Beyond Cargo Cult Statistics

In an increasing variety of areas, getting it wrong brings huge dangers!

John Maindonald

31 October 2018

Questions

1) What is statistics? How is the role of statistical professionals changing?
2) What is, and is not, working in the use of statistical methodology in commerce, government, and science?
3) There has been extensive research in pychological research on the limits of untrained human intuition? What does in tell us? How does in impact on the training of statisticians, and on our work?
4) Where does statistics sit relative to developments in what is being termed "artificial intelligence"?
5) What are the other impacts from technological change?
6) What changes seem needed in statistics education?
7) What are the research challenges?

Section 2 will add further notes on:

1) Historical background to reproducibility issues
2) Machine learning
3) Statistics as a component of collaborative science that builds on the past

1 Brief comments on the questions

1.1 What is statistics?

In a paper entitled "Cargo-cult statistics and scientific crisis", Stark and Saltelli (2018), comment:

> Statistics was developed to root out error, appraise evidence, quantify uncertainty, and generally to keep us from fooling ourselves.

This seems accurate and to the point. They go on to say:

> Increasingly often, it is used instead to aid and abet weak science . . ."

There is no lack of work that melds effective use of statistically methodology with strong science. That melding should be the standard for all areas of statistical application. The old challenge, to ensure that statistical analysis gives insightful and defensible results, remains. This will remain the case, however the contexts for that challenge may change and widen.

Statistics, as I perceive it for purposes of the present discussion, is a mathematical science in its own right. It draws heavily from mathematics and from computer science, but is not mathematics and is not computer science. I'd prefer the name *statistical science*.

From global warming risks to earthquake risks to medical risks (in some cases as much from treatment as from allowing nature to take its course), risk is everywhere. Statistical insights have never been more important for the conduct of public affairs, of business, and of science.

1.2 More scientifically defensible scientific processes

Challenges facing the profession, which Stark and Saltelli (2018) highlight, mesh with challenges facing the scientific enterprise more generally. Specific issues are:

- p-values have been treated as a substitute for independent replication of experimental work. p-values, or other statistics that might be used in their place, must instead be used as aids to interpreting the results of the replication process.
- Careful, incisive, and informed critique is a crucial part of statistical model fitting. Use of model diagnostics – checks that the data is reasonably in line with model assumptions – is one component of the needed checking. Just as important is to match the model against the real world context in which model results are interpreted.
- Earthquake risk is an example of an area where models are developed and checked out using all available relevant data worldwide. The same ought to happen much more widely, e.g., for checking out models of insect response to a fumigant.

Public health is one of a number of areas where there is heavy reliance on regression methods. Results are often based on extensive non-experimental data. One analysis of one set of data is rarely enough to establish a result — solid results will require insights from multiple sources of data, and from competing analyses of the same data. Here, post-publication critique ought to become standard practice. Or, papers might be put out on the web for discussion, for a time prior to formal publication.

Reproducible reporting, with access provided to both code and data, should be standard practice.

1.3 Know thyself (and your scientific co-workers)

Psychological research has important insights to offer on the human capacity for flawed judgment. Note in particular Kahneman (2011), in a book that should be compulsory reading for all statistics and data science students.

1.4 General AI, strong AI, and Statistics

These issues gain a sharper edge with the attempts now in place, under the name of AI, to automate data collection and associated analysis. What is currently called AI is "general AI" – really a form of black box regression, with the same issues. In "The Book of Why" (Pearl and Mackenzie 2018) Pearl comments that:

> Many researchers in artificial intelligence would like to skip the hard step of constructing or acquiring a causal model and rely solely on data for all cognitive tasks. [p.16]
>
> Deep learning has given us machines with truly impressive abilities, but no intelligence. [p.30]

Pearl's interest is in "strong AI" systems that will be able to use real world information to move from data and model to reach real world conclusions – something that current systems cannot do. His examples are all, interestingly, of a statistical inference kind. He takes the link between smoking and lung cancer as an major example.

1.5 Impacts from technological change

We see new data sources, often very large datasets, widespread automation, and unparalled opportunities to work collaboratively and pool ideas and resources. The world is far more connected than at any previous time in history. Development of statistical software has been greatly helped by the resources that the internet provides for cooperative work.

The development of large databases that can be accessed internationally has been an important driver of progress in areas that include climate science, earthquake science, genomics, and astronomy. There would be a large potential benefit from making data from quarantine research similarly openly available.

1.6 Statistical education

Key challenges, as I see them, are:

- Teach limits of widely used statistical models
  - Economic models are a well-researched place to start. Make Thaler (2015) compulsory reading.
- Teach example-based model criticism. Some useful resources are:
  - O'Neil (2016) ("Weapons of Math Destruction")
  - Pearl and Mackenzie (2018) ("The Book of Why")
  - Smith (2014) ("Standard Deviations: Flawed Assumptions, Tortured Data, . . .") Smith's entertainingly written book comments on examples, from published papers and from the media, of common types of data misinterpretation. It is one of several such recent semi-popular books
- Move beyond treating *p*-values as the sign and seal that a difference has been generated.

There is a great deal to learn, and it increases with time. PhDs should move closer to the US model, with extensive coursework.

An apprenticeship model, where new graduates work for a time under the supervision of experienced statisticians, has much to commend it. This should happen in more than one application area. Statistics seems to me unusual in the extent to which practitioners are likely to benefit from exposure to multiple areas of application, carrying over insights gained from one area to another.

The managers of the future, as much as the statisticians who work with them, need to understand where statistics fits in a world where the messages come increasingly from those who present themselves as AI practicioners. They should understand these issues well enough that they will know where to go, as need may arise, for comment and/or help.

1.7 Research challenges

Progress in statistical modeling software in the past 20 years has been, as I see it, painfully slow in some areas. Standard software for handling overdispersed binomial type data has annoying limitations – certainly if one is working in the mixed models context and wants to model the scale parameter as a function of explanatory variables. Problems with the needed nonlinear optimization code remain a serious sticking point. We need software that is much better at exploring alternative strategies, and that provides better diagnostics on how and why optimization has failed. There has, by contrast, been remarkable progress in what are now termed *Integrated Development Environments*, with RStudio as a prime example.

Commercial support and demand has been the big driver of progress in such areas as robotics, image recognition, and automatic language translation. Statistical modeling software development does not attract the same level of commercial interest and support.

2 Some further notes

2.1 Reproducibility – history

Over the course of the later 16th and the seventeenth centuries, there were dramatic changes in the world view of educated people. The outcome was a recognizably modern scientific world view, where educated Europeans no longer believed in witches, or in werewolves, or that mice are spontaneously generated in piles of straw. Reproducibility, or explanation in terms of processes that have been demonstrated to be reproducible, became the criterion for accepting broadly scientific claims.

Fisher (1926) suggested the use of p-values as a mechanism for checking whether, in the face of statistical variability, a result has been convincingly reproduced. Since that time, the practice has grown of using p-values to do a job for which they are not fitted, as a substitute for replication. One p-value, generated from work by one experimenter or research group in one place, has been treated as enough.

Statistical methods developed to deal with cases where there is evident statistical variation from one replication to another. (Fisher 1937), repeating in different words (Fisher 1926), stated:

> . . . we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result.

In subsequent decades, the demand for replication was replaced by the much weaker demand to demonstrate, on the basis of a single study, $p < 0.05$. It is much weaker because there is no check on the idiosyncracies of what an individual researcher or research team, working in a single laboratory, may have done. In the areas of experimental science where this remains the standard, it is now impossible to put trust in what is published.

A radical new initiative is needed, paralleling that of the 16th century, to change the world by changing the way that statistical evidence is used in large areas of science.

It is not as easy as finding a good alternative to p-values. Whatever statistical methodology and resulting statistics are used, they must be used with informed critical insight. As Stark and Saltelli (2018) argue, paraphrasing David Freedman, much (and by implication, far too much) “of frequentist statistics is about what you would do if you had a model, and much Bayesian statistics is about what you would do if you had a prior.”

2.2 Machine learning

There has been huge progress in specialist engineering domains such as robotics and automated guidance systems, where machines directly interrogate and respond to their environment. In these areas, system failures have relatively immediate, and perhaps dangerous, consequences. The methods used do not directly carry across to systems that rely on human data input, and on human checks on the the relevance and accuracy of output.

The widespread use of machine learning software makes it easier than ever to identify spurious as well as real associations, deceiving untrained human intuition in ways such as are documented in Kahneman (2011). The critical evaluative skills needed to negotiate this territory appear to be in short supply within the science scene and within Government.

As Pearl and Mackenzie (2018) comment, "Data is dumb". Mathematical equations, and statistics calculated using mathematical formula are likewise, though Pearl does not precisely say this, dumb. Scientific understanding must drive their use and interpretation.

Strong AI remains the domain of humans, and sets limits to present possibilities for automation. There will be increasing help from machines, in ways that may in future well include the kinds of system that it has been Pearl's aim to build.

Comments in Cliff (1983) are apt:

> ... beautiful computer programs do not really change anything fundamental. Correlational data are still correlational, and no computer program can take account of variables that are not in the analysis. Causal relationships can only be established through patient, painstaking attention to all relevant variables ...

2.3 Statistics as a component of collaborative science that builds on the past

Areas of science where the nature of the work requires close collaboration internationally, in data collection as well as in modeling and analysis, strike me as in much better shape than laboratory science. Each new investigation has to carefully critique, and build on, what has gone before. This provides a level of protection that, for one-off laboratory experiments, has to be more explicitly built in. Such areas include oceanography, geoscience, climate science, and much of clinical (but not pre-clinical) medical science. We need to make laboratory science, in ways that are not at the moment standard practice, part of a cooperative process that very explicitly builds on what has gone before, where it takes more than a single laboratory experiment to give credence to claims made, and where work gets the same kind of critique from multiple disciplinary perspectives as happens in areas where the nature of the work more directly ensures such scrutiny.

Each new set of data should add to what has gone before, adding both to earlier scientific insights and to our ability to build and critique models. In many contexts, checking models out across a wide range of comparable datasets is needed as a basis for effective judgments on what models are defensible and give plausible results. Too often, model choice and analysis rely on part only of the relevant data.

References

Cliff, Norman. 1983. "Some Cautions Concerning the Application of Causal Modeling Methods." *Multivariate Behavioral Research* 18 (1): 115–26. doi:10.1207/s15327906mbr1801_7.

Fisher, R A. 1926. "The Arrangement of Field Experiments." *Journal of the Ministry of Agriculture GB* 33: 503–13.

———. 1937. *The Design of Experiments.* 2nd ed. Oliver; Boyd.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow.* 1st ed. Penguin Books.

O'Neil, Cathy. 2016. *Weapons of Math Destruction.* 1st ed. Crown.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why. the New Science of Cause and*

Effect. 1st ed. Basic Books.

Smith, G. 2014. *Standard Deviations: Flawed Assumptions, Tortured Data, and Other Ways to Lie with Statistics.* Duckworth Overlook.

Stark, Philip B., and Andrea Saltelli. 2018. "Cargo-Cult Statistics and Scientific Crisis." *Significance* 15 (4): 40–43. doi:10.1111/j.1740-9713.2018.01174.x.

Thaler, Richard H. 2015. *Misbehaving.* Allen Lane.