Three-Toed Sloth

Slow Takes from the Canopy (My Very Own Internet Tradition)

July 25, 2012

« Acknowledgments | Main | Undergraduate Capacity Control »

No, Really, Some of My Best Friends Are Data Scientists

> *Attention conservation notice*: 3500+ words on the scope of statistics, the teaching of statistics, and "data science". Drafted in October, forgotten until now, when I stumbled across it again in the course of prepping for next week's talk at JSM.

A follow-up, because lurkers[1] demanded it in e-mail, and Cathy posted a response of sorts as well.

I'll start, as I perhaps should have done, with the case *against* statistics. Suppose you were exposed to that subject as a sub-cabalistic ritual of manipulating sums of squares and magical tables according to rules justified (if at all) only by a transparently false origin myth — that is to say, you had to endure what is still an all-too-common sort of intro. stats. class — or, perhaps worse, a "research methods" class whose content had fossilized before you were born[2]. Suppose you then looked at the genuinely impressive things done by the best of those who call themselves "data scientists". Well then no wonder you think "This is something new and wonderful"; and I would not blame you in the least for not connecting it with statistics. Perhaps you might find some faint resemblance, but it would be like comparing a child's toy wagon to a Ducati.

Modern statistics is not like that, and has not been for decades. Statistics has been, since its beginning, a branch of applied mathematics which designs and analyses methods for drawing reliable inferences from imperfect (incomplete, limited, distorted, noisy) data. ("Applied mathematics" doesn't quite have the right connotations; "mathematical engineering" would be better[3]. Other branches of mathematical engineering include optimization, computer science, and, recursively, numerical analysis.) Just as mechanical or electrical engineers, in order to design good machines and circuits, need to be able to analyze all kinds of machines and circuits, for us to design good inferential procedures, we need to analyze all sorts of methods of inference, and all sorts of reliability. We have found stochastic models to be very useful for this, so we use probability; we have developed some specialized mathematical concepts and techniques for the analyses, which we call "statistical theory". These are, or should be, always in the service of crafting methods of inference, being honest and careful about how reliable they are, and getting as much out of the data as we can.

To repeat myself, for a long time our theory about statistical-inference-in-general ran way, way ahead of what we could actually *do* with data, given the computational resources available to us. Over the last, say, thirty years, those computational constraints have, of course, drastically relaxed, and suddenly powerful and flexible methods which were once merely theoretical objects became executable software. This in turn opened up new theoretical avenues, which led to new methods of inference, and so on. This doesn't help with the fossils, and no doubt there are many places where it has not made its way down into the undergraduate curriculum, especially not into what is taught to non-statisticians.

None of this changes the fact that the skills of a "data scientist" are those of a modern statistician. Let me quote from Cathy's first post at some length:

> Here are some basic skills you should be looking for when you're hiring a data scientist. They are general enough that they should have some form of all of them (but again don't be too choosy about exactly how they can address the below needs, because if they're super smart they can learn more):
>
> - Data grappling skills: they should know how to move data around and manipulate data with some programming language or languages.
> - Data viz experience: they should know how to draw informative pictures of data. That should in fact be the very first thing they do when they encounter new data
> - Knowledge of stats, errorbars, confidence intervals: ask them to explain this stuff to you. They should be able to.
> - Experience with forecasting and prediction, both general and specific (ex): lots of variety here, and if you have more than one data scientist position open, I'd try to get people from different backgrounds (finance and machine learning for example) because you'll get great cross-pollination that way
> - Great communication skills: data scientists will be a big part of your business and will contribute to communications with big clients.

I will not pretend to speak about what is taught to undergraduates in *every* statistics department, but I can be quite specific about what we teach here. Students typically come to us taking a year-long basic statistics sequence; then a year of probability and mathematical statistics (some students enter here, and take more electives later); then modern regression and advanced data analysis; and then, or simultaneously, electives. (We require five upper-division statistics electives.) The electives run over statistical graphics, statistical computing, data mining, stochastic processes, surveys and sampling, epidemiology, experimental design, unsupervised learning and clustering, multivariate methods, complex systems, and potentially other subjects. (Not all of my colleagues, it seems, are as good about putting class materials on the Web as I'd like.) Graphics and computing are not required, but are popular and typically strongly urged on them. In the spring of their senior year (or sometimes even their junior year), as many of our majors as we can handle take the research projects class, where they spend a semester working in groups of 2--3 on data provided by investigators from outside the department, with a real-world analysis problem that needs to be solved to the investigator's satisfaction. (For various reasons, we have to limit that class to only about 15 students a year.) In terms of math, we *require* the usual multidimensional calculus and differential equations sequence, plus linear algebra; I usually try to steer my advisees to getting some exposure to Fourier analysis.

Now take Cathy's desiderata in turn:

- "Data grappling skills" are things we teach along the way in modern regression and advanced data analysis, which between them guarantee at least a year of intensive R usage. These are things we explicitly teach in statistical computing, with even more R.
- "Data viz experience" begins with our intro. statistics classes, and then goes on in great depth in statistical graphics and visualization, with even more of the accompanying R. The habit of starting to understand any new data by drawing pictures is certainly something we inculcate.
- "Knowledge of stats, errorbars, confidence intervals" needs no elaboration.
- "Experience with forecasting and prediction": Again, both regression and advanced data analysis are full of this, and data mining (which is not mandatory but is quite popular) is about little else. The point of the data mining class, in fact, is that their statistical skills can almost immediately achieve a great range of real-world prediction tasks.
- "Great communication skills": Graphics, regression, and advanced data analysis all require, and grade on, the ability to write comprehensible and useful data analysis reports. The research projects class involves a lot of this, as well as regular oral presentations. It would be good if we did more on this front, however.

We could demand more programming, but as Cathy says very well,

> don't confuse a data scientist with a software engineer! Just as software engineers focus on their craft and aren't expected to be experts at the craft of modeling, data scientists know how to program in the sense that they typically know how to use a scripting language like python to manipulate the data into a form where they can do analytics on it. They sometimes even know a bit of java or C, but they aren't software engineers, and asking them to be is missing the point of their value to your business.

However, we are right to demand *some* programming. It is all very well to be able to use someone else's software, but (again, to repeat myself) "someone who just knows how to run canned routines is not a data analyst but a drone attached to a machine they do not understand"[4]. This is why I insist on programming in all my classes, and why I encourage my advisees to take real computer science classes.

People with certain sorts of computer science backgrounds often have these skills: machine learning, obviously, but also information retrieval, or natural language processing, or even databases. (The developments in database techniques which make this kind of work possible on an industrial scale do not get enough love.) From the perspective of someone coming from computer science, a lot of what we teach The Kids now looks a lot more like machine learning than statistics as it was taught circa 1970, or even circa 1980 [5]. And then of course there is signal processing, some branches of applied math, experimental physics, and so on and so forth.

So of course you don't need a statistics degree to learn these things. You don't even need to have taken a single statistics class (though it's taking ~~food out of my non-existent children's mouths~~ tuna out of my cat's bowl for me to say so). *I've* never taken a statistics class[6]. Everything I *know* about statistics I've learned without formal instruction, from reading books, reading papers, listening to talks, and arguing with people.

Like the rest of the applied mathematical sciences, learning the most useful parts of statistics is not, in my experience, intrinsically hard for anyone who already has a decent grounding in some other mathematical science.

Or rather, the hard parts are finding good references, finding the time to tackle them, and maintaining the motivation to keep doing so, especially as mastering them really does mean trying to do things and failing, and so feeling like an idiot. (The beginner's mind fits uncomfortably, once you've out-grown it.) My job as a statistics professor, teaching classes, is to provide those references, force the time to be carved out, try to maintain the motivation, and provide feedback. But so far from wanting to discourage people who aren't statistics students from learning these things and doing data analysis, I think that they should be encouraged to do so, which is part of why I spend a lot of time and effort in working up freely-accessible teaching materials.

Now, beyond the sheer knowledge of established methods and techniques, there is, as Cathy rightly says in her follow-up post, the capacity to figure out what the important problems to solve are, and the capacity to use established principles to devise solutions — in the limit, if necessary, to establish new principles. It's easier to deal with these in reverse order.

I think we are doing our students, even at our undergrads, a profound dis-service if we do *not* teach them those general principles, and how canned procedures for particular cases come from them. There are books with titles like *A Gazillion Statistical Tests*, and they serve a role, just like the tables of emission spectra in physics handbooks, but memorizing such things is not what learning the subject is about in either field. If this is the kind of thing Cathy thinks I mean by "undergraduate statistics", then yes, she's right to say it's nowhere near enough. (And there are schools where the curriculum doesn't really go beyond training students to be human interfaces to such tables, and/or commercial software.) But I am frankly skeptical about the utility of much of what is in such references, having never had a real problem where even so much as an *F*-test was useful. As for the cookbook distributional assumptions of Gaussianity and the like, well, Fisher himself was quite shaken to realize that people were taking them as defaults, and there is a reason I spend so much time on non-parametrics and resampling.

The real goal, however, is not even to teach about (say) specific cross-validation techniques, but, once again, to teach about the general principles and how to use them in specific cases. I don't know of a better way, yet, of doing this than showing how I'd do it on a range of examples and helping students work through examples on their own, the more real the better. (How, after all, does one learn to set up a model in physics?) Drilling on the most abstract theoretical ideas isn't a substitute for this, since it leads to the "But, sir, what Hamiltonian should I diagonalize?" problem. The capacity to devise solutions on the basis of known principles is a kind of craft knowledge, a collection of technical habits of the mind resting on pattern recognition and heuristics, and like any such knowledge it improves strongly with feedback and practice. (ObInvocationOfLocalCultureHero: *The Sciences of the Artificial* is very sound on this.) Twenty-one year olds with fresh undergraduate degrees won't have ten years of practice in data analysis, but here we should have given them at least two.

The other issue Cathy raises is the capacity to formulate a good problem. Getting this right is logically prior to solving problems, and it's a key capacity of a researcher or even any skilled and autonomous professional. Sadly, it seems a lot harder to help people develop than mere problem-solving. (Good old fashioned cognitive psychology gives us pretty good accounts of problem solving; problem formulation, not so much.) Indeed, to some extent schooling-as-usual seems actively harmful, since it rewards solving what are already well-defined problems — at most, guessing what ambiguous statements mean to the graders — and not reducing a confused situation into a clear problem (and revising the formulation as necessary). Nonetheless, some people are better at problem-formulation than others, and people tend to improve with practice. (Or perhaps those who are bad find other lines of work, and this impression is a survivorship bias?) If anyone has a better idea of how to teach it than to give learners chances to formulate their own problems, and then critiquing their efforts, please let me know, because I'd find it very helpful.

Let me just add that there is something in Cathy's posts, especially the follow-up, which seem well-intentioned but *potentially* hazardous. This is the idea that all that really matters is being "smart" (and "nerdy"), and that if you've got that, you can pick up whatever it is you need to know. Rather than rehash the question of a one-dimensional smartness *again*, I will just make two observations, one anecdotal and the other more scientifically grounded.

The less-personal one invokes the work of Carol Dweck and associates about learning. Over-simplifying, what they find is that it is actually counter-productive for students to attribute their success or failure in learning about something to an innate talent, ability or knack, apparently because it makes it harder to use feedback successfully. If, after seeing that they bombed question no. 3 on the exam, a student thinks, "Wow, I guess I really don't have a head for this", well, what can they do about that? But the student who explains their failure by inadequate preparation, poor study skills, etc., has something they can work on productively. In other words, explaining outcomes by innate talent inside themselves actually seems to make success harder to achieve, by leading to neglect of other causes which they *can* control. There is however a role here for an unshakeable conviction on the learner's part that they are smart and nerdy enough to learn any nerdy thing, if that keeps them from falling for the "I don't have a head for this" line.

The ancedotal observation is to mount one of my hobby-horses again. Why have physicists (and others in their wake) made so many so very many dubious-to-ridiculous claims about power laws? It's not that they aren't smart enough or nerdy enough — certainly someone like Gene Stanley is a lot smarter than I am, and doing things right is just not that hard. I am pretty sure the reason is that they think smarts and physics nerdiness removes the need to learn about other fields before trying to solve their problems, which, as usual, is wrong.

Again, the last thing I want is to discourage people from learning and practicing data analysis. I think statistics is beautiful and useful and the world would be a better place if its were better known. It would be absurd to insist that only Duly Credentialed Statisticians should do it. (Some people have suggested introducing standardized certification for statisticians; I would oppose this if it seemed a serious threat.) I agree whole-heartedly with Cathy about what goes into a good "data scientist", and I am not trying to claim that any freshly-minted statistics BA (even from here) is as good as anyone now in a "data scientist" job; that would be silly. I *am* saying that the skills and habits of mind which go into such jobs are the ones we try to teach, I hope with some success. If people want to call those who do such jobs "data scientists" rather than "statisticians" because it sounds more dignified[7], or gets them more money, or makes them easier to hire[8], then more power to them. If they want to avoid the suggestion that you need a statistics *degree* to do this work, they have a point but it seems a clumsy way to make it. If, however, the name "statistician" is avoided because that connotes not *a powerful discipline which transforms profound ideas about learning from experience into practical tools*, but rather, *a meaningless conglomeration of rituals better conducted with twenty-sided dice*, then we as a profession have failed ourselves and, more importantly, the public, and the blame lies with us. Since what we have to offer is really quite wonderful, we should not let that happen.

[1]: No, am still not going to add comments. It's all I can do to write posts, never mind tending a comments section, and I know only too well what would happen if I left one alone. ^

[2]: If you doubt that such fossilization actually happens, observe that in *2011*, a respectable academic publisher, specializing in research methods for psychology, published a book hailing confidence intervals — confidence intervals! — as "new statistics", with every appearance that they are in fact new to that audience. ^

[3]: I know I lifted the idea that statistics is a sort of mathematical engineering from a conversation with Andy Gelman, but he's not to blame for the use I've made of the notion. ^

[4]: I am classist enough to look down on someone who *chooses*, when they have an alternative, to be a mere fleshly interface to a dumb tool, to be enslaved by one of our own creations. Too often, of course, people have no choice in this. ^

[5]: On the other hand, machine learning itself was drastically transformed in the 1980s and 1990s by importing lots of ideas from statistics, ranging from cross-validation to empirical process theory. Compare, say, Tou and Gonzalez's *Pattern Recognition Principles* from 1974 to Ripley's *Pattern Recognition and Neural Networks* from 1996, and you see not just 20 years of technical development along the same lines, but really deep changes in aims and approaches, in what counts as good work on pattern recognition. (Perhaps this is why Amazon has Tou and Gonzalez categorized under "Books > History > Ancient > Early Civilization"?) This was a scientific revolution which immediately affected industrial practice, and it would be great to have a good history of it. ^

[6]: My own degrees are in theoretical physics, as I may have mentioned once or twice, and along the whole curriculum from shooting the monkey to entropy-driven forces between topological defects in low-temperature spin systems, there were only three classes with any statistical content. These were: (i) undergraduate physics lab, where we were taught about standard errors, propagation of error, and fitting a line through three points by least squares; (ii) a graduate "pattern recognition" class, reproducing what was taught under that heading to Russian mechanical engineers in the Brezhnev era (the names "Vapnik" and "Chervonenkis" never passing the professor's lips); and (iii) some asides on estimation in my revered probability guru's stochastic processes classes. ^

[7]: Similarly, if people who pick up garbage want to be called "sanitation engineers", because it makes it a bit easier to do an unpleasant and ill-paid but necessary job, why not? (It's not as though we could imagine a world where they'd have pay and dignity.) ^

[8]: Anecdotally, I have been told that our insane immigration system gives you an easier time if your job description says "scientist" rather than "engineer" or "analyst". I have also been told of HR departments which would object to hiring someone as "a statistician" if they don't have a statistics degree, but can't object to hiring someone as "a data scientist" to do the same job, because there are no data science degrees. (Yet.) If so, this may be the first intellectual movement which is really just regulatory arbitrage. ^

Update, next day: I should have made clear that, as usual, I'm not speaking for the CMU statistics department. Also, I'd like to thank reader R.W. for bringing @ML_hipster's "Yeah, so I'm actually a data scientist. I just do this barista thing in between gigs." to my attention. It's funny because it's a bit too cruel to be true.

28 July: Minor typo fixes, thanks to a different R.W.

31 July: A characteristically generous reply from Cathy.

Manual trackback: Flowing Data; Fernando Pereira (whose friendly amendment I happily accept); Numbers Rule Your World (deserves a response, probably won't get one anytime soon); Paperpools; Synthetic Daisies; Since it is not...; Synthetic Analytics; Equitablog

Enigmas of Chance

Posted at July 25, 2012 09:10 | permanent link

Three-Toed Sloth