PERSPECTIVES

STATISTICS

The future lies in uncertainty

Sophisticated statistical insights are crucial for gaining knowledge from ever-larger data sets

By **D. J. Spiegelhalter**

Statisticians have celebrated a lot recently. 2013 marked the 300th anniversary of Jacob Bernoulli's *Ars Conjectandi,* which used probability theory to explore the properties of statistics as more observations were taken. It was also the 250th anniversary of Thomas Bayes' essay on how humans can sequentially learn from experience, steadily updating their beliefs as more data become available (*1*). And it was the International Year of Statistics (*2*). Now that the bunting has been taken down, it is a good time to take stock of recent developments in statistical science and examine its role in the age of Big Data.

Much enthusiasm for statistics hangs on the ever-increasing availability of large data sets, particularly when something has to be ranked or classified. These situations arise, for example, when deciding which book to recommend, working out where your arm is when practicing golf swings in front of a games console, or (if you're a security agency) deciding whose private e-mail to read first. Purely data-based approaches, under the title of machine-learning, have been highly successful in speech recognition, real-time interpretation of moving images, and online translation.

Statistical science does produce some excellent machine-learning tools, but it is concerned with more than just classification or ranking: It explicitly tries to deal with the uncertainty about what can be concluded from data, be it a prediction or scientific inference. The revolutionary ideas of Bernoulli, Bayes, and others form the historical basis for such learning from large data sets (*3*).

To assess the uncertainty about unknown or future quantities, statisticians tend to build probability models of underlying processes. For example, Microsoft's TrueSKILL ranking system produces a probability distribution for the unknown skill of an online gamer. Similarly, trading models can assess both the expectation and volatility of future commodity prices. Detailed risk assessment is also vital to the insurance industry, requiring complex statistical models. Sophisticated simulation-based statistical methods, such as particle filters, are increasingly used in areas such as signal processing, dynamic economic models, and systems biology (*4*).

Traditional statistical problems could be termed "large n, small p": There were many observations (n), such as participants in a clinical trial, but few parameters were measured (p), and just a handful of hypotheses tested. More recently attention has turned to "small n, large p" problems, such as a few brain scans but with millions of voxels in each, or the expression of tens of thousands of genes in a limited number of tissue samples. To deal with these problems, statisticians developed models for complex interactions—for example, when learning the structure of a network describing gene relationships (*5*), and handling observations that are not just single data points but may comprise shapes, functions, images, or phylogenetic trees (*6*).

Many such "small n, large p" problems require screening of vast numbers of hypotheses, for which the naïve use of statistical significance is inappropriate; the standard "$P < 0.05$" criterion means that 1 in 20 nonexistent relationships will be declared significant, so that if you do enough tests, some apparent discoveries will always pop up. Procedures have been developed to control the false discovery rate (FDR); that is, the proportion of apparent discoveries that turn out to be wrong (*7*). For example, the confidence required before announcing the discovery of the Higgs boson was couched in statistical terms as 5 sigma, and this assessment included an adjustment for how many hypotheses were examined (the "look elsewhere effect").

As an example of the modern use of statistics, the June 2014 issue of *Bioinformatics* features a huge range of statistical and machine-learning techniques, many using the *Bioconductor* suite of software packages. For example, Berry *et al.* use "scan" statistics to identify regions of the human genome targeted by retroviruses, assessed with FDRs that allow for the fact that the whole genome has been searched (*8*). We recently used similar methods to show that a cluster of six London cyclist deaths in a fortnight had only around a 1 in 40 chance of occurring in any fortnight over 8 years (*9*).

Statistical principles are also routinely applied in A/B experiments on Web sites, in which small tweaks in layout and design are tested by using alternative versions at random and seeing whether the innovation has, for example, increased click-through rates. More sophisticated adaptive experimental designs are making drug testing more efficient, because treatment regimes that show little promise can be rapidly dropped without compromising the overall chance of a false discovery (*10*).

As medical databases grow and large randomized trials become more expensive and difficult to conduct, there is increasing demand for trying to make causal inferences from observational data. This is a minefield: Even with masses of data, there is no automatic technique for turning correlation into causation. But statistical science has developed frameworks for clarifying the careful analysis necessary, for example, when taking into account changing interventions

ILLUSTRATION: PETER AND MARIA HOEY/WWW.PETERHOEY.COM

Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Cambridge, CB3 0WB, UK. E-mail: david@statslab.cam.ac.uk

over time and indirect effects of therapy through mediating factors, such as viral load in HIV (*11*).

Big Data means that we can get more precise answers; this is what Bernoulli proved when he showed how the variability in an estimate goes down as the sample size increases. But this apparent precision will delude us if issues such as selection bias, regression to the mean, multiple testing, and overinterpretation of associations as causation are not properly taken into account. As data sets get larger, these problems get worse, because the complexity and number of potential false findings grow exponentially. Serious statistical skill is required to avoid being misled.

As measurement becomes ever faster and cheaper, the trend will be toward "large n, really large p" problems as, for example, images, genomes, and electronic health records become linked together. Other challenging new areas for statistical science include those traditionally handled with deterministic models, such as weather, climate, extreme natural hazards, and epidemics. In each of these, a stochastic element can be added, such as combining weather projections from randomly perturbed starting points to provide an ensemble forecast, although the appropriate role for stochastics in climate models is still contested (*12*). In these tricky areas, statisticians can show that they are prepared to deal with messy data, but still think in terms of generalizable principles (*13*). They can enable others to produce knowledge from data by promoting their own particular skills and insights, particularly by understanding both the strengths and limitations of models.

The title of this article is deliberately ambiguous: Not only is the future uncertain, but also it will be vital to understand and promote uncertainty through the appropriate use of statistical methods rooted in probability theory. Careful application of statistical science will be essential. ■

REFERENCES

1. B. Efron, *Science* **340**, 1177 (2013).
2. M. Davidian, *J. Am. Stat. Assoc.* **108**, 1141 (2013).
3. National Academy of Sciences, *Frontiers in Massive Data Analysis* (2013); see www.nap.edu/catalog.php?record_id=18374.
4. C. Andrieu, A. Doucet, R. Holenstein, *J. R. Stat. Soc. Series B Stat. Methodol.* **72**, 269 (2010).
5. B. Li, H. Chun, H. Zhao, *J. Am. Stat. Assoc.* **107**, 152 (2012).
6. B. Aydın *et al.*, *Ann. Appl. Stat.* **3**, 1597 (2009).
7. J. Fan, X. Han, W. Gu, *J. Am. Stat. Assoc.* **107**, 1019 (2012).
8. C. C. Berry *et al.*, *Bioinformatics* **30**, 1493 (2014).
9. J. Aberdein, D. Spiegelhalter, *Significance* **10**, 46 (2013).
10. B. Gaydos *et al.*, *Drug Inf. J.* **43**, 539 (2009).
11. O. O. Aalen, K. Røysland, J. M. Gran, B. Ledergerber, *J. R. Stat. Soc. Ser. A Stat. Soc.* **175**, 831 (2012).
12. C. Macilwain, *Science* **344**, 1221 (2014).
13. http://simplystatistics.org/2013/05/29/what-statistics-should-do-about-big-data-problem-forward-not-solution-backward/

10.1126/science.1251122