The Identity of Statistics in Data Science

Nov. 1, 2015

Tommy Jones

Tommy Jones is the director of data science at Impact Research, LLC. He holds an MS in mathematics and statistics from Georgetown University and a BA in economics from the College of William and Mary. He is a PhD student in the George Mason University Department of Computational and Data Sciences. He specializes in statistical models of language and time series modeling.

Data science has been generating considerable interest inside and outside of the statistics community. Within the statistics community, there is a debate about whether data science and statistics are distinct disciplines. This conversation about data science betrays an anxiety about our (statisticians') identity.

In a July 2013 article in Amstat News, "Aren't We Data Science?" former ASA president Marie Davidian summarizes these concerns: "I've been told of university administrators who have stated their perceptions that statistics is relevant only to 'small data' and 'traditional' 'tools' for their analysis, while data science is focused on Big Data, Big Questions, and innovative new methods."

Similarly, Norman Matloff titled his November 2014 editorial in Amstat News "Statistics Losing Ground to Computer Science." He raised many good points, but his title cuts to the heart of our anxiety. Does "data science" mean we're being replaced?

I believe this anxiety stems from an overly-broad definition of statistics and an unclear definition of data science. For my part, I've come to see data science as supply chain management for "data products." This supply chain starts with real-world problems and ends with a report, business decision, or software. The middle contains lots of statistics, databases, programming, communicating, etc. Data science is fundamentally multidisciplinary. "But," you may ask, "isn't that just statistics?"

Davidian's article is titled "Aren't We Data Science?" after all. Randy Bartlett answered "We Are Data Science" in a subsequent Statistician's View. This "everything data" definition of statistics is popular among statisticians. Former ASA President Robert Rodriguez championed this view in 2012, offering ASA as a "big tent." The popular blog Simply Statistics states, "Whenever someone does something with data, we should claim them as a statistician."

There is historical precedent for this claim. Statistics as a discipline originated in the 18th century. Least squares dates to the early 1800s with Gauss and Legendre. We statisticians were the only data game in town, even as statistics became tied with mathematical probability in the 19th and early 20th centuries.

Yet times have changed. Judging by current statistics curricula, statistics is more closely tied to the mathematics of probability than to fundamentals of data management. Survey the requirements of most graduate statistics programs. There is a core of courses in measure-theoretic probability, theoretical statistics, and linear models. I am not saying computation, database management, and application foci are absent. But the degree to which such courses are emphasized, or even offered at all, is highly variable. What proportion of programs require a scientific databases course or a high-performance computing course? We are well trained in quantifying uncertainty and deriving asymptotics. We are poorly trained in the tools of modern data management.

What has driven this structural break? Data have proliferated. This isn't about the volume of data in a "Big Data" sense, but rather that data are more popular. More data sets exist. More people are analyzing data. It is no longer the case (if it ever was) that only scientists, trained to deal with complexity, are the consumers of data products. The need for compelling visualizations and narratives to convey complicated stories has increased.

As models have become more accurate, they have also become more complex.

Ensembles of models are often better predictors than any single model. Ensembles are empirically accurate, but their asymptotic properties are often unknown. And an additional question arises: Asymptotic to what? One could take any or all of the number of observations, predictors, models in the ensemble, etc. to infinity and possibly arrive at different solutions. In the age of Big Data, asymptotic properties matter.

Finally, data are bigger in a Big Data sense. Storing, moving, and processing terabytes of data is neither simple nor all "statistical" in nature. There has long been a working relationship between statistics and computer science. But now software engineering knowledge is required if any useful analysis is to come from a Big Data project.

Whither statistics?

The More Things Change, the More They Stay the Same

In an age of Big Data, I believe statistics' focus on probability and asymptotic properties is more valuable, not less. As we move toward more complex statistical and machine-learned models, there is still a need to understand the properties of and to get inferences from these models. A (computational) data scientist once told me "statisticians will be the ones to help us figure this mess out." These are questions at the heart of theoretical statistics.

And in a world that is streaming data, careful research design and data collection are as important as ever. A biased sample is still biased if it has a million observations. This is especially important when the data are born of the Internet and people implicitly or explicitly opt in. These are challenges survey statisticians face regularly.

Recent research by statisticians is tackling some of these issues. Gerard Biau, Luc Devroye, and Gabor Lugosi have demonstrated the consistency of averaging classifiers. Stefan Wager, Trevor Hastie, and Bradley Efron propose methods to get prediction errors of bootstrapped and bagged learners. Abhijit Dasgupta et. al show how to estimate effect size using nonparametric, "black box" models. Andrew Womack, Elias Moreno, and George Casella have shown that a popular model for text mining is an inconsistent estimator.

But Sometimes, Things Just Change

While many of the fundamental problems facing statisticians are the same, the applications and environment are different. Statistics education, particularly at the graduate level, must adapt. As data get "bigger" and research and applications become more multidisciplinary, the need for statisticians to communicate and collaborate with a wide range of professionals and laypeople increases.

Statistics education should require minimum competency in fundamentals of computer science. ASA's recent statement, "The Role of Statistics in Data Science" highlights three data science skillsets: database management, statistics and machine learning, and distributed and parallel systems. Statisticians must work closely with software engineers to develop solutions that scale. We must understand the code so that scaled solutions still have desirable statistical properties. I believe that statisticians should have minimum foundational training in database management and high-performance computing.

In addition, examples and applications in introductory statistics courses may need updates. For example, ensemble methods will be at least as important as linear regression in the coming years. We may consider teaching concepts like Zipf’s and Heap’s laws early on, as analyses of linguistic data are growing more common.

It is an exciting time to be a statistician. Statistical models and methods are applied in ways unimaginable only a decade ago. Airplanes fly themselves; doctors use statistical models to aid diagnoses; scientific research involves mining massive data sets. The importance of these tasks makes understanding our models an imperative. Yet, fundamental statistical properties of these models remain little understood.

I am not convinced that statistics is data science. But I am convinced that the fundamentals of probability and mathematical statistics taught today add tremendous value and cement our identity as statisticians in data science.