiology of Graphics, University of Wisconsin Press, 1983].

Börner, K. (2015). *Atlas of knowledge: Anyone can map*. Cambridge, MA: MIT Press.

Brinton, W. C. (1939). *Graphic presentation*. New York, NY: Brinton Associates.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis*. Belmont, CA: Wadsworth.

Cleveland, W. S. (1985). *The elements of graphing data*. Summit, NJ: Hobart Press.

Cook, D., & Swayne, D. F. (2007). *Interactive and dynamic graphics for data analysis with R and GGobi*. New York, NY: Springer.

Friendly, M. (2000). *Visualizing categorical data*. Cary, NC: SAS Institute.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (1991). *Fundamentals of exploratory analysis of variance*. New York, NY: Wiley.

Kosslyn, S. M. (1994). *Elements of graph design*. New York, NY: W. H. Freeman.

Murrell, P. (2005). *R graphics*. Boca Raton, FL: Chapman & Hall/CRC.

Tukey, J. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Tukey, J. W. (1993). Graphic comparisons of several linked aspects: Alternatives and suggested principles. *Journal of Computational and Graphical Statistics*, *2*, 1–33.

Tufte, E. R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Unwin, A. (2015). *Graphical data analysis with R*. Boca Raton, FL: Taylor & Francis.

Wainer, H. (2005). *Graphic discovery: A trout in the milk and other visual adventures*. Princeton, NJ: Princeton University Press.

Wilkinson, L. (2005). *The grammar of graphics* (2nd ed.). New York, NY: Springer.

Young, F. W., Valero-Mora, P., & Friendly, M. (2006). *Visual statistics: Seeing data with dynamic interactive graphics*. Hoboken, NJ: Wiley-Interscience.

Authors

HOWARD WAINER is Distinguished Research Scientist at the National Board of Medical Examiners. His latest book is *Truth or Truthiness: Distinguishing Fact from Fiction by Learning to Think like a Data Scientist,* which was published by Cambridge University Press in January 2016. He and Michael Friendly met as undergraduates in the Fall of 1962 and have been friends ever since.

MICHAEL FRIENDLY is professor of psychology, and founding chair of the Graduate Program in Quantitative Methods at York University, Toronto. His research interests are the development of graphical methods for categorical and multivariate data and the history of data visualization. His latest book (with David Meyer) is *Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data,* Chapman & Hall, January 2016.

PERE MILLÁN-MARTÍNEZ is doctoral candidate at Universitat de Valencia and currently doing a research stay at York University. His research interests are the characterization of statistical graphics, their automation and optimization from the receiver point of view.