Centre for Environmental Informatics

Uncertainty, Statistical Science, and Black Swans

Noel Cressie,
Centre for Environmental Informatics,
National Institute for Applied Statistics Research Australia,
University of Wollongong, Australia (ncressie@uow.edu.au)

A US Secretary of Defense in 2002 talked famously about "knowns" and "unknowns." He used the words as both a noun and an adjective. In Science, we talk about uncertainty (a noun) that spans the whole scale from completely certain (i.e., known) to totally uncertain (i.e., unknown). In Statistical Science, we quantify that scale using probabilities: An event that is completely certain gets a probability of 1, and an event that is uncertain, to different degrees, gets a probability that is less than 1.

What sort of events get a probability of 0? These would include events that we are certain can never happen, but they might also be events that we have not even thought about. In any statistical model we are working with, such unimagined events have been implicitly assigned a probability of 0. What about an event of probability 0.5? In this simplified situation, where it is only presence or absence of an event that is of interest, such an event has maximum uncertainty in that its occurrence could be predicted as well by a coin toss as by having gigabytes of data at hand.

The world we live in and try to understand is not so dichotomous; there are typically many possible events, each of which comes with its own quantified uncertainty (i.e., probability). Statistical Science uses data and the laws of probability to describe, to model, to test hypotheses, and to predict the less (and even more) certain parts of our world.

As was already mentioned, statistical models need to be constructed carefully, since an unimagined event, an event that is implicitly given a probability of 0, will not emerge out of the "ether" in subsequent inferences. So-called "unknown unknowns" would not be accounted for.

Consider the existence of a swan that is black; in the sixteenth century in Europe, this was given a probability of 0. In 1697, Dutch explorers became the first Europeans to see black swans in Western Australia. At the time, Terra Australis Incognito was a land full of unknown flora and fauna. To these explorers, the existence of black swans went from an event of probability 0 to an event of probability 1. However, to its indigenous population in Western Australia, of course swans could be black.

Probabilities are personal, yet a consensus is possible, albeit possibly incorrect, and hence statistical models can be built and defended. Uncertainties are everywhere, in different degrees, and a compelling way to handle these sources is through conditional probabilities in a hierarchical statistical model.

The Black Swan watercolour by Richard Browne for Lieut. Thomas Skottowe's manuscript, "Select Specimens from Nature, 1813" [Browne was an Irish convict transported to Australia in 1811.]

Nassim Nicholas Taleb published a book in 2007 about "Black Swan" events that, when they happen, are unpredicted and have a major impact (e.g., on our environment). The implication of Taleb's book is that in 1600 the scientific world could not predict black swans, so how can scientists predict global warming in the present day.

My response is that naturalists of the past were too certain about their science, that is, that black swans did not exist. The lessons we can learn for the future is that scientific knowledge is never perfect (i.e., is uncertain to different degrees), that modellers need to be open-minded, and that uncertainty can be quantified with probabilities. Probability theory has rules that are mathematically coherent: Joint probabilities, marginal probabilities, and conditional probabilities, particularly those obtained from Bayes' Theorem, can be calculated, and scientific conclusions can be stated with quantifications of their uncertainties (e.g., standard errors, confidence intervals, prediction intervals, quantiles) based on these probabilities. In short, Statistical Science is the science of uncertainty.

Our world is uncertain. Our attempts to explain our world (Science) are uncertain. Our measurements of our (uncertain) world are uncertain. In the past, scientists have been able to navigate their way, with more or less success, through this sea of uncertainty, but we can do better in the twenty-first century. I believe that hierarchical statistical models that recognize and quantify uncertainty through a series of conditional probability models, along with unprecedented access to big and diverse datasets, should be part of the solution to the grand challenges of this century.

Last reviewed: 21 February, 2017