Journal of Official Statistics, Vol. 28, No. 3, 2012, pp. 463–468

Book Reviews

In order to secure more complete coverage of books on survey methodology and official statistics, JOS is soliciting volunteer book reviewers. Those interested may contact the Book Review Editor Jaki McCarthy (jaki_mccarthy@nass.usda.gov) with a brief description of their area(s) of expertise. Readers and publishers who would like to suggest books for JOS to review are likewise welcome to contact Jaki McCarthy or send the books to the postal address below. – Editors-in-Chief.

Books for review are to be sent to the Book Review Editor Jaki S. McCarthy, USDA/NASS, Research and Development Division, Room 305, 3251 Old Lee Highway, Fairfax, VA 22030, U.S.A.
Email: jaki_mccarthy@nass.usda.gov

Peter J. Huber. *Data Analysis. What Can be Learned from the Past 50 Years*. New York: Wiley, 2011 ISBN 978-1-118-01064-8, 234 pp, $110USD.

"I have no data yet. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."

[Sherlock Holmes to Watson: Sir Arthur Conan Doyle; A scandal in Bohemia (1891)]

The fact that the title Huber had first in mind for this book was "Prolegomena to the Theory and Practice of Data Analysis" gives a hint on the nature of the book itself: this is not a manual of statistical methods, it is rather an account of Huber's "philosophy of data analysis" (p. xiii). This is a text concerned with the issues of the overall procedure of data analysis, with pitfalls, and with some statistical methods. Indeed, Huber intends to convey the principles of data analysis rather than a comprehensive overview of statistical techniques: "My principal aim is to distill the most important lessons I have learned from half a century of involvement with data analysis, in the hope to lay the groundwork for future theory" (p. 3). The premises are intriguing given Peter Huber's renowned experience in data analysis. He is a preeminent authority and has laid the ground for the concept of robust statistics – his 1981 book on Robust Statistics (Huber, 1981) is a classic and a pioneering work – and has provided relevant contributions to computational statistics in the past decades.

To the fundamental question "How do you learn data analysis?" Huber answers with the following citation from Box (1990): "You do not learn to swim from books and lectures on the theory of buoyancy" and, as a consequence, the approach followed in the book is that of using anecdotes and case studies. Indeed, the way in which I myself have learnt how to swim the troubled waters of data analysis is by looking at a few

masters confronting themselves with real data. In a similar attitude, Huber opens his book with a chapter in homage to John Tukey. In particular, it is devoted to a description of the path of statistics in the past 50 years, reporting Huber's impression of how the state of statistics has developed in the five decades after Tukey's paper "The future of data analysis" (Tukey 1962).

The structure of the book is as follows. After this first introductory chapter, there are four central chapters – Chapters 2 through 5 – based on four workshop lectures held by Huber in the past years dealing with strategy, massive data sets, computing languages and statistical models, respectively. There are overlaps between the four chapters that come from allowing for the possibility of reading each of them independently. Then Chapter 6 is devoted to pitfalls, Chapter 7 to some multivariate analysis tools and a final chapter on further case studies and particular statistical and numerical methods.

Chapter 2 follows a particularly effective comparison of data analysis with tactics and strategy in war: the former is for battles, while the latter for campaigns. Here the focus is not on single statistical techniques, but rather on the whole strategy, on a holistic view of data analysis. Then, given that the latter ranges from planning the data collection to presenting the conclusions of the analysis, a useful warning is that "The war may be lost already at the planning stage" (p. 21). This is the chapter where many readers of JOS will feel at ease, because particular attention is drawn on planning the data collection, choosing the appropriate design method, avoiding systematic (measurement) error: "Quantity never compensates for quality" (p. 23).

Chapter 3 deals with the very actual issue of developing statistically sound techniques to analyze massive data sets. I do agree with Huber's view on data mining, that is not particularly benevolent: "By the late 1990's, Data Mining became the fashion word of the decade and touted as a cure-all for the problems caused by data glut. [. . .] Most of the so-called data mining tools are nothing more than plain and simple, good old-fashioned methods of statistics, with a fancier terminology and in a glossier wrapping" (p. 38). However, Huber notes that massive data sets differ from smaller ones ontologically, not only by size: they are not just more of the same, they have to be larger and are in general much more heterogeneous. For this reason, the analysis of large data sets either begins with task and subject matter specific, complex preprocessing, or by extracting systematic subsets on the basis of a priori considerations, or a combination of the two. Therefore, tools for preprocessing massive data sets are those of real interest: "Data analysis is a detective work. [. . .] After perusing some of the literature on data mining, I have begun to wonder: too much emphasis is put on futile attempts to automate non-routine tasks, and not enough effort is spent on facilitating routine work" (p. 55).

Very much linked with these issues are those covered in Chapter 4. It provides a sort of review of computer languages for data analysis, in the search for the optimal one: "An ideal system should cover the full range smoothly, from launching canned applications, to improvising new applications, and to canning them" (p. 62). Huber provides checklists and seems often to be describing the features of R, but indeed never mentions it.

Chapter 5 is the core of Huber thoughts on statistical modeling. Not by any chance the title is Approximate Models: "the header of this chapter is an intentional pleonasm: by definition, a model is not an exact counterpart of the real thing, but a judicious approximation." (p. 90). He introduces the discussion about the role of statistical models

by describing the historical development of the mathematical model for planetary motion. Kepler introduced the fundamental new model of elliptical motions after 1500 years of epicyclic modeling because he could not obtain an adequate fit of his data on planet motion. This sets the example for all statisticians that the method is driven by the object under study and, therefore, to avoid the temptation to squeeze one's data into a wrong (but well-established and preconceived) model class. In other words, it reminds me of the well known statement that electric lamps were not invented by improving candles.

The discussion of the concept and the use of goodness of fit of a model pervades the whole chapter and sets the basis also for a comparison between the frequentist and the Bayesian approach, that, I am sure, will leave many statisticians using the latter unsatisfied with the sharp statement that "Bayesian statistics lacks a mechanism for assessing goodness-of-fit in absolute terms" (p. 92). On the other side, I do agree when it is highlighted that, in general, treating a model that has not been rejected as correct can be misleading and dangerous. In fact, in real-life situations the interpretation of the results of goodness-of-fit tests must rely on judgement of content rather than on p-values only. In a world made of increasingly larger and more complex and structured data sets, Huber shifts the perspective from that of model validity to that of model adequacy, where it is important to assess the quality and the interpretation of the fit.

Chapter 6 deals with three pitfalls that Huber deplores that are overlooked in much statistical literature. The first is the Simpson's paradox, that is handled, however, very quickly by using two examples based on synthetic data. The second faces unrecognized missingness. This is an issue that many readers of JOS are acquainted with, although here it is dealt with using two examples from astronomy and physics that provide a quite different perspective to it. Finally, some warning thoughts on uses and misuses of linear regression are provided.

Chapter 7 proposes the use of multivariate analysis techniques to create order in the data by providing qualitative intuitive insights with the help of interpretable graphical layouts. The focus is on principal components methods based on singular value decomposition and on multidimensional scaling and its similarities with correspondence analysis. The methods are sketched and then applied to a set of intriguing examples from history, astronomy and Latin poetry. Chapter 8 provides more case studies on non-standard examples of dimension reduction through nonlinear local modeling (also via splines) and comparison of spatial point configurations. The end of the chapter provides some simple and clear notes on numerical optimization. Note that the statistical techniques considered in these two chapters are limited and targeted to the examples considered. One could argue that there could be many more methods and possibly more up-to-date (it is hard to find in the book a non-Huber reference from the 2000's). However, I believe that the whole point of the methods presented (in line with that of the whole book) goes beyond the particular method employed and is that of conveying an overall approach to data analysis.

Two characteristics in my opinion permeate all the chapters of the book. First, Huber is mostly known for his works on robustness. One does not find a survey of them in this book, however, the concept of robustness pervades the whole book and, in particular, the discussion on the role of models in data analysis: "For the sake of robustness we may sometimes prefer a (slightly) wrong to a (possibly) right model" (p. 96). Second, in many

pages of the book his dislike of canned computer procedures that took over much of the tactical use of statistical tools becomes apparent. I particularly appreciated the following sentence: "The human act of creating subjective order by separating the wheat from the chaff, and throwing out the latter, also creates insight, and no machine can (or should!) replace the creative process going on in the human mind" (p. 36). Reading this book is very different from reading the other books on statistical methods: it is more like taking Huber's classes on data analysis. I would recommend it to any statistician willing to be provoked by and confronted with Huber's views on statistics and on its new challenges.

References

Box, G. (1990). Applications in Business and Economic Statistics: Some Personal Views. Comment Statistical Science, 83, 390–391.

Huber, P.J. (1981). Robust Statistics. New York: Wiley.

Tukey, J.W. (1962). The Future of Data Analysis. Annals of Mathematical Statistics, 33, 1–67.

M. Giovanna Ranalli
Dipartimento di Economia, Finanza e Statistica,
University of Perugia, Italy,
06123 Perugia, Italy.
Email: giovanna.ranalli@stat.unipg.it

Steven K. Thompson. *Sampling.* Third Edition. Hoboken, NJ: John Wiley & Sons, Inc., 2012. ISBN 978-0-470-40231-3, 393 pp, $115USD.

Now in its third edition, Thompson's *Sampling* textbook provides a comprehensive introduction to basic and advanced topics in survey sampling and design-based inference. The intended audience for this review includes instructors teaching advanced-undergraduate or graduate-level courses on research design or applied researchers trained in statistics seeking texts for self-learning. I have used earlier editions of Thompson's textbook for a PhD-level course on social science research design. For these and similar purposes, I think it is the best textbook on sampling among those currently in print, although there are some limitations that I note below. The text is too advanced for beginning undergraduates, and likely inadequate for statisticians or methodologists seeking an advanced discussion of design-based inference. For the latter, the reader would likely prefer Särndal et al. (1992) or the monograph by M.E. Thompson (1997).

The textbook works primarily within the classical randomization- or design-based paradigm. The design-based paradigm treats outcome data as fixed with arbitrary distribution over a target population, with stochasticity entering the picture only through the non-deterministic sampling of units from this population. The appeal of the design-based approach is that it develops principles for estimation and inference that are applicable in any substantive domain. For this reason, Thompson's textbook is equally relevant for researchers in the natural and social sciences. Examples in the textbook range

from problems in ecology to actuarial studies to household surveys. Such diversity helps students to appreciate essential concepts and to think creatively about research design problems. On occasion the text applies model-based principles when it is illuminating or especially convenient to do so. This I think is a reasonable reflection of current practice in survey sampling.

In its coverage of sampling designs, Thompson's textbook has remarkable breadth. The first half of the book develops principles for canonical problems in survey sampling: simple random sampling, unequal probability sampling, covariate-assisted design, and clustered, stratified, or multistage designs. The presentation in the first half is highly succinct, perhaps to a fault. What distinguishes this book from other sampling textbooks is the second half, which covers advanced methods for hard-to-detect populations, spatial sampling, and adaptive sampling. Much of the material here draws on Thompson's own research, which has been highly imaginative in extending the reach of design-based methods. My students have responded to these sections with delight.

Most remarkable about the new edition is Thompson's addition of worked examples coded for the R statistical computing environment. The examples work with estimators programmed "from scratch," rather than relying on pre-programmed routines. This is welcomed. It provides students with a second take on the form of estimators and allows them to comprehend the anatomy of estimation problems. The examples also contain simulations to illustrate crucial, but difficult concepts such as finite population central limit theorems. R is arguably the most intuitive environment for first-time statistical programmers and it is increasingly the environment of choice for statistical programming in the social and natural sciences. These additions greatly enhance the value of the textbook.

Thompson's textbook requires supplementation for an effective course. After reviewing dozens of sampling textbooks for my graduate course, I had considered Thompson's textbook closely alongside the textbook by Lohr (2009) as well as classic works by Cochran (1977) and Särndal et al. (1992). I decided that Lohr's textbook moved too slowly for a graduate class, while the Särndal et al. textbook went well beyond what could be grasped for students new to survey sampling. What I have settled on is Thompson's book supplemented by sections from Cochran (1977) that cover the canonical problems listed above. Unfortunately, Thompson's coverage of these problems is too succinct for students who are coming to design-based inference for the first time (who have typically been trained in regression and model-based inference). Thompson tends to treat the analytical derivation of estimators, biases, and variances secondarily. Final expressions are summarily displayed in the main text and then derivations are relegated to appendix-like sections at the end of chapters. Unfortunately, this manner of presentation promotes a sense that estimators and their properties are to be learned by rote rather than understood. This is in contrast to Cochran (1977), who integrates such derivations into the flow of the text, allowing the derivations to incrementally reveal the logic of design-based inference. My experience suggests that Cochran's manner of presentation is far more effective and that students appreciate the detail. In addition, Cochran includes many real data examples, whereas Thompson's examples tend to be highly simplified abstractions. These limitations could be addressed in future editions. Nonetheless, Thompson's textbook is, in my opinion, the best single volume among all options currently in press and is therefore highly recommended.

References

Cochran, William G. (1977). Sampling Techniques, Third Edition. New York, NY: John Wiley & Sons.

Lohr, Sharon (2010). Sampling: Design and Analysis, Second Edition. Boston, MA: Brooks/Cole.

Särndal, Carl-Erik, Swensson, Bengt, and Wretman, Jan (1992). Model Assisted Survey Sampling. New York, NY: Springer-Verlag.

Thompson, M.E. (1997). Theory of Sample Surveys. New York, NY: Chapman & Hall.

Cyrus Samii
New York University,
Department of Politics, 19 West 4th Street,
New York, NY 10012, USA.
Phone: 1-212-998-8500
Email: cds2083@nyu.edu