The statistics style report

Ben Klemens

19 December 2009

It may sound like an oxymoron, but there is such a thing as fashionable statistical analysis. Where did this come from? How is it that our tests for Truth, upon which all of science relies, can vacillate from season to season like hemlines?

Before discussing those questions, let me tap on the brake, and point out that statistics as a whole is not arbitrary. The Central Limit Theorem is a mathematical theorem like any other, and if you believe the basic assumptions of mathematics, you have to believe the CLT. The CLT and developments therefrom were the basis of stats for a century or two there, from Gauss on up to the early 1900s when the whole system of distributions (Binomial, Bernoulli, Gaussian, t, chi-squared, Pareto) was pretty much tied up. Much of this, by the way, counts not as *statistics* but as *probability*.

Next, there's the problem of using these objective truths to describing reality. That is, there's the problem of writing models. Models are a human invention to describe nature in a human-friendly manner, and so are at the mercy of human trends. Allow me to share with you my arbitrary, unsupported, citation-free personal observations.

Number crunching The first thread of trendiness is technology-driven. In every generation, there's a line you've got to draw and say 'everything after this is computationally out of reach, so we're assuming it away', and the assume-it-away line drifts into the distance over time. Here's a little something from a 1939 stats textbook on fitting time trends [Arkin and Colton, 1939, p 43]:

> To fit a trend by the freehand method draw a line through a graph of the data in such a way as to describe what appears to the eye to be the long period movement. ... The drawing of this line need not be strictly freehand but may be accomplished with the aid of transparent straight edge or a "French" curve.

As you can imagine, this advice does not appear in more recent stats texts. In this respect, a stats text can actually become obsolete. But as time passes, approximations like this are replaced by new techniques that were before just written off as impossible. [Now reading: Hastie and Tibshirani [1990], who offer a few hundred pages on computational methods to do what was done by freehand above.]

Computational ability has brought about two revolutions in statistics. The first is the linear projection (aka, regression). Running a regression requires inverting a

matrix, with dimension equal to the number of variables in the regression. A two-by-two matrix is easy to invert (¿remember all that about $ad - bc$?) but it gets significantly more computationally difficult as the number of variables rises. If you want to run a ten-variable regression using a hand calculator, you'll need to set aside a few days to do the matrix inversion. My laptop will do the work in 0.002 seconds. [It's still in under a second up to about 500 by 500, but 1,000 by 1,000 took 8.9 seconds. That includes the time it took to generate a million random numbers.]

So revolution number one, when computers first came out, was a shift from simple correlations and analysis of variance and covariance to linear regression. This was the dominant paradigm from when computers became common until a few years ago.

The second revolution was when computing power became adequate to do searches for optima. Say that you have a simple function to take in inputs and produce an output therefrom. Given your budget for inputs, what mix of inputs maximizes the output? If you have the function in a form that you can solve algebraically, then it's easy, but let us say that it is somehow too complex to solve via Lagrange multipliers or what-have-you, and you need to search for the optimal mix.

You've just walked in on one of the great unsolved problems of modern computing. All your computer can do is sample values from the function—if I try these inputs, then I'll get this output—and if it takes a long time to evaluate one of these samples, then the computer will want to use as few samples as possible. So what is the method of sampling that will find the optimum in as few samples as possible? There are many methods to choose from, and the best depends on enough factors that we can call it an art more than a science.

In the statistical context, the paradigm is to look at the set of input parameters that will maximize the likelihood of the observed outcome. To do this, you need to check the likelihood of every observation, given your chosen parameters. For a linear regression, the dimension of your task was equal to the number of regression parameters, maybe five or ten; for a maximum likelihood calculation, the dimension is related to the number of data points, maybe a thousand or a million. Executive summary: the problem of searching for a likelihood function's optimum is significantly more computationally intensive than running a linear regression.

So it is no surprise that in the last twenty years, we've seen the emergence of statistical models built on the process of finding an optimum for some complex function. Most of the stuff below is a variant on the search-the-space method. But why is the most likely parameter favored over all others? There's the Cramer-Rao Lower Bound and the Neyman-Pearson Lemma, but in the end it's just arbitrary. Gauss had no theorems that this framework gives superior models relative to linear projection, but it does make better use of computing technology.

Hemlines The second thread of statistical fashion is whim-driven like any other sort of fashion. Golly, the population collectively thinks, everybody wore hideously bright clothing for so long that it'd be a nice change to have some understated tones for a change. Or: now that music engineers all have ProTools, everything is a wall of sound; it'd be great to just hear a guy with a guitar for a while. Then, a few years later, we collectively agree that we need more fun colors and big bands. Repeat the cycle until

civilization ends.

Statistical modeling sees the same cycles, and the fluctuation here is between the parsimony of having models that have few moving parts and the descriptiveness of models that throw in parameters describing the kitchen sink. In the past, parsimony won out on statistical models because we had the technological constraint.

If you pick up a stats textbook from the 1950s, you'll see a huge number of methods for dissecting covariance. The modern textbook will have a few pages describing a Standard ANOVA (analysis of variance) Table, as if there's only one. This is a full cycle from simplicity to complexity and back again. Everybody was just too overwhelmed by all those methods, and lost interest in them when linear regression became cheap.

Along the linear projection thread, there's a new method introduced every year to handle another variant of the standard model. E.g., last season, all the cool kids were using the Arellano-Bond method on their time series so they could assume away endogeneity problems. The list of variants and tricks has filled many volumes. If somebody used every applicable trick on a data set, the final work would be supremely accurate—and a terrible model. The list of tricks balloons, while the list of tricks used remains small or constant. Maximum likelihood tricks are still legion, but I expect that the working list will soon find itself pared down to a small set as optimum finding becomes standardized.

In the search-for-optima world, the latest trend has been in 'non-parametric' models. First, there has never been a term that deserved air-quotes more than this. A 'non-parametric' model searches for a probability density that describes a data set. The set of densities is of infinite dimension. If all you've got a hundred data points, you ain't gonna find a unique element of $\Re^\infty$ with that. So instead, you specify a certain set of densities, like sums of Normal distributions, and then search for that subset that leads to a nice fit to the data. You'll wind up with a set of what we call *parameters* that describe that derived distribution, such as the weights, means, and variances of the Normal distributions being summed.

But 'non-parametric' models allow you to have an arbitrary number of parameters. Your best fit to a 100-point data set is a sum of 100 Normal distributions. If you fit 100 points with 100 parameters, everybody would laugh at you, but it's possible. In that respect, the 'non-parametric' setup falls on the descriptive end of the descriptive-to-parsimonious scale. In my opinion.

I don't want to sound mean about 'non-parametric' methods, by the way. It's entirely valid to want to closely fit data, and I have used the method myself. But I really think the name is false advertising. How about *distribution-fitting methods* or *methods with open parameter counts*?

Bayesian methods are increasingly cool. If you want to assume something more interesting than Normal priors and likelihoods, then you need a computer of a certain power, and we beat that hurdle in the 90s as well, leaving us with the philosophical issues. In the context here, those boil down to parsimony. Your posterior distribution may be even weirder than a multi-humped sum of Normals, and the only way to describe it may just be to draw the darn graph. Thus, Bayesian methods are also a shift to the description-over-parsimony side.

[Method of Moments estimators have also been hip lately. I frankly don't know where that's going, because I don't know them very well.

Also, this guy really wants multilevel modeling to be the Next Big Thing in the linear model world, and makes a decent argument for that. He likes it because it lets you have a million parameters, but in a structured manner such that we can at least focus on only a few. I like him for being forthright (on the blog) that the computational tools he advocates (in his books) will choke on large data sets or especially computationally difficult problems.]

Increasing computational ability invites a shift away from parsimony. Since PCs really hit the world of day-to-day stats recently, we're in the midst of a swing toward description. We can expect an eventual downtick toward simpler models, which will be helped by the people who write stats packages—as opposed to the researchers who caused the drift toward complexity—because they write simple routines that implement these methods in the simplest way possible.

So is your stats textbook obsolete? It's probably less obsolete than people will make it out to be. The basics of probability have not moved since the Central Limit Theorems were solidified. In the end, once you've picked your paradigm, not much changes; most novelties are just about doing detailed work regarding a certain type of data or set of assumptions. Further, those linear projection methods or correlation tables from the 1900s work pretty well for a lot of purposes.

But the fashionable models that are getting buzz shift every year, and last year's model is often considered to be naïve or too parsimonious or too cluttered or otherwise an indication that the author is not down with the cool kids—and this can affect peer review outcomes. A textbook that focuses on the sort of details that were pressing five years ago, instead of just summarizing them in a few pages, will have to pass up on the detailed tricks the cool kids are coming up with this season—which will in turn affect peer reviews for papers written based on the textbook's advice.

A model more than a few years old has had a chance to be critiqued while a new model has not. So using an old technique gives peer reviewers the opportunity to use their favorite phrase: *the author seems to be unaware*, in this case that somebody has had the time to find flaws in the older technique and propose a new alternative that fixes those flaws—while the new technique is still sufficiently novel that nobody has had time to publish papers on why it has even bigger flaws.

All this is entirely frustrating, because we like to think that our science is searching for some sort of true reflection of constant reality, yet the methods that are acceptable for seeking out constant reality depend on the whim of the crowd.

References

Herbert Arkin and Raymond R Colton. *An Outline of Statistical Methods as Applied to Economics, Business, Education, Social and Physical Sciences, Etc.* Barnes & Noble, 4th edition, 1939.

T J Hastie and R J Tibshirani. *Generalized Additive Models*. Number 43 in Monographs on Statistics and Applied Probability. Chapman & Hall, 1990.