Marina Dobrota
University of Belgrade, Faculty of Organizational Sciences, Serbia

BOOK REVIEW

DOI: 10.7595/management.fon.2018.0028

Abstract:

Book review of: *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science* by Bradley Efron and Trevor Hastie**,** Cambridge University Press, 2016, Series: Institute of Mathematical Statistics Monographs (5), 495pp., ISBN13: 9781107149892, ISBN10: 1107149894, Online ISBN: 9781316576533, DOI:10.1017/CBO9781316576533

As many other scientific fields and disciplines today, statistics has been under a great and significant influence of digitization and ICT. In this aspect, Bradley Efron and Trevor Hastie, professors at the Department of Statistics, Stanford University, California, have published a book that very closely describes the state of present days in terms of modern statistics, *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. As described by the authors on the very first page of a preface, the discipline of statistics has a 250-year old history (since 1760s) and we are now in its most recent quarter (1950s to the present). This quarter is the *computer and digital age*, "the time when computation, the traditional bottleneck of statistical applications, became faster and easier by a factor of a million" (Efron& Hastie, 2016, p.xv). The book explores how statistics has evolved during the past years, under the influence of computers and digitalization, giving us, as authors please to say: "…an aerial view of a vast subject, but seen from the height of a small plane, not a jetliner or satellite"(Efron& Hastie, 2016, p.xv).

The publication covers a broad set of topics regarding the statistical methods and statistical inference, examined from multiple viewpoints. Efron and Hastie observe these topics through the frame of electronic computation, which is, as they refer to it, central for their book. Due to the very nature of statistics, not every topic in the book was computer-related. Nevertheless, an extensive number of topics were the descendants of the computer and digital age, such as bootstrap, proportional hazards, large-scale hypothesis testing, big data, and the machine learning algorithms. Above all, one of the most popular statistical concepts processed in the book is Data Science, "a particularly energetic brand of the statistical enterprise" as authors like to call it. "Algorithms are what statisticians do while inference says why they do them," and the concept of Data Science is "emphasizing algorithmic thinking rather than its inferential justification" (Efron& Hastie, 2016, p.xvi).

The publication consists of three parts. *Part I: Classic Statistical Inference* deals with the typical traditional viewpoint of the inferential theory. The authors explain the need for handling the concept of Statistical Inference, by referring to it as the system of mathematical logic for guidance and correction when interpreting the observations of natural scientists in an attempt to "judge the accuracy of their nature science ideas". The Part I includes the following topics: *Algorithms and Inference*, *Frequentist Inference* (with Frequentism and frequentist optimality), *Bayesian Inference* (withBayesian/Frequentist comparison list), *Fisherian Inference and Maximum Likelihood Estimation* (with Fisher Information, MLE, Conditional Inference, Randomization), and *Parametric Models and Exponential Families* (with Univariate and Exponential Families, Multivariate Normal Distribution, Multinomial Distribution). In this part, Efron and Hastie debate on the strengths and weaknesses of the three prime statistical inference theories: Frequentist, Bayesian, and Fisherian Inference. They make an immense divide and give a detailed comparison of the Bayesian versus Frequentist disputes. An interesting outlook of the authors states the following: "Sir Ronald Fisher was arguably the most influential anti-Bayesian of all time, but that did not make him a conventional frequentist. His key data analytic methods—analysis of variance, significance testing, and maximum likelihood estimation—were almost always applied frequentistically. Their Fisherian rationale, however, often drew on ideas neither Bayesian nor frequentist in nature, or sometimes the two in combination" (Efron& Hastie, 2016, p.38). In this chapter overall, Efron and Hastie encourage the reader to look at the method or a technique from one of the three viewpoints, and then to raise a question and to compare the benefits of each approach.

Dissimilar to restrictions of slow mechanical computation that sculpted classical statistics, the new computational possibilities from the end of the 20th century have opened the door to fresh, more expansive, and useful statistical methodologies. *Part II: Early Computer-Age Methods* consists of nine chapters: *Empirical Bayes*, *James-Stein*

Estimation and Ridge Regression, *Generalized Linear Models and Regression Trees*, *Survival Analysis and the EM Algorithm*, *The Jackknife and the Bootstrap*, *Bootstrap Confidence Intervals*, *Cross-Validation and Cp Estimates of Prediction Error*, *Objective Bayes Inference and MCMC*, and *Postwar Statistical Inference and Methodology*. As authors remark, while "journals of the 1950s continued to emphasize classical themes: pure mathematical development typically centered around the normal distribution, …in the 1990s a new statistical technology, computer enabled, was firmly in place" (Efron& Hastie, 2016, p.75). Chapters in Part II describe major developments from this time period. The ideas of statistical methods of the "postwar era" are not crucially different from those classical methods, but their computational demands, which judging by the authors have grown 100 or 1000 times, undoubtedly are. So, rather than enlarging the classical methods, for example the James–Stein estimator, the jackknife, or the bootstrap merely speak for a different use of modern computer capabilities: "they extend the reach of classical inference" (Efron& Hastie, 2016, p.181).

Further on, the authors get to "the story of statistics in the twenty-first century", where computational demands have grown again "by the factor of thousands" (Efron& Hastie, 2016). In the third part of the book, large-scale inference and prediction algorithms are targeted, such as boosting and deep learning. As opposed to the typical traditional viewpoint, these topics mostly illustrate the Data Science point of view. *Part III: Twenty-First-Century Topics* covers seven chapters: *Large-Scale Hypothesis Testing*, *Sparse Modeling and the Lasso*, *Random Forests and Boosting*, *Neural Networks and Deep Learning*, *Support Vector Machines and Kernel Methods*, *Inference After Model Selection*, and *Empirical Bayes Estimation Strategies*. In the 21stcentury, statistical algorithms continue to pursuit massive data sets, while inferential analysis strives to rationalize the algorithms (Efron& Hastie, 2016, p.271). We now face colossal data sets, both in terms of entities and variables. We were anticipating it for years, and the time has finally come, we are there now: tons of data and the necessity to determine how to use them. "Something important changed in the world of statistics in the new millennium. Twentieth-century statistics … could still be contained within the classic Bayesian–Frequentist–Fisherian inferential triangle (Efron& Hastie, 2016, p.265). This is not so in the twenty-first century" (Efron& Hastie, 2016, p.446). Topics like false-discovery rates, post-selection inference, empirical Bayes modeling, or the lasso, still fit within the abovementioned triangle but others, such as neural nets, deep learning, boosting, random forests, and support-vector machines, are rushing towards the computer science.

Each chapter is concluded with the *Notes and Details* section, the touch that significantly enhances and builds up the publication. They hold the derivation details, links to Frequentist, Bayesian, or Fisherian inference, and comments on historical relevance. Finally, in *Epilogue*, authors give a very short time-line, starting from the 1900s with Karl Pearson's chi-square paper, all the way through to 2016 with modern Data Science and Big Data. It is drafted in the epilogue how the spotlight of statistical progress has converted between Applications, Mathematics, and Computation throughout the 20th and 21st century. As authors like to say, it is intriguing but tricky to surmise on the future of statistics. A demand for statistical analyses is growing incessantly, and as a result the Data Science has boomed. "A hopeful scenario for the future is one of an increasing overlap that puts data science on a solid footing while leading to a broader general formulation of statistical inference" (Efron& Hastie, 2016, p.452).

Authors intended to preserve a technical level of discussion in the publication, appropriate to Masters'-level statisticians or first-year PhD students. Nevertheless, by reading this book, the veteran statisticians can find a compact summary of chronological statistics, students can find a guide to statistical inference that can upgrade the most books of an introductory statistics, and other interested audience can find vast discussions on modern topics in statistics.

Accepted: 2018-11-01

About the Authors

Marina Dobrota
University of Belgrade, Faculty of Organizational Sciences, Serbia
dobrota.marina@fon.bg.ac.rs

Marina Dobrota is an Assistant Professor at the University of Belgrade, Faculty of Organizational Sciences, where she completed her PhD thesis in the field of Operational Research and Computational Statistics. She has published a number of research papers in scientific journals, conference proceedings, or monographs, both international and national. The main subjects that she teaches are the Theory of Probability and Statistics. Her research focus lies in Composite Indicators, Econometric Modeling, and Data Science. Her research interests include Statistical inference, Data Analysis, Data Mining, and Time Series Analysis.