SINGH, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9** 1187–1195.
TU, D. S. (1992). Approximating the distribution of a generalized functional statistic with that of jackknife pseudo values. In *Exploring the Limits of Bootstrap* (R. LePage, and L. Billard, eds.) 279–306. Wiley, New York.
WESTFALL, P. H. and YOUNG, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley, New York.
WU, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). *Ann. Statist.* **14** 1261–1350.
WU, C. F. J. (1990). On the asymptotic properties of the jackknife histogram. *Ann. Statist.* **18** 1438–1452.
YOUNG, G. A. and DANIELS, H. E. (1990). Bootstrap bias. *Biometrika* **77** 179–185.

Comment

Rudolf Beran

G. Alastair Young's essay states as its theme: "We will discuss reasons why, though a theoretical success, the bootstrap may be judged to have been a less spectacular success in recent years than many might have expected or than should be possible." Expectations are a personal matter, not widely shared. Young's specific concerns include the following:

(a) "...bootstrap procedures which have been developed to handle more complex problems, such as those involving dependent data, are generally not automatic in that they require choice of some form of design parameter" (Section 3).
(b) "Patch-ups of the basic bootstrap involving devices such as modification of resampling size, while understood theoretically, suffer still from a lack of practicality" (Section 4.2).
(c) "Published applications of the bootstrap are now numerous..." but the latest discoveries of bootstrap theory have not made their way into such data analyses (Section 3).
(d) "Researchers have succumbed too much, perhaps, to the temptation to devote their efforts to squeezing even better performance from the bootstrap...rather than focusing their efforts on more fundamental issues concerning basic reliability of the approach" (Section 3).
(e) "Schenker (1985) illustrates the poor small-sample performance of procedures, which have asymptotic justification, when constructing [bootstrap] confidence intervals for a population variance" (Section 3). "Only recently has attention been paid to the practically crucial question of providing the user with some means of assessing how well-determined, or accurate, the bootstrap estimator is" (Section 4.2).
(f) "...there is still much theoretical analysis of bootstrap required before we can be confident of its value. Second, there is need for readily accessible software" (Section 6).
(g) "The very term 'bootstrap,' rightly or wrongly, evokes qualms with many, as producing something out of nothing. Many will feel on firmer ground with nonparametric likelihood" (Section 7).

Let us examine these assertions more closely. Statement (a), that the bootstrap is not automatic, is surely true, more deeply than Young discusses. Data does not follow a statistical model. Random variables are a mathematical construct, as are stationary time series and more complex models. The goal of statistical theory is to analyze procedures in hypothetical situations that mimic aspects of data. Even the most complete theory is easily misapplied. The first part of statement (f) founders on this reality. The use of bootstrap or other statistical procedures, like the use of surgical instruments, is an *empirical* business that offers no guarantees or refunds. This does not preclude success in skilled hands.

Statement (b) hastens to judge a very active topic. The modification of bootstrap resampling size has received closer scrutiny in recent technical reports by D. Politis and J. Romano and in a prominent invited lecture by F. Götze at the 1993 Annual Meeting of the IMS. The study of the wild bootstrap and generalized bootstrap is likewise moving ahead rapidly, for instance, in work by E. Mammen. Each of these strategies handles examples where simple bootstrapping fails. Early numerical results support the theory.

Statement (c) can be set against the prehistory of

Rudolf Beran is Professor of Statistics, Statistics Department, University of California at Berkeley, Berkeley, California 94720.

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-