the bootstrap. For example, the first edition of *Numerical Recipes: the Art of Scientific Computing* by Press et al. (1986) sketches, in Section 14.5, a construction of bootstrap "pivotal" confidence limits for model parameters. The book cites as references two astrophysical papers published in 1976. In comparing these astrophysical papers with Efron (1979a) and with later bootstrap work, one sees again the historical role of the statistician in formulating, sharpening and developing a primitive new data-analytic idea. The bootstrap is not just a notion inflicted by theoretical statisticians upon reluctant data analysts. Also the reverse holds. Incidentally, the second edition of *Numerical Recipes* cites Efron.

Broadcasting bootstrap methods requires updating statistical education. Education goes beyond accessible software, mentioned in statement (f). Many undergraduate statistics texts fail to treat the Behrens–Fisher problem adequately, let alone developments of recent decades such as nonparametric regression, statistical graphics, generalized linear models or bootstrap. Why? I suggest the following: (a) Comprehension of modern statistical methods benefits from an actual need to analyze complex data. (b) Statistical theory relies on the mathematics of the twentieth century. (c) Using modern statistical methods, such as bootstrap, is computer-intensive. Meeting these three requirements is not so easy in large undergraduate classes. However, computing costs continue to drop as PC's become more powerful; students face a growing need to analyze the ambient information flood; and careful analysis of simple cases can develop statistical intuition. Meanwhile, MA-level courses can be effective in spreading modern statistical ideas to students in other fields. On bootstrap methods, we now have several trustworthy monographs.

Statements (d) and (e) flirt with double-think. The main thrust of bootstrap research, from 1979 onward, has been to understand what form of bootstrap works for what kind of statistical model. Young himself mentions the steady development of bootstrap techniques for time-series analysis. In preprints, this time-series research dates back to at least 1988. The work on squeezing better performance from bootstrap methods that is denigrated in assertion (d) resolved problems neglected according to statement (e), and these results are part of the ongoing research into diagnostics of bootstrap reliability. It is a noteworthy success that intuitive bootstrap critical values achieve the good small-sample performance of Welch's solution to the Behrens–Fisher problem or, more generally, of the Bartlett adjustment to likelihood ratio confidence sets and tests.

Statement (g) illustrates the numbing effect of familiar terminology. The word "nonparametric" is a blind description of what is actually a function-valued parameter. The word "likelihood" is equally a misnomer. Consider the three parameter lognormal model—smooth in the parameters and possessing finite Fisher information—whose likelihood function climbs to infinity at a most unlikely place. Bootstrap and empirical likelihood are complementary techniques rather than competitors. For instance, after empirical likelihood determines the shape of a confidence region, bootstrap provides a more accurate critical value for that region.

I conclude by mentioning two useful references not cited in Young's essay. The proceedings of the 1990 Trier conference (Jöckel, Rothe and Sendler, 1992) contain papers on random number generation and Monte Carlo tests as well as on bootstrap theory and applications. Janas (1993) surveys some of the earlier work on bootstrapping time series.

Comment

B. Efron

> *"My general feeling about bootstrapping is that I don't like it very much. It's easy for me to say that, because nowadays I don't have to do practical problems for a living."*—Henry Daniels, *Statistical Science, August 1993.*

B. Efron is Professor of Statistics and Biostatistics, Statistics Department, Stanford University, Stanford, California 94305-4065.

In 1980 I gave a talk at Ann Arbor called "Six influential papers and what ever became of them." The six papers were classics of the postwar literature: Wilcoxon on rank tests, Huber on robust estimation, Robbins on empirical Bayes, James and Stein on shrinkage estimates, Cox on proportional hazards and Tukey on the jackknife variance estimate. The question raised in the talk, but not settled, was why two of these papers, Wilcoxon's and Cox's, seemed to leap into applied use, while the others com-

paratively languished. I was particularly interested in the case of the jackknife. This was an elementary, nonparametric, completely automatic way of computing variances, which could not even make it into most nonparametric textbooks.

Alastair Young's nice article suggests that I should have saved some of my concern for the bootstrap. In fact the bootstrap (or perhaps, following Young, "the backstab") *is* used quite a bit and shows signs of considerably more employment in the near future. Nevertheless the article raises some provocative points concerning the relationship between statistical theory and statistical practice.

The bootstrap is a method for extending point estimates to more ambitious inferential statements such as confidence intervals or likelihoods. A probability model $\mathbf{P}$ involving many or an infinite number of unknown parameters produces some observed data $\mathbf{x}$. Often it is easy to construct a point estimate for $\mathbf{P}$ from $\mathbf{x}$, say, $\widehat{\mathbf{P}}$. For example, in a one-sample nonparametric problem, $\widehat{\mathbf{P}}$ is usually taken to be the empirical distribution of the data. We can then use the bootstrap to obtain from $\widehat{\mathbf{P}}$ standard errors and confidence intervals for a parameter of interest $\theta = t(\mathbf{P})$.

Most bootstrap research has focused on extracting accurate inferential statements from the point estimate $\widehat{\mathbf{P}}$. Monte Carlo techniques are often required for the extraction process, but that is not essential to the basic idea of the bootstrap. This research has been surprisingly successful. The 1975-model jackknife produced standard errors for smooth statistics, pretty much restricted to a one-sample nonparametric framework. The 1994 bootstrap, direct descendant of the jackknife, produces highly accurate confidence intervals in most parametric and many nonparametric situations.

Young nicely describes the vigorous expansion of the research effort: to more complicated models $\mathbf{P}$, alternative point estimates $\widehat{\mathbf{P}}$ and better ways of extracting general inferences from $\widehat{\mathbf{P}}$. A side benefit of the bootstrap work has been an increased interest in (or at least a tolerance of) other computer-intensive inferential methods such as Gibbs sampling, multiple imputation and empirical likelihoods.

Do I use the bootstrap in my own applied work? Yes, but not as much as I use the t-test, linear regression or the standard intervals $\widehat{\theta} \pm 1.645\widehat{\sigma}$. However, my bootstrapping has increased considerably with the switch to S, a modern interactive computing language. My guess is that the bootstrap (and other computer-intensive methods) will really come into its own only as more statisticians are freed from the constraints of batch-mentality processing systems like SAS.

The fact is that applied statisticians got along fine without the bootstrap before and can still do so now. The trouble with this statement is the definition of "got along fine." The statistics profession has been very successful in getting our clients to ask only those questions we can answer. Now we are prepared to answer harder questions, but the clients will not ask these until we tell them to do so.

Here is an example of what I mean, taken from preliminary calculations for Efron and Feldman (1991). A cholesterol-reducing drug was given to 165 men, many of whom took only a small fraction of the intended dose. The plotted points are

$$(x, y) = (\text{proportion dose taken, cholesterol decrease}).$$

It was important to estimate θ_{60}, the true cholesterol decrease at $x = 0.60$, the average dose taken.

The solid curve in Figure 1 is "lowess," a locally weighted smoother developed by W. S. Cleveland, available in S. It gives the estimate $\widehat{\theta}_{60}$ = lowess(0.60) = 33.99. Simple bootstrap replications, resampling the 165 (x, y) points, were used to assess the accuracy of $\widehat{\theta}_{60}$: 200 replications gave estimated standard error $\widehat{\sigma} = 4.25$; 2,000 replications showed a nearly normal histogram, with 90% BC_a confidence interval (26.8, 40.7) for θ_{60}. The entire analysis took about 30 minutes, including programming time and graphical printout.

The analysis in Efron and Feldman (1991) actually features ordinary least squares quadratic regression. This was a lot more familiar to me and

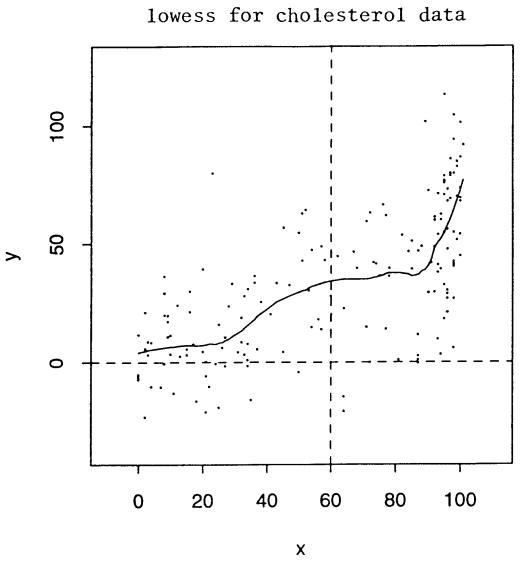

FIG. 1. *A cholesterol-reducing drug was tried on* 165 *men*; x = *proportion of intended dose taken*; y = *decrease in total cholesterol level from baseline measurement. Average compliance* = 0.60. *Solid curve is "lowess," a weighted moving average scatterplot smoother*; *window width* 0.3. *How accurate is the estimate* lowess(0.60) = 33.99? [*From preliminary calculations for Efron and Feldman* (1991).]

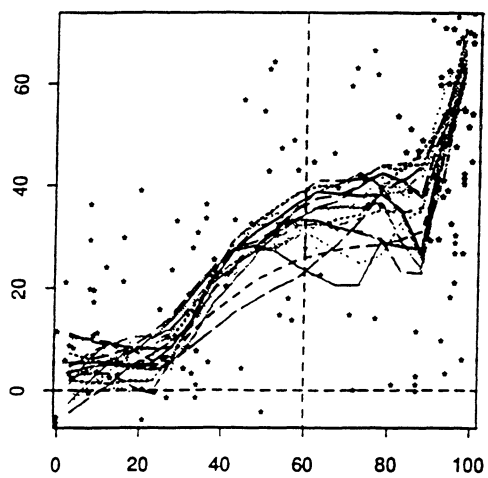

FIG. 2. *The first 20 bootstrap lowess curves; the sharp break at 0.85 seen in the original lowess curve is validated by the bootstrap replications.*

to the readers. Lowess is probably better for this situation. Figure 2 shows the first 20 of the 2,000 bootstrap lowess curves. The sharp break in the response function at $x = 0.85$ is a dependable feature of the replications. It is easy to quantify "dependable" with a bootstrap confidence interval for, say, $\gamma_{\text{break}} = \log((\theta_{100} - \theta_{85})/(\theta_{85} - \theta_{50}))$.

Without making too much of this small example, it does illustrate some encouraging trends in modern data analysis: more flexible fitting techniques than ordinary least squares polynomial regression; better confidence intervals than $\hat{\theta} \pm 1.645\hat{\sigma}$; and attention to "difficult" but interesting parameters like γ_{break}.

Theoreticians naturally focus on pathologies, which test a theory to its limits. Real applications tend less to be pathological than clumsy, awkward and difficult, as illustrated by the cholesterol example. In other words, they do not easily fit the simple mathematical models of classical statistical analysis. Computer-intensive methods like the bootstrap greatly extend the range of classical methods, and this is the way I believe that they will most dramatically affect 21st century statistics. Young's knowledgeable delineation of the limits of current bootstrap theory should not obscure an important fact: that these limits are already wide enough to permit a much more flexible approach to statistical practice.

Comment

Patricia M. Grambsch, Mary Kathryn Cowles and Thomas A. Louis

Young's review provides an informative history of the development of the bootstrap and discusses recent developments. We let others comment on technical issues, and briefly discuss Young's warnings related to the bootstrap. His principal worry is that the bootstrap invites mispractice by many users in that it has the reputation of an all-purpose procedure that will provide at least approximately valid inferences. Developers and generators of the bootstrap literature understand that, as with all statistical procedures, the bootstrap performs extremely well in many contexts (basically those where large-sample Gaussian asymptotics hold), but can fall on its face in nonregular contexts. Embellishments such as bias-correction and the nested bootstrap have improved small and moderate sample performance, but bring with them additional complications and decisions. Also, they strongly refute Efron's original claim that the bootstrap is "A statistical procedure devoid of intellectual content"!

Although Young's concerns are valid, are they any more compelling for the bootstrap than for other procedures such as the t-test, multiple regression or the Cox model? Our answer is both yes and no. Any statistical procedure frequently used will be frequently abused. Availability in a user-friendly computing package facilitates use and abuse. So, Young's criticisms unfairly single out the bootstrap. On the other hand, especially in its nonparametric, vanilla form, the bootstrap is relatively easy to apply to a limitless class of problems. All one has to do is decide on the sampling unit (or not decide and just get on with it), put the relevant data on actual or symbolic tokens and let the Monte Carlo run. Unlike the t-test, regression or Cox model, there are no explicit or implicit limits to the models or methods that comprise the "black box" around which one bootstraps. Sometimes the bootstrap will provide valid assessments of

Patricia M. Grambsch and Thomas A. Louis are faculty in the Division of Biostatistics, School of Public Health, University of Minnesota, Box 197 Mayo, Minneapolis, Minnesota 55455. Mary Kathryn Cowles is a faculty member in Preventive and Societal Medicine, University of Nebraska Medical Center, 600 South 42nd Street, Omaha, Nebraska 68198.