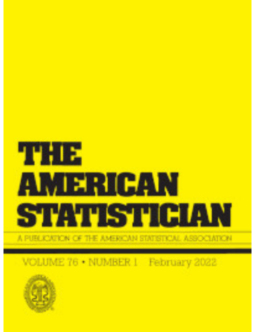

THE AMERICAN STATISTICIAN

A PUBLICATION OF THE AMERICAN STATISTICAL ASSOCIATION

VOLUME 76 • NUMBER 1 February 2022

The American Statistician

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

The Phantom Pattern Problem: The Mirage of Big Data,

by Gary Smith and Jay Cordes. Oxford, United Kingdom: Oxford University Press, 2020, vii + 227 pp., $32.95(H), ISBN: 978-0-19-886416-5

Emilija Perković

Chapters 2 and 3 respectively discuss linear and nonlinear dimension reduction techniques. Dimension reduction aims to represent a high-dimensional dataset with a low-dimensional one, which expedites subsequent analyses and allows for visualization in 2-D or 3-D. For linear dimension reduction, the book covers principal component analysis, singular value decomposition, nonnegative matrix factorization, factor analysis, linear discriminant analysis, random projection, and intrinsic dimensionality estimation. For nonlinear dimension reduction, multidimensional scaling, manifold learning, and neural network-based approaches (including autoencoders) are introduced.

Chapter 4 presents an interesting topic on data tours. The idea of data tours is to view a dataset from multiple perspectives. This is in contrast to dimension reduction techniques where only one view is typically returned. The chapter includes several methods for data tours: grand tours, interpolation tours, projection pursuit, and independent component analysis.

Chapters 5 and 6 respectively cover algorithm-based and model-based clustering methods. The goal of a clustering analysis (sometimes also known as unsupervised classification) is to divide data into several homogeneous groups (called clusters). The identified clusters may be directly interpretable in certain contexts and/or may be useful for subsequent supervised learning tasks. For algorithm-based clustering, the book introduces hierarchical clustering, k-means, spectral clustering, document clustering, and minimum spanning tree-based clustering. For model-based clustering, the focus is on mixture models, their estimation procedure, and their utility in density estimation and discriminant analysis. In addition, Chapter 5 also covers various metrics to assess clusters.

Chapter 7 discusses several smoothing techniques including loess, smoothing splines, and bivariate smoothing. These techniques are useful in visualizing the relationships between pairs of variables.

Part III of the book contains four chapters covering EDA as visualization.

Chapters 8–10 introduce various methods for visualizing clusters, univariate and multivariate densities, dependencies, summary statistics, and dimension reduction, with the primary focus on continuous data.

Chapter 11, by contrast, covers visualization for discrete data. Both univariate and multivariate approaches are discussed; the focus of the latter is on bivariate cases (i.e., two-way contingency tables) but references for general multivariate cases are provided at the end of the chapter.

Overall, the book is very easy to follow and I enjoyed reading it. All the companion codes are available at the CRC website *https://www.crcpress.com/9781498776066*, which makes it easy for readers to implement the EDA techniques covered in the book to their own datasets. I believe that the book can benefit students as well as researchers from many disciplines. It can also be a good reference book for statisticians and data scientists who routinely use MATLAB. As mentioned by the authors in the Preface, their next task is to write an R Shiny app implementing many of the methods in this book. I look forward to this future contribution that can benefit an even broader audience.

Yang Ni
Texas A&M University

The Phantom Pattern Problem: The Mirage of Big Data, by Gary Smith and Jay Cordes. Oxford, United Kingdom: Oxford University Press, 2020, vii+227 pp., $32.95(H), ISBN: 978-0-19-886416-5

The *Phantom Pattern Problem* is a timely and well-written submission dealing with the growing issue of pattern misattribution in the era of big data. As the authors point out, we people are pattern-finding creatures. Often, we jump to assigning meaning to patterns since a pattern in itself appears to contain meaning. Furthermore, more patterns are waiting to be discovered than ever before in the era of big data!

This book is written for a reader with limited background knowledge in statistics but who is interested in applying data mining and machine learning methods to some aspect of their life and work. The manuscript is full of humorous and insightful examples of pitfalls and flawed reasoning resulting from our desire to make sense of the world. The book's introduction examines these main points through an excellent "correlation is not causation" example.

This book consists of an introduction, nine chapters, and an epilogue. Besides the introduction and the epilogue, all chapters end with a section titled: "How to avoid being misled by phantom patterns." This section presents a review of the main points of the chapter and summarizes the key takeaways.

Chapter 1 examines the history of the natural human inclination of pattern finding. The review of several historical and contemporary examples helps explain why this desire to order the world is so profoundly human and emphasizes the need for the authors' submission. Chapter 2 explains, in simple terms, why correlation does not imply causation and the steps one needs to take to show causation in practice (randomized experiments, i.e., A/B testing). Chapter 3 focuses on the concept of confirmation bias and various ways we rationalize away apparent issues when they do not fit with the pattern we are trying to see.

Chapter 4 examines the pattern of randomness. More concretely, this chapter describes the difficulty we humans have in perceiving and accept randomness. Randomness often does not look as orderly as we would expect, and we often perceive truly random events as exhibiting a sort of pattern.

Chapter 5 describes the perils of data mining, or as Ronald Coase and the authors put it, data torturing. Given all the data scientists' methods at our disposal, making the data confess to what we want to hear is tempting. One section of the chapter looks at data mining Trump tweets; since this article has been submitted, the former president has lost access to Twitter and had his account. So this section likely requires some minor revision to ensure the language used is up to date.

Chapter 6 looks at the vast amount of data available for free online and the temptation to use this information to, for example, make stock market predictions. I found the running stock market examples particularly compelling as they often built on each other. This connection allows the reader to become fully immersed as the illustrating example does not require us to jump from topic to topic. Though I found the examples in this chapter particularly compelling, the topic covered matches very closely to Chapter 5. One suggestion to the authors is to use the last section of these two chapters to emphasize the different takeaways of the two chapters more strongly.

Chapter 7 covers the reproducibility crisis that is still shaking the social and natural sciences. This is a crucial chapter for this book as it deals with the consequences of sloppy and bad science described previously in Chapters 5 and 6. Chapter 8 builds on the previous chapters by describing even more instances of bad or careless data science. I do think that the examples covered in this chapter are critical and illuminating. However, it would be great to emphasize the differences between examples covered in this chapters and those in Chapters 5 and 6.

Chapter 8 circles back to examples of using observational data to make conclusions about cause and effect as well as the gold standard of a randomized experiment. I thought these reminders of good practices were a great way to end the book as they leave the reader with a clear example of how to conduct a good data analysis or experimental design.

Finally, the authors include an epilogue focusing on Bayes' theorem. Including Bayes' theorem in the epilogue was a prudent decision as it is a pretty complex concept. I also enjoyed the historical remarks and thought the authors managed to explain a complicated topic in a digestible way.

Overall, my thoughts on this submission are very positive. From a reader's perspective, I found the book to be delightful. I think it provides an excellent introduction to the importance of quality data science and statistics. I also found the examples about the authors' own statistical research adventures to be very personable. This is a beneficial book that undoubtfully has a potentially large readership base.

Emilija Perković
University of Washington

Statistics for Making Decisions, by Nicholas T. Longford. Boca Raton, FL: Chapman & Hall/CRC Press, 2021, xv+292 pp., $120.00(H), ISBN: 978-0-36-734267-8

Decision-making is a ubiquitous activity in our everyday lives, and plays a crucial role in science, business, and governance. One of the core tasks of applied statisticians is to advise stakeholders in various fields, such as medicine, finance, or education, with the goal of improving their decisions in the face of uncertainty. Nicholas T. Longford's book *Statistics for Making Decisions* caters to a statistically well-versed audience interested in approaching common statistical problems from a new decision-centered perspective. The author criticizes conventional hypothesis testing procedures, and proposes to replace them with cost-benefit analyses that take sampling uncertainty, as well as clients' individual loss functions into account.

The book comprises 11 chapters. The first five chapters establish basic concepts and describe the statistical decision-making framework that the author puts forward. The framework is introduced through an example case of determining whether parameter values are positive or negative; first for normally distributed estimators (Chapter 3), then for statistics with non-normal distributions (Chapter 4). In Chapter 5, the framework is extended to scenarios with multiple decision options, particularly in relation to verdicts about the size of effects. The following chapter is dedicated to experimental design. In addition to a discussion of sample size determination as a function of costs and expected knowledge gain, the chapter also thematizes the importance of avoiding a decision impasse. Chapters 7–10 provide examples for practical areas of application, such as medical screening, sequential decision making, performance rating of institutions, and clinical trial evaluation. The final chapter of the book presents model selection in the context of linear regression from a decision-theoretic perspective.

The most striking feature of the book is its bold subjectivist approach that unapologetically puts the client at the center of all statistical endeavors. The author emphatically denies the usefulness of default statistical procedures, and urges the reader to find tailored solutions for specific application scenarios. In fact, the proposed decision-making paradigm is impossible to apply without the elicitation of a customized loss function assigning costs to different kinds (and magnitudes) of errors. The author acknowledges that the exact form of these loss functions might be difficult to elicit in practice, and that different stakeholders may disagree about the costs associated with certain errors. However, he contests the idea that it may be impossible to assign costs to decision outcomes. The solution he provides is conducting sensitivity analyses using a range of plausible values (e.g., Chapters 3.7, 7.1).

In my view, one potential drawback of the book is that it does not provide a discussion of practical methods for the elicitation of loss functions despite their evident importance for the approach. Value judgements of immaterial outcomes can be highly controversial, and fear of controversy or dissent among stakeholders might keep readers from applying the proposed paradigm in practice. However, fields such as health economics and actuarial science are well known for complex value judgments and their incorporation in mathematical models. It would have been interesting to see some compelling examples for loss elicitation from these or other fields that can provide the reader with the confidence that it is possible to reach agreements about value judgments in practice.

An interesting aspect of the book is that it is not committed to the Bayesian or frequentist paradigm. Expected losses are computed based on fiducial or posterior distributions that are combined with elicited cost functions. Given the strong focus on informed client-centered analyses in the book, it came as a surprise to me that the author seems reluctant to promote the incorporation of clients' prior knowledge in Bayesian parameter estimation via the prior distribution. In my view, the ease of integrating prior knowledge would have been a