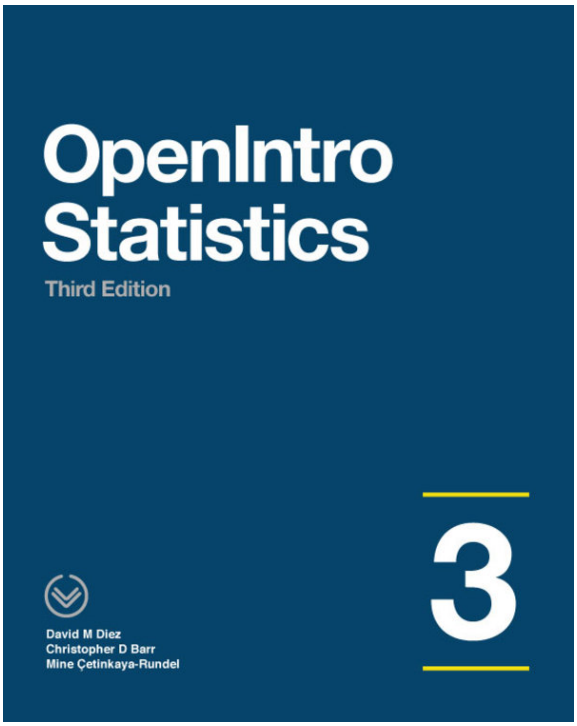
Book Review

OpenIntro Statistics, 3ed

by David M. Diez, Christopher D. Barr and Mine Çetinkaya-Rundel

Pedro J. Aphalo, ORCID: 0000-0003-3385-972X
ViPS, Department of Biosciences, University of Helsinki, Helsinki, Finland
DOI: 10.19232/uv4pb.2016.2.90



Before I start the actual review of the book authored by Diez et al. (2015), I will start by describing what are my expectations for a great introductory textbook of statistics.

1. Focus on principles, stressing how variations of the same ideas reappear across different data analysis methods.
2. Use of numerical exercises to demonstrate not so much the practical details of the use of methods, but rather the overall ideas behind statistics.
3. Very strongly emphasize the dangers of miss-interpreting and over-interpreting the results from tests of significance.
4. Keeping common sense and other criteria for relevance of results "on the table" when discussing and selecting examples and designing exercises.
5. Use of illustrations whenever they can help understanding.
6. Availability of supplementary materials for students and teachers.
7. Affordable and easy to acquire.

The OpenIntro book consists of eight chapters and two appendixes. The titles of the chapters follow a rather traditional path: 1. Introduction to data, 2. Probability, 3. Distributions of Random Variables, 4. Foundations for Inference, 5. Inference for numerical data, 6. Inference for categorical data, 7. Introduction to linear regression and 8. Multiple and logistic regression.

Throughout the book the focus of the discussion, although following a rather traditional organization, is different from other books in that the students are encouraged to get the grasp of the principles through experimentation or examples (see Fig. 1.13 from Diez et al. 2015, p. 18, reproduced in the appendix). Most exercises promote deep understanding, rather than simply exemplifying the mechanics of how a certain type of analysis is done. With examples, the approach used is new to me, but I expect it to be

very effective: there are both "Examples" and "Guided practice" items within the main text. These are meant to be done while reading, and answers are provided for them. In addition to these, at the end of chapters, there are sets of traditional exercises for home work.

The book is clearly written by authors that are familiar with the usual mistakes with regards to statistical analysis in the published literature. Rarely nowadays do errors concern the calculations themselves. They are related mostly to misinterpretation of results of tests or model fits, either because of misunderstanding of assumptions or ignoring that they are not fulfilled. In my experience, students tend to first or only look at *P*-values, rather than starting by considering more broadly the practical or biological significance of the result. Many times, recurse to common sense, is forgotten. This is an area where the book excels. These points are highlighted and many examples used to demonstrate them, including figures (see Fig. 7.13 from Diez et al. 2015, p. 342, reproduced in the appendix).

The book is modern in that many numerical examples are used (417 examples and guided practice items, 350 end-of-chapter exercises), together with many illustrations (213 figures, not counting those in exercises). Example data sets are interesting, sometimes even very entertaining, and plots and diagrams are used as part of the analyses. This is my personal view, but it makes me happy that most of the illustrations are not "cartoons", but instead plots of actual example data.

Supplementary material includes video overviews linked directly from within the PDF file of the book. Data sets used in the book are available as an R package. Furthermore, the OpenIntro name comes from open-source, as both the PDF of the book and the LaTeXsource of the book are available for free! They are licensed under a Creative Commons BY-SA 3.0 license. This book may be downloaded as a free PDF at http://openintro.org. However, I encourage you to make a voluntary payment to help sustain this project, and an easy way of doing this is by buying a copy of the PDF at https://leanpub.com/openintro-statistics. A printed version is available through Amazon for a low price as a black and white paperback, and at a higher price as a full-colour hard cover book.

To summarize this review in a few words I can say that I wish I had studied statistics with a book like this, as I feel that many of the insights about data analysis that took me several years to acquire through practice and reading assorted literature have been distilled into this introductory statistics text book. This book is engaging and clearly written, and teaches what is really important, and includes enough information about modern approaches, such as re-sampling, to give a solid foundation for further study of modern statistical methods.

References

Diez, D. M., C. D. Barr, and M. Çetinkaya-Rundel (2015). *OpenIntro Statistics*. 3rd ed. OpenIntro, Inc. 436 pp. ISBN: 978-1943450053.

OpenIntro.org declares the following aims: 'The mission of OpenIntro is to make educational products that are free, transparent, and lower barriers to education.' Please, visit http://www.openintro.org/ for up-to-date information.

Editorial-board-reviewed article.
Published on-line on 2017-02-01.
Edited by: T. K. Kotilainen.

Appendix: reproduced figures

Two figures out of the 213 in the book (Diez et al. 2015) are reproduced below under Creative Commons BY-SA license 3.0.

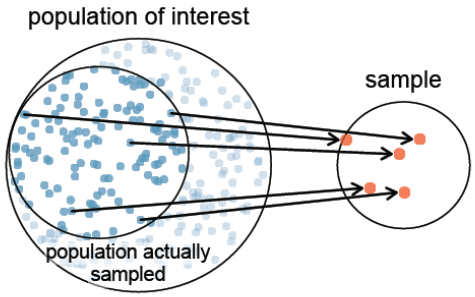


Figure 1.13: Due to the possibility of non-response, surveys studies may only reach a certain group within the population. It is difficult, and often times impossible, to completely fix this problem.

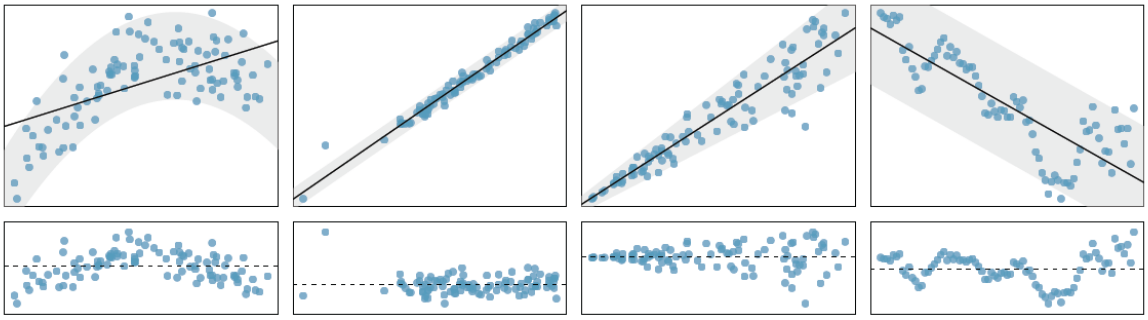


Figure 7.13: Four examples showing when the methods in this chapter are insufficient to apply to the data. In the left panel, a straight line does not fit the data. In the second panel, there are outliers; two points on the left are relatively distant from the rest of the data, and one of these points is very far away from the line. In the third panel, the variability of the data around the line increases with larger values of x. In the last panel, a time series data set is shown, where successive observations are highly correlated.