JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 523, 1391–1394
https://doi.org/10.1080/01621459.2018.1513232

Taylor & Francis
Taylor & Francis Group

Observation & Experiment: An Introduction to Causal Inference. Paul R. Rosenbaum. Cambridge, MA: Harvard University Press, 2017, xvi+374 pp, $35.00(H), ISBN: 978-0-67-497557-6.

Observation and Experiment, by Paul Rosenbaum, lives up to its subtitle: it provides an excellent *Introduction to Causal Inference.* Using language and concepts accessible to any non-math-phobic layperson, it explains the special difficulties and statistical techniques of causal inference in (primarily) observational contexts. This text provides a broad survey of the ideas in the field, patiently explaining concepts in simple, accessible language. The book is a well-written and thoughtful reflection on the doing of causal inference from one of causal inference's noted experts.

Rosenbaum's central focus is on how to build a case for a causal effect in a world where you can never see a perfect counterfactual. The underlying framework is of course that of potential outcomes, familiar now to most statisticians. In this framework, he argues that evidence for any but the most clear-cut causality never rises above the level of "reasonably compelling" (p. 107), so several of the chapters focus on ways to build a case by strengthening or layering different lines of evidence.

The book begins with the gold standard of causal evidence: the randomized controlled study. But this material is really just laying the groundwork for how we will think about observational studies later on. It is integral for serving the larger agenda of what makes evidence; the chapter is there to provide a clear description of the aspirational goal of the efforts that come after. To underscore this point, consider that Part I (on randomized experiments) is around 50 pages and the rest (on observational studies) spans more than 200. But this foundational effort is important: in Part I, Rosenbaum focuses primarily on the concept of balance (and why randomization provides it) and on the logic of permutation tests, both critical concepts on which the remainder of the book is built.

The RCT dispensed with, Rosenbaum then moves on to write one chapter each on natural experiments, the benefits of more-elaborate theories with multiple predictions, quasi-experimental devices (e.g., control outcomes), sensitivity analysis and design sensitivity, matching techniques, techniques for correcting for bias due to general disposition, and, finally, instruments/instrumental variables. In each case, attention is focused on developing a solid intuition for the underlying logic of a technique or approach, rather than on developing practice-ready recipes for using the technique. In each case there are real examples that motivate the methods and the thinking.

And it is the thinking that is the key focus: this text encourages thinking deeply about the doing of statistics. Rosenbaum talks about Popper, even Wittgenstein, and ponders what constitutes evidence. For instance, Chapter 7, devoted to elaborate theories, discusses why we should seek multiple, statistically independent predictions, so that, in the words of Mervyn Susser, "evidence is strengthened when diverse approaches produce similar results." In this chapter (and some others) we have abandoned any classic conception of what a statistics textbook should look like, and are instead receiving a lesson on how to do good thinking, how to avoid pitfalls and strengthen evidence.

While the book is very broad about the patterns of thinking, it can be arguably narrow about the statistical techniques discussed and presented. Some of the presented ideas are quite unusual and not well known, a surprising feature if one thinks of this as a general introduction to causality. But these ideas do deserve their space. They underscore the larger point that one way to obtain good evidence is to be very clever and deeply attentive to confounding; see, for example, the presentation of his idea of correcting for "biases due to general dispositions" (Rosenbaum 2006) in chapter 12. Here, he suggests that (for example) in a study of bicycle injuries, to find a matched control for a helmetless cyclist, we should prefer to match helmetless cyclists who text while driving to helmeted cyclists who do not in order to get a population with comparable underlying levels of cautious behavior, despite differing on both specific behaviors measured. This is equivalent to estimating the treatment effect difference between the observational "treatment" of wearing a helmet and that of avoiding texting while driving; both indicate a general level of caution, but only the helmet has the direct physical effects in which the researcher is interested. It is true that propensity-score-based methods might perform this correction "naturally," but only if the researcher explicitly measures other reckless behaviors like texting and includes them in the propensity model. This articulation of what the appropriate match is is quite deep. What makes the overall chapter even more stimulating and enjoyable for those who appreciate the power of statistics is that, in order to formalize these concepts in a statistical model, he repurposes the Rasch model from measurement theory.

Much of the text covers concepts familiar to many statisticians, but sometimes not in the way they might expect. For example, pages 162 through 166 cover the logic of difference in differences, but as with much of Rosenbaum's writing, these core ideas are presented in his own way before circling back to tie these ideas to classic or more widespread conceptions. Due to his clear articulation, in this case, the difficulties of how measurement scale of the outcome can contaminate a diff-in-diff approach is made abundantly clear, as we are shown how parallel trends are both an aspect of the world and an aspect of how we measure it. This danger is often ignored or missed in the usual treatment of this method, and is certainly not as clear to understand under classic views of this technique. For Rosenbaum to have an alternative vantage point is quite common; for those with prior exposure to what he is discussing, his take can be a prompt to reflect on what one does, but it can also be disorienting. (See, also, his discussion of instruments.)

While his vantage point may feel odd, it does not come from caprice. Let no one say Rosenbaum does not have a sense of what is right. This is amusingly captured, for example, with note #1 on page 336, where—buried in one of the 59 pages of endnotes, references, examples, and comments that supplement the main body of the text—he provides a sharp critique of fixating on the database "variable" part of "instrumental variable" rather than the "instrument" itself, acting in the world. In his opinion, we should use only "instrument" when talking about instruments because that maintains focus on the core concerns: whether there is in fact a haphazard nudge towards treatment that allows insight into the impact of treatment. We deeply appreciate this commitment to the meaning of what we do. Down to the very words we use, Professor Rosenbaum creates a system for causal analysis that encourages this focus.

But this does make the book narrow in some sense. In the causal inference world there are many debates about different approaches to estimating causal effects (e.g., using directed acyclic graphs to represent causal systems, or inverse weighting rather than matching). With propensity scores in particular there is a divide between using matching or inverse weighting. Rosenbaum lands strongly on the matching side, and perhaps one reason why is its deliberate construction of a comparison group, an activity which lays bare all of one's assumptions in such a manner as to make them tangible, easily understood and communicated. Nevertheless, in this work it appears as though Rosenbaum presents only what he thinks is right, and thus many things (such as weighting) are left out. Even the much used and often maligned workhorse of what he calls "analytical methods to adjust for observed covariates x" (presumably, regression-based adjustments) is only obliquely mentioned, being briefly compared to "buying a house without seeing it, being offered instead an audio recording of the sounds of hammering and sawing" (p. 217). Rosenbaum then adds, making his preference for matching techniques over this approach clear, "With so much fine hammering and sawing like that, how could the house be less than perfect?" (p. 217).

Perhaps the toughest thing about writing such a technical work for a popular public is finding the right tone for the audience. Is this book for laypeople? In that case, if, as Stephen Hawking's publisher told him while editing *A Brief History of Time* (Hawking 1996), each equation in such a book will halve the sales, then this book's sales will not be good. Rosenbaum does not shy away from including equations, although he is careful to first introduce each concept in intuitive, nonmathematical terms, and to ensure that no more than high school algebra is necessary. Sometimes, however, the explanation is broken down so completely that we worry that people may get lost before the end is reached. Or perhaps this book is for people already knowledgeable in statistics? In that case, they would likely find the lengthy explanations of common concepts, however well-written and leavened by appropriate concrete examples, to be tedious. In the end, Rosenbaum's lively writing style does as good a job as possible at threading this needle.

Regardless, the overall structure of the book is quite pleasing. Each chapter ends with a "Taking Stock" section summarizing what was accomplished. There is a glossary and index, and they are both of real use, not an afterthought. Most causal concerns one has can be easily looked up, and when you find the relevant page you will find something subtle to think about, with good references for further reading. That being said, there is possibly a gap between the raw academic nature of many of the references, and the extremely accessible prose of the text. Depending on who is reading the book, they might experience quite a shock when trying to follow up some of these ideas as the gap is at times considerable.

The ideal audience for the book would therefore probably be a researcher with a basic statistical background, working in a concrete field such as education or health in which causal inference is needed. Such a researcher could use this book as a roadmap to the deeper concerns behind causal inference, absorbing concept-specific chapters one at a time as needed. With a solid conceptual understanding of a given chapter's topic, the researcher could then easily turn to the suggestions for further reading at the end for technical details. Each chapter is capably summarized in a few short paragraphs at the end, so by the time this hypothetical researcher had read all of them, they would have a quick and easy reference for reminding themselves of any details they later forgot.

Considered as a possible undergraduate textbook, this book is probably too lengthy to complete in a one-semester course, but would provide an excellent additional resource, both for students wanting to learn more, and for a professor looking for inspiration in translating sophisticated ideas into simple language. The many excellent examples, taken primarily from Rosenbaum's long history of collaborative work with many different scientists, all come with meaningful discussion of the importance of the studies themselves, and are a rich resource. If nothing else, this book provides a great collection of stories, a series of compelling statistical investigations and a constellation of concepts about doing the work of causal inference.

Though these reviewers are probably not the books primary intended audience, we found its grounding of each technique in solid intuitive logic, as well as the selection of interesting examples, to be quite engaging. We also found the deeper philosophical points regarding the activity of doing causal inference to be quite thought-provoking, a series of lovely prompts to reflect on our own understandings of causality. Few will be wasting their time by reading this text. We certainly did not.

References

Hawking, S. (1996), *A Brief History of Time*, New York: Bantam Books. [1393]

Rosenbaum, P. R. (2006), "Differential Effects and Generic Biases in Observational Studies," *Biometrika*, 93, 573–586. [1392]

Jameson A. Quinn and Luke W. Miratrix
Harvard University

Repeated Measures Design With Generalized Linear Mixed Models for Randomized Controlled Trials.

Toshiro Tango. Boca Raton, FL: Chapman & Hall/CRC Press, 2017, xv+359 pp., $93.95(H), ISBN: 978-1-49-874789-9.

This book provides a summary of generalized linear mixed models for the setting of randomized clinical trials (RCTs) with repeated measurements of outcomes. Toshiro Tango presents an S:T repeated measures design which can accommodate RCTs with more than one baseline measurement. Chapters 1–5 review naïve pre–post analysis, ANOVA models, and ANCOVA models. In Chapter 6, missing data in the context of longitudinal studies is discussed. Chapters 7–9 provide in-depth analysis and sample size calculation for continuous, logistic, and Poisson mixed models, including SAS code and output. The remaining