



[Statistical Modeling: The Two Cultures]: Comment

Author(s): D. R. Cox

Source: *Statistical Science*, Vol. 16, No. 3 (Aug., 2001), pp. 216-218

Published by: Institute of Mathematical Statistics

Stable URL: <http://www.jstor.org/stable/2676682>

Accessed: 24-05-2016 14:37 UTC

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at

<http://about.jstor.org/terms>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.



Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve and extend access to *Statistical Science*

Comment

D. R. Cox

Professor Breiman's interesting paper gives both a clear statement of the broad approach underlying some of his influential and widely admired contributions and outlines some striking applications and developments. He has combined this with a critique of what, for want of a better term, I will call mainstream statistical thinking, based in part on a caricature. Like all good caricatures, it contains enough truth and exposes enough weaknesses to be thought-provoking.

There is not enough space to comment on all the many points explicitly or implicitly raised in the paper. There follow some remarks about a few main issues.

One of the attractions of our subject is the astonishingly wide range of applications as judged not only in terms of substantive field but also in terms of objectives, quality and quantity of data and so on. Thus any unqualified statement that "in applications..." has to be treated sceptically. One of our failings has, I believe, been, in a wish to stress generality, not to set out more clearly the distinctions between different kinds of application and the consequences for the strategy of statistical analysis. Of course we have distinctions between decision-making and inference, between tests and estimation, and between estimation and prediction and these are useful but, I think, are, except perhaps the first, too phrased in terms of the technology rather than the spirit of statistical analysis. I entirely agree with Professor Breiman that it would be an impoverished and extremely unhistorical view of the subject to exclude the kind of work he describes simply because it has no explicit probabilistic base.

Professor Breiman takes data as his starting point. I would prefer to start with an issue, a question or a scientific hypothesis, although I would be surprised if this were a real source of disagreement. These issues may evolve, or even change radically, as analysis proceeds. Data looking for a question are not unknown and raise puzzles but are, I believe, atypical in most contexts. Next, even if we ignore design aspects and start with data,

key points concern the precise meaning of the data, the possible biases arising from the method of ascertainment, the possible presence of major distorting measurement errors and the nature of processes underlying missing and incomplete data and data that evolve in time in a way involving complex interdependencies. For some of these, at least, it is hard to see how to proceed without some notion of probabilistic modeling.

Next Professor Breiman emphasizes prediction as the objective, success at prediction being the criterion of success, as contrasted with issues of interpretation or understanding. Prediction is indeed important from several perspectives. The success of a theory is best judged from its ability to predict in new contexts, although one cannot dismiss as totally useless theories such as the rational action theory (RAT), in political science, which, as I understand it, gives excellent explanations of the past but which has failed to predict the real political world. In a clinical trial context it can be argued that an objective is to predict the consequences of treatment allocation to future patients, and so on.

If the prediction is localized to situations directly similar to those applying to the data there is then an interesting and challenging dilemma. Is it preferable to proceed with a directly empirical black-box approach, as favored by Professor Breiman, or is it better to try to take account of some underlying explanatory process? The answer must depend on the context but I certainly accept, although it goes somewhat against the grain to do so, that there are situations where a directly empirical approach is better. Short term economic forecasting and real-time flood forecasting are probably further exemplars. Key issues are then the stability of the predictor as practical prediction proceeds, the need from time to time for recalibration and so on.

However, much prediction is not like this. Often the prediction is under quite different conditions from the data; what is the likely progress of the incidence of the epidemic of v-CJD in the United Kingdom, what would be the effect on annual incidence of cancer in the United States of reducing by 10% the medical use of X-rays, etc.? That is, it may be desired to predict the consequences of something only indirectly addressed by the data available for analysis. As we move toward such more ambitious tasks, prediction, always hazardous, without some understanding of underlying process and linking with other sources of information, becomes more

D. R. Cox is an Honorary Fellow, Nuffield College, Oxford OX1 1NF, United Kingdom, and associate member, Department of Statistics, University of Oxford (e-mail: david.cox@nuffield.oxford.ac.uk).

and more tentative. Formulation of the goals of analysis solely in terms of direct prediction over the data set seems then increasingly unhelpful.

This is quite apart from matters where the direct objective is understanding of and tests of subject-matter hypotheses about underlying process, the nature of pathways of dependence and so on.

What is the central strategy of mainstream statistical analysis? This can most certainly not be discerned from the pages of *Bernoulli*, *The Annals of Statistics* or the *Scandinavian Journal of Statistics* nor from *Biometrika* and the *Journal of Royal Statistical Society, Series B* or even from the application pages of *Journal of the American Statistical Association* or *Applied Statistics*, estimable though all these journals are. Of course as we move along the list, there is an increase from zero to 100% in the papers containing analyses of “real” data. But the papers do so nearly always to illustrate technique rather than to explain the process of analysis and interpretation as such. This is entirely legitimate, but is completely different from live analysis of current data to obtain subject-matter conclusions or to help solve specific practical issues. Put differently, if an important conclusion is reached involving statistical analysis it will be reported in a subject-matter journal or in a written or verbal report to colleagues, government or business. When that happens, statistical details are typically and correctly not stressed. Thus the real procedures of statistical analysis can be judged only by looking in detail at specific cases, and access to these is not always easy. Failure to discuss enough the principles involved is a major criticism of the current state of theory.

I think tentatively that the following quite commonly applies. Formal models are useful and often almost, if not quite, essential for incisive thinking. Descriptively appealing and transparent methods with a firm model base are the ideal. Notions of significance tests, confidence intervals, posterior intervals and all the formal apparatus of inference are valuable tools to be used as guides, but not in a mechanical way; they indicate the uncertainty that would apply under somewhat idealized, may be very idealized, conditions and as such are often lower bounds to real uncertainty. Analyses and model development are at least partly exploratory. Automatic methods of model selection (and of variable selection in regression-like problems) are to be shunned or, if use is absolutely unavoidable, are to be examined carefully for their effect on the final conclusions. Unfocused tests of model adequacy are rarely helpful.

By contrast, Professor Breiman equates mainstream applied statistics to a relatively mechanical

process involving somehow or other choosing a model, often a default model of standard form, and applying standard methods of analysis and goodness-of-fit procedures. Thus for survival data choose a priori the proportional hazards model. (Note, incidentally, that in the paper, often quoted but probably rarely read, that introduced this approach there was a comparison of several of the many different models that might be suitable for this kind of data.) It is true that many of the analyses done by nonstatisticians or by statisticians under severe time constraints are more or less like those Professor Breiman describes. The issue then is not whether they could ideally be improved, but whether they capture enough of the essence of the information in the data, together with some reasonable indication of precision as a guard against under or overinterpretation. Would more refined analysis, possibly with better predictive power and better fit, produce *subject-matter* gains? There can be no general answer to this, but one suspects that quite often the limitations of conclusions lie more in weakness of data quality and study design than in ineffective analysis.

There are two broad lines of development active at the moment arising out of mainstream statistical ideas. The first is the invention of models strongly tied to subject-matter considerations, representing underlying dependencies, and their analysis, perhaps by Markov chain Monte Carlo methods. In fields where subject-matter considerations are largely qualitative, we see a development based on Markov graphs and their generalizations. These methods in effect assume, subject in principle to empirical test, more and more about the phenomena under study. By contrast, there is an emphasis on assuming less and less via, for example, kernel estimates of regression functions, generalized additive models and so on. There is a need to be clearer about the circumstances favoring these two broad approaches, synthesizing them where possible.

My own interest tends to be in the former style of work. From this perspective Cox and Wermuth (1996, page 15) listed a number of requirements of a statistical model. These are to establish a link with background knowledge and to set up a connection with previous work, to give some pointer toward a generating process, to have primary parameters with individual clear subject-matter interpretations, to specify haphazard aspects well enough to lead to meaningful assessment of precision and, finally, that the fit should be adequate. From this perspective, fit, which is broadly related to predictive success, is not the primary basis for model choice and formal methods of model choice that take no account

of the broader objectives are suspect in the present context. In a sense these are efforts to establish data descriptions that are potentially causal, recognizing that causality, in the sense that a natural scientist would use the term, can rarely be established from one type of study and is at best somewhat tentative.

Professor Breiman takes a rather defeatist attitude toward attempts to formulate underlying processes; is this not to reject the base of much scientific progress? The interesting illustrations given by Beveridge (1952), where hypothesized processes in various biological contexts led to important progress, even though the hypotheses turned out in the end to be quite false, illustrate the subtlety of the matter. Especially in the social sciences, representations of underlying process have to be viewed with particular caution, but this does not make them fruitless.

The absolutely crucial issue in serious mainstream statistics is the choice of a model that will translate key subject-matter questions into a form for analysis and interpretation. If a simple standard model is adequate to answer the subject-matter question, this is fine: there are severe hidden penalties for overelaboration. The statistical literature, however, concentrates on how to do

the analysis, an important and indeed fascinating question, but a secondary step. Better a rough answer to the right question than an exact answer to the wrong question, an aphorism, due perhaps to Lord Kelvin, that I heard as an undergraduate in applied mathematics.

I have stayed away from the detail of the paper but will comment on just one point, the interesting theorem of Vapnik about complete separation. This confirms folklore experience with empirical logistic regression that, with a largish number of explanatory variables, complete separation is quite likely to occur. It is interesting that in mainstream thinking this is, I think, regarded as insecure in that complete separation is thought to be a priori unlikely and the estimated separating plane unstable. Presumably bootstrap and cross-validation ideas may give here a quite misleading illusion of stability. Of course if the complete separator is subtle and stable Professor Breiman's methods will emerge triumphant and ultimately it is an empirical question in each application as to what happens.

It will be clear that while I disagree with the main thrust of Professor Breiman's paper I found it stimulating and interesting.

Comment

Brad Efron

At first glance Leo Breiman's stimulating paper looks like an argument against parsimony and scientific insight, and in favor of black boxes with lots of knobs to twiddle. At second glance it still looks that way, but the paper *is* stimulating, and Leo has some important points to hammer home. At the risk of distortion I will try to restate one of those points, the most interesting one in my opinion, using less confrontational and more historical language.

From the point of view of statistical development the twentieth century might be labeled "100 years of unbiasedness." Following Fisher's lead, most of our current statistical theory and practice revolves around unbiased or nearly unbiased estimates (particularly MLEs), and tests based on such estimates. The power of this theory has made statistics the

dominant interpretational methodology in dozens of fields, but, as we say in California these days, it is power purchased at a price: the theory requires a modestly high ratio of signal to noise, sample size to number of unknown parameters, to have much hope of success. "Good experimental design" amounts to enforcing favorable conditions for unbiased estimation and testing, so that the statistician won't find himself or herself facing 100 data points and 50 parameters.

Now it is the twenty-first century when, as the paper reminds us, we are being asked to face problems that never heard of good experimental design. Sample sizes have swollen alarmingly while goals grow less distinct ("find interesting data structure"). New algorithms have arisen to deal with new problems, a healthy sign it seems to me even if the innovators aren't all professional statisticians. There are enough physicists to handle the physics case load, but there are fewer statisticians and more statistics problems, and we need all the help we can get. An

Brad Efron is Professor, Department of Statistics, Sequoia Hall, 390 Serra Mall, Stanford University, Stanford, California 94305-4065 (e-mail: brad@stat.stanford.edu).