Probability Theory: The Logic of Science

by E.T. Jaynes

CAMBRIDGE, CAMBRIDGE UNIVERSITY PRESS, 2003

727pp. $65.00 HARDBACK ISBN 0-521-59271-2

The Fundamentals of Risk Measurement

by Chris Marrison

BOSTON, McGRAW-HILL, 2002

415pp. $44.95 HARDBACK ISBN 0-07-138627-0

The Elements of Statistical Learning: Data Mining, Inference and Prediction

by Trevor Hastie, Robert Tibshirani and Jerome Friedman

NEW YORK, SPRINGER, 2001

533pp. $82.95 HARDBACK ISBN 0-387-95284-5

REVIEWED BY JAMES FRANKLIN

Mathematical Intelligencer 27 (2) (Spring 2005), 83-5

A standard view of probability and statistics centers on distributions and hypothesis testing. To solve a real problem, say in the spread of disease, one chooses a "model", a distribution or process that is believed from tradition or intuition to be appropriate to the class of problems in question. One uses data to estimate the parameters of the model, and then delivers the resulting exactly specified model to the customer for use in prediction and classification. As a gateway to these mysteries, the combinatorics of dice and coins are recommended; the energetic youth who invest heavily in the calculation of relative frequencies will be inclined to protect their investment through faith in the frequentist philosophy that probabilities are all really relative frequencies. Those with a taste for foundational questions are referred to measure theory, an excursion from which few return.

That picture, standardised by Fisher and Neyman in the 1930s, has proved in many ways remarkably serviceable. It is especially reasonable where it is known that the data are generated by a physical process that conforms to the model. It is not so useful where the data is a large and little-understood mess, as is typical in, for example, insurance data being investigated for fraud. Nor is it suitable where one has several speculations about possible models and wishes to compare them, or where the data is sparse and there is a need to argue about prior knowledge. It is also weak philosophically, in failing to explain why information on relative frequencies should be relevant to belief revision and decision-making.

Like the Incredible Hulk, statistics has burst out of its constricting garments in several directions. In the foundational direction, Bayesians, especially those of an objectivist stamp like E.T. Jaynes, have reconnected statistics with inference under uncertainty, or rational degree of belief on non-conclusive evidence. In the direction of engagement with the large and messy data sets thrown up by the computer revolution, the disciplines of data mining and risk measurement, represented by the books of Hastie *et al*. and Marrison, have developed data analysis and tools well outside the restraints of traditional boundaries.

The essence of Jaynes's position is that (some) probability is logic, a relation of partial implication between evidence and conclusion. According to this point of view, statistical inference is in the same line of business as "proof beyond reasonable doubt" in law and the evaluation of scientific hypotheses in the light of experimental evidence. Just as "all ravens are black and this is a raven" makes it logically certain that this is black, so "99% of ravens are black and this is a raven" makes it logically highly probable that this is black (in the absence of further relevant evidence). That is why the results of drug trials give rational confidence in the effects of drugs. Galileo and Kepler used the language of objective probability about the way evidence supported their theories, and in the last hundred years a number of books have filled out the theory of logical probability – Keynes' *Treatise on Probability* (the great work of his early years, before he went on to easier pickings in economics), D.C. Williams's *The Ground of Induction*, George Polya's *Mathematics and Plausible Reasoning*, and now E.T. Jaynes's posthumous masterpiece, *Probability Theory: The Logic of Science*. Jaynes's school are called "objective Bayesians" or "maxent Bayesians", to distinguish them not only from frequentists but from "subjective Bayesians", who think that any degrees of belief are allowable, provided they are consistent (that is, obey the axioms of probability such as that the probability of a proposition and its negation must sum to one). The objectivists emphasise, on the contrary, that one's degree of belief ought to conform to the degree to which one's evidence does logically support the conclusion. In asking why evidential support should satisfy the axioms of probability theory, objectivists have been much impressed by the proof of R.T. Cox (*American Journal of Physics*, 1946) that any assignment of numbers to the relation of support between propositions which satisfies very minimal and natural logical requirements must obey the standard axioms of conditional probability. They have been correspondingly unimpressed by supposed paradoxes of logical probability that purport to demonstrate that one cannot consistently assign initial probabilities. In some of his most entertaining pages, Jaynes exposes these “paradoxes” as exercises in pretending not to know what everyone really does know. His reliance on symmetry principles to assign initial probabilities shows its worth, however, well beyond such philosophical polemics. In inverse problems like image reconstruction, where the data grossly underdetermines the answer, it is essential to assign initial probabilities as non-dogmatically as possible, in order to give maximum room for the data to speak and point towards the truth. Jaynes’s maximum entropy formalism allows that to be done.

In the business world, there is the same need as in science to learn from data and make true predictions. But among other forces driving the expansion of commercial statistics are the new compliance regimes in banking and accounting. Following a number of corporate scandals and unexpected collapses, the world governing bodies in banking and accounting have decided on standards that include, among other things, risk measurement. The Basel II standard in banking says, in effect, that banks may used any sophisticated statistical methodology to measure their overall risk position (in order to determine the necessary reserves), provided they disclose their methods to their national banking regulator (the Federal Reserve in the U.S., the Bank of England in the U.K.) Marrison’s book is an excellently written introduction to the standard ideas in the field. It avoids the unnecessary elements in usual statistics courses and goes immediately to the most applicable concepts. These include the “value-at-risk” formalism, which measures the loss such that worse losses occur exactly 1% (say) of the time, and the concepts needed for precision in handling rare losses, such as heavy-tailed distributions and correlations between the losses of different financial instruments. It is significant, for example, that foreign exchange rate changes resemble a random walk, but are heavy-tailed, heteroskedastic (variable in “volatility”, that is, standard deviation) and have some tendency to revert to the mean. It is perhaps surprising to learn that credit ratings are intended to mean absolute probabilities – a AAA rating means one chance in 10,000 of failure within a year; naturally it is hard to ground so small a probability in data, so one presumes that credit rating agencies will need to use priors and qualitative evidence (a euphemism for market rumors?) in the style of Jaynes. Marrison’s insights into how bank risk teams really work is enlivened by occasional dry humour: in pointing out that profits from risky trades need to be discounted, he adds “Convincing traders that their bonuses should be reduced according to *Allocated Capital* $\times H_T$ is left as an exercise for the reader.”

In accounting, the forthcoming IFRS (International Financial Reporting Standards) compliance standard will play a role similar to Basel II in banking, in enforcing higher standards of mathematical competence. It will be necessary to price options reasonably in the interests of their truthful display on balance sheets, for example. The book for accountants corresponding to Marrison’s appears not yet to be written, so there may be a gap in the market for an ambitious textbook writer who would like to become very rich very quickly.

If there is a dispute in statistics as heated as that between frequentists and Bayesians, it is that between traditional statisticians and data miners. Data mining, with its roots in the neural networks and decision trees developed by computer scientists in the 1980s, is a collection of methods aiming to understand and make money from the massive data sets being collected by supermarket scanners, weather buoys, intelligence satellites and so on. “Drink from the firehose of data”, says the science journalist M. Mitchell Waldrop. It is not easy – and especially not with the model-based methods developed by twentieth century statisticians for small and expensive data sets. With a large data set, there is a need for very flexible forms to model the possibly complicated structure of the data, but also for appropriate methods of smoothing so that one does not “overfit”, that is, learn the idiosyncracies of the particular data set in a way that will not generalise to other sets of the same kind. Are specialists in data mining (or “analytics” as they now often prefer) pioneers of new and exciting statistical methodologies, or dangerous cowboys lacking elementary knowledge of statistical models? Those who enjoy vigorous intellectual debate will want to read data miner Leo Breiman’s pugnacious `Statistical modelling: the two cultures’, *Statistical Science* 16 (2001), 199-219, with a marvellously supercilious reply on behalf of the traditionalists by Sir David Cox. As an introduction to the field for practitioners in the business world, Michael Berry’s *Mastering Data Mining* (New York, Wiley, 2000) is often recommended, but for mathematicians interested in understanding the field, Hastie *et al.*’s *Elements of Statistical Learning* is the ideal introduction. Assuming basic statistical concepts and an ability to read formulas, it runs through the methods of supervised learning (that is, generalisation from data) that have come from many sources: neural networks, kernel smoothing, smoothed splines, nearest-neighbour techniques, logistic regression and newer techniques like bagging and boosting. The unified treatment and illustration with well-chosen (and well-graphed) real datasets makes for efficient understanding of the whole field. It is possible to appreciate how different methods are really attempting the same task – for example, that classification trees developed by computer scientists to suit their discrete mindset are really performing non-linear regression. But the differences between methods are well laid out too: the table on p. 313 compares the methods with respect to such crucial qualities as scalability to large datasets, robustness to outliers, handling of missing values, and interpretability. The less-tamed territory of unsupervised learning, such as cluster analysis, is also well covered. One topic of current interest missing is the attempt to infer causes from data, but, as is clear from Richard Neapolitan's *Learning Bayesian Networks* (Harlow, Prentice Hall, 2004), that theory is still in a primitive state. Spatial statistics and text mining are not covered either; they too await readable textbooks of their own.

Mathematicians, pure and applied, think there is something weirdly different about statistics. They are right. It is not part of combinatorics or measure theory but an alien science with its own modes of thinking. Inference is essential to it, so it is, as Jaynes says, more a form of (non-deductive) logic. And unlike mathematics, it does have a nice line in colorful polemic.