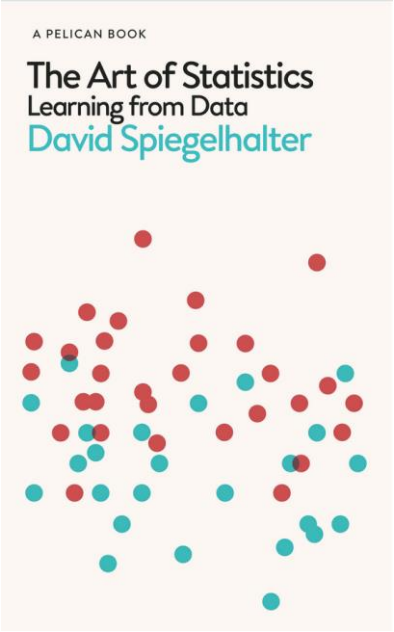
DOI: 10.13140/RG.2.2.28462.46400

Book Review: David Spiegelhalter's "The Art of Statistics: How to Learn from Data" Pelican Books, 2019

Norman Fenton[1], 20 August 2019



A superb, timely overview of the benefits and limitations of statistics in the era of big data and machine learning

David Spiegelhalter (subsequently referred to as DS) has gained a deserved reputation as a masterful communicator of statistics and risk through his media work and writings. I believe this timely book is the best introduction to the benefits and limitations of statistics that I have seen and is DS's most important work yet in public communication. Any of the minor concerns explained below that I have about the book (including the understated role of causal models and the role of the likelihood ratio in courts) are the inevitable result of having to be selective about which more detailed material has to be left out to satisfy both the page and audience constraints.

The book is primarily targeted at lay people but, unlike most popular science books, it manages to explain clearly a range of very challenging topics that even many professional statisticians either do not know about or do not understand. DS does this using highly motivating and engaging examples, some of which run across multiple chapters. For example, Chapter 6 on "Algorithms analytics, and prediction" is the best overview of modern data science and big data analytics that I have seen. DS uses the example of data on Titanic passengers – and the problem of predicting whether or not

[1] Queen Mary University of London, London, UK, E1 4NS, E-mail address: n.fenton@qmul.ac.uk

they would survive based on the various factors (like sex, age, class) known about them - to explain the full range (and limitations) of machine learning and other algorithms as well as to explain important and difficult concepts such as: error rates, performance, ROC curves, calibration, overfitting and cross validation. Hence, in addition to anybody wishing to understand more about risk and statistics, I would recommend this book to the increasing number of students studying 'data science' (at both undergraduate and postgraduate level) and would go so far as to say they should read it before becoming indoctrinated into particular techniques.

The book starts with the example of the use of statistics in identifying unusually poor hospital treatment outcomes and, in particular: "What happened to children having heart surgery in Bristol between 1984 and 1995". The number of 'excess deaths' at Bristol during this period were sufficiently high to merit an enquiry, which seems to have led to general improvements in procedures. As it so happens, I had a slight concern about this example – I felt that the later 2012-2015 data (which demonstrated the overall national improvements) contained an 'outlier' hospital that was relatively equally as 'bad' as Bristol between 1984 and 1995, but which was not discussed. When I raised this concern by email with DS he recognized the issue and also pointed out the risk-adjusted information on this site:

https://childrensheartsurgery.info/data/table

This website (which DS helped design) is a great example of how well statistical and risk information can be communicated using clever interactive graphics i.e. data visualization. Of course, as DS himself suggests several times during the book, there is a limitation to what can be communicated in two dimensional black and white images as he is limited to here. Nevertheless, the multiple images throughout the book generally do a very good job and have clearly been chosen based on years of experience about what works and what does not. A good selection of such graphics appears in Chapters 2 and 3 that demonstrate ways to summarize complex data.

Chapter 4 contains a very good introduction to the difficult but important topic of causation and the limitations of what can be evaluated from observational data alone. Ideally, this is a topic that I feel should play a greater role than it currently does in the book. DS refers to the work of Pearl[2] here (and also in later chapters) that highlights the need for causal models in order to answer questions about interventions and counterfactuals (and hence achieve 'true AI'). However, there is no mention of the basic graphical models that are driving the current 'causal revolution' in statistical and probabilistic analysis and which should precede data analysis. Instead, whereas for example graphical models and Pearl's do-calculus can establish causal effects from observational data, DS focuses on the randomized controlled trial (RCT) (and, later, regression analysis) as 'the' way to establish causal effects. While the example of the RCT results for statin use are very powerful[3], the example addressing the question "Is prayer effective" simply highlights the fact that RCTs may also be very limited. In this particular RCT patients who were about to undergo cardiac bypass surgery were

[2] Pearl, J., & Mackenzie, D. (2018). "The book of why : the new science of cause and effect". New York: Basic Books.

[3] it made me rethink my own objections to taking statins - every statin I have tried have caused severe muscle or joint pain and I considered that the marginal benefits of taking statins were insufficient to balance these side effects.

randomly assigned to three groups: patients in group 1 were prayed for (but did not know it), patients in group 2 were not prayed for (but did not know it), and patients in group 3 were prayed for and knew it. DS reports that:

> *"The only apparent effect was a small increase in complications in the group that knew they were being prayed for: one of the researchers commented, 'It may have made them uncertain wondering, "Am I so sick that they had to call in their prayer team".'*

The problem with this conclusion is that I believe the RCT design was fundamentally flawed because of a failure to consider underlying causal factors: it is certain that religious people are more likely to believe in prayer, but they also believe that such prayer needs to come from themselves or a trusted personal chaplain – not some anonymous person as in the study. The RCT completely failed to recognize this most important feature of religious belief and prayer. To properly answer the original question, religious belief should have been a control factor and prayers should have been administered by a trusted chaplain. Moreover, there should also have been consideration of whether there are any confounding factors for religious belief (for example, people may become religious as a result of having overcome previous serious health problems). Drawing a causal graphical model would have addressed these issues before the RCT was undertaken.

As I would expect of any introduction to causal effects, Chapter 4 contains a discussion of Simpson's paradox and DS covers this brilliantly using the real example of Cambridge University admissions data for STEM subjects in 1996. The concern was that the overall acceptance rate for women on the five STEM subjects was lower than for men. Yet, for each of the five subjects individually, the acceptance rate for women was ***higher*** than for men. Most people find this apparent paradox impossible to believe, but as DS explains:

> *The explanation is that the women were more likely to apply for the more popular and therefore more competitive subjects with the lowest acceptance rates such as medicine and veterinary medicines and tended not to apply to engineering which has a higher acceptance rate.*

DS highlights this as an important reason why we cannot rely on observational data alone for determining causal effects. However, it is also important to note that 'the gold standard' alternative of RCTs also cannot completely overcome the possibility of Simpson's paradox invalidating their results due to a confounder not considered in the study.

Much of the focus of the book is on what would be described as 'classical statistics' (Chapter 5 on regression modelling, Chapter 7 on confidence intervals, and Chapter 8 on frequentist probability). In Chapter 10, DS provides an extensive overview of classical statistical significance hypothesis testing. This chapter includes the very interesting example of the case of Harold Shipman (a GP who was convicted of murdering 215 mainly elderly patients by lethal injection), including the observation that standard statistical hypothesis testing for 'excess deaths' would have flagged an alert on Shipman as early as 1984, saving 175 lives. It is also interesting that the statistical monitoring system for GPs that was piloted after the Shipman enquiry (in which DS was involved) identified a GP with excess death rates even higher than Shipman's. However, it turned out that there was a perfectly reasonable causal

explanation in this case - the doctor practiced in a south-coast town with a large number of retirement homes with many old people, and he conscientiously helped many of his patients to remain out of hospital for their death.

While classical statistical significance hypothesis testing remains the standard method in much of the sciences and social sciences, DS highlights the limitations of this approach including the much-misunderstood meaning of confidence intervals[4] and the much misused notion of P-values. It is not surprising, therefore, that DS's own attachment to the Bayesian approach (as an alternative to the classical approaches) comes to the fore in Chapter 11. Crucially, as DS makes clear, the Bayesian approach to hypothesis testing is not only more natural but also avoids all the fundamental limitations and counterintuitive assumptions of the classical approach. Perhaps the most important quote from the book is the one that starts Chapter 11 ("Learning from experience the Bayesian way"):

> *"I must now make an admission on behalf of the statistical community. The formal basis for learning from data is a bit of a mess. Although there have been numerous attempts to produce a single unifying theory of statistical inference, none has been fully accepted. It is no wonder mathematicians tend to dislike teaching statistics."*

This chapter is an excellent overview of Bayesian probability and inference. DS makes clear that he considers the Bayesian subjective approach as the best way to define the meaning of probability. I do have a couple of minor concerns, however, in this chapter in relation to the material about the use of Bayes in the law[5]. DS explains that, although Bayes is a natural means of determining whether a suspect is guilty (since we have to revise our prior belief as we observe evidence), Bayes is 'essentially prohibited' in UK courts, but the use of the likelihood ratio (LR) is not. In the legal context the LR is the probability of the evidence (such as DNA trace at the crime scene matching the defendant) given the prosecution hypothesis (the DNA is from the defendant) divided by the probability of the evidence given the defence hypothesis (the DNA is not from the defendant). What DS does not explain is that the notion that the LR is a measure of the probative value of the evidence is only meaningful because of Bayes Theorem (which tells us that the posterior odds of the prosecution hypothesis are equal to the LR times the prior odds); so the idea that using the LR is not using Bayes is a misconception by the judiciary (and many forensic scientists). Moreover, what is also not stated (and which is the source of enormous confusion) is that, unless the defence hypothesis is the exact negation of the prosecution hypothesis then the LR tells us nothing about the probability of the prosecution hypothesis. Indeed, this problem can be highlighted by DS's DNA example on page 321 which says:

$$\text{likelihood ratio} = \frac{\text{probability of DNA match assuming suspect left the trace}}{\text{probability of DNA match assuming someone else left the trace}}$$

[4] While the book provides a detailed explanation I recommend that readers also look at the explanation here: http://www.eecs.qmul.ac.uk/~norman/papers/probability_puzzles/confidence_intervals.shtml

[5] A detailed explanation of the issues I raise here can be found in: Fenton, N. E., Neil, M., & Berger, D. (2016). "Bayes and the law". Annual Review of Statistics and Its Application, 3(1), 51–77. https://doi.org/10.1146/annurev-statistics-041715-033428

In this case the defence hypothesis (someone else left the trace) is indeed the negation of the prosecution hypothesis (suspect left the trace); so, if the denominator really can be accurately determined by forensic scientists then the LR would be valid as a measure of probative value. However, in practice DNA experts have to use a *different* defence hypothesis, namely: "someone ***unrelated*** to the suspect left the trace" because the statistical basis for assessing the probability relies on this assumption. But, because this excludes anybody ***related*** to the suspect, the LR tells us nothing about the probability of the suspect leaving the trace, even if we have a prior probability for that. In the extreme case the suspect may have an unknown twin for whom the LR would be the same. In cases where only tiny amounts of DNA are detectable – so that only partial matches are possible, we can get a very high LR value even though a different (possibly related) person is more likely than the suspect to have left the trace. For mixed profile DNA 'matches' – especially those involving tiny amounts of DNA - these kinds of errors of interpreting the meaning of a high LR can (and do) lead to miscarriages of justice.

It is also worth noting that DS mentions that, for independent pieces of evidence, one can simply multiply the LRs together to get a combined LR for the entirety of the evidence. However, in practice, different pieces of evidence will not be independent. In such cases to compute the LR we need to model the relationship between different pieces of evidence and different hypotheses as a graph – a **Bayesian network**[6]. There are standard algorithms for computing the LR in such cases. DS was one of the pioneers of such Bayesian network algorithms[7] (along with people like Pearl[8]), so it is a shame that there was no room in the book for any mention of Bayesian networks, especially as they are the ideal formalism for modelling causal relationships between variables.

The last two main chapters (12 and 13) deal respectively with what DS calls the 'dark side of statistics' and with how to do things better. Chapter 12 includes many fascinating examples of misuse of statistics (including academic fraud) such as the saga of the experiments demonstrating the existence of ESP (extra sensory perception). What is common to many of the problems identified is misuse or misunderstanding of P-values. DS discusses the 2005 claim made by Ioannidis that 'most published research findings are false'. The key issue seems to be that P-value driven research (among its many other possibilities for abuse) forces an artificial crude classification on research findings as 'significant' or 'not significant'. Moreover, there is built-in bias in the scientific publication process, whereby studies that find 'no significance' tend to be either not submitted/reported or not accepted[9].

To counter the kind of problems identified in Chapter 12 (and elsewhere in the book) Chapter 13 provides useful advice and checklists for both consumers of statistical information and those who produce and communicate it. One especially important 'rule' (which was also discussed in Chapter 1) is to report absolute rather than relative risk, especially when describing medical risks. As an example of good practice, the

[6] Fenton, N. E., & Neil, M. (2018). "Risk Assessment and Decision Analysis with Bayesian Networks" (2nd ed.). CRC Press, Boca Raton.

[7] Lauritzen, S. L., & Spiegelhalter, D. J. (1988). "Local Computations with Probabilities on Graphical Structures and their Application to Expert Systems (with discussion)". *J. R. Statis. Soc. B, 50, No 2, Pp 157-224*.

[8] Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Francisco: Morgan Kaufmann Publishers In

[9] For the same kind of reason, it is quite rare to find negative book reviews. Indeed, I would not have taken the time to write this review if I did not like this book!

chapter describes the way in which one team produced fast and accurate predictions of the 2017 UK election result based on exit-polling.

It is interesting to link the advice in Chapter 13 back to what DS says In Chapter 2 about the Royal Statistical Society (of which DS is now the President):

> *"When the Statistical Society of London (later the Royal Statistical Society) was set up in 1834 by Charles Babbage, Thomas Malthus and others, they loftily declared that 'The Statistical Society will consider it to be the first and most essential rule of its conduct to exclude carefully all opinions from its transactions and publications - to confine its attention rigorously to facts – and, as far as it may be found possible, to facts which can be stated numerically and arranged in tables.' From the very start they took no notice whatsoever of this scripture and immediately started inserting their opinions about what their data on crime, health and the economy meant and what should be done in response to it. Perhaps the best we can now do is recognise this temptation and do our best to keep our opinions to ourselves."*

Sadly, too many people working in statistics have failed to heed this advice and too many statistical analyses are driven by political ideology above scientific excellence[10].

The book ends with a very brief Conclusions chapter and an extensive and very useful 25-page Glossary and 12 pages of Chapter notes.

In summary, this book is a must have for a) anybody who wants to better understand statistics and risk; b) anybody involved in the communication of statistics and risk; and c) anybody undertaking a course in data science and machine learning.

[10] This includes the Royal Statistical Society with which I and Martin Neil raised a concern over their 'statistic of the year' in 2017. See: Fenton, N. E., Neil, M., & Thieme, N. (2018). "Lawnmowers versus terrorists". Significance, 15(1), 12–15. https://doi.org/10.1111/j.1740-9713.2018.01104.x and Fenton, N. E., & Neil, M. (2018). Response to Nick Thieme's: "Statistic of the Year", not "Statistic of the Next Ten Years." https://doi.org/10.13140/RG.2.2.30958.72002