THE BOOTSTRAP
AND
MODERN STATISTICS

by

Bradley Efron

Technical Report No. 206
August 1999

Division of Biostatistics
STANFORD UNIVERSITY
Stanford, California
THE BOOTSTRAP
AND
MODERN STATISTICS

by

BRADLEY EFRON

Department of Statistics
Stanford University

Technical Report No. 206

August 1999

This research was supported in part by
National Science Foundation grant DMS-9504379 and
National Institutes of Health grant 2R01 CA59039

Division of Biostatistics
STANFORD UNIVERSITY
Stanford, California
The Bootstrap and Modern Statistics
Bradley Efron

I once began an address to a mathematics conference with the following preposterous question: suppose you could buy a really fast computer, one that could do not a billion calculations per second, not a trillion, but an infinite number. So after you unpacked it at home you could numerically settle the Riemann hypothesis, the Goldbach conjecture, Fermat’s last theorem (this was a while ago) and still have time for breakfast. Would this be the end of mathematics?

My question wasn’t a very tactful one but its intentions were honorable. I was trying to communicate the current state of statistical theory. From a pre-World War II point of view our current computational abilities are effectively infinite, at least in terms of answering many common questions that arise in statistical practice. And no, this hasn’t spelled the end of statistical theory, though it certainly has changed, for the better in my opinion, what constitutes a good question and a good answer.

The bootstrap provides striking verification for the “infinite” capabilities of modern statistical computation. Figure 1 shows a small but genuine example, discussed more carefully in DiCiccio and Efron (1996) and Efron (1998). Twenty AIDS patients have received an experimental anti-viral drug, with the results shown in the left panel. The Pearson sample correlation coefficient between the 20 (Before, After) pairs of measurements is \( \hat{\theta} = .723 \). What inferences can we draw concerning the true population correlation \( \theta \)?

An immense amount of pre-war effort, much of the best by Fisher himself, was devoted to answering this

![Figure 1](image-url) 20 AIDS patients received an experimental anti-viral drug; left panel shows their CD4 counts in hundreds before and after treatment; Pearson correlation coefficient \( \hat{\theta} = .723 \). Right Panel histogram of 4000 nonparametric bootstrap replications \( \hat{\theta}^* \); solid lines indicate central 90% BC\( \alpha \) interval for true correlation \( \theta \); dashed lines are standard interval endpoints. Bootstrap standard error .0921, compared with nonparametric delta method .0795.
question. Most of this effort assumed a bivariate Gaussian probability model, the classic example being Fisher's z-transform for normalizing the correlation distribution. The bivariate Gaussian model, a poor fit to the AIDS data, was pushed far beyond its valid range because there was essentially no alternative.

An exception to this statement, almost the only one, was the nonparametric delta method estimate of standard error, given in terms of sample central moments of various mixed powers by this heroic formula:

\[
\hat{\sigma} = \sqrt{n} \left\{ \hat{\mu}_{22} + \frac{1}{4} \left( \hat{\mu}_{40} + \hat{\mu}_{04} + 2\hat{\mu}_{22} \right) - \left( \frac{\hat{\mu}_{31}}{\hat{\mu}_{11}\hat{\mu}_{20}} + \frac{\hat{\mu}_{13}}{\hat{\mu}_{11}\hat{\mu}_{02}} \right) \right\}^{1/2}.
\]  

(1)

Formulas like (1) were an important part of the applied statistician's tool-kit, heavily used for approximating standard errors, confidence intervals, and hypothesis tests. They still are (sometimes unfortunately) even though we now are armed with more potent weaponry.

The power of modern computation is illustrated in the right panel of Figure 1. It shows the histogram of 4000 nonparametric bootstrap correlation coefficients \( \hat{\theta}^* \). Each \( \hat{\theta}^* \) was calculated by drawing 20 points at random, with replacement, from the 20 actual data points in the left panel, and then computing the Pearson sample correlation coefficient for this bootstrap data set. (A variant of this algorithm would have been used if we wished to bootstrap the Gaussian model.) In total, about 1,000,000 elementary numerical calculations were required. This is less than one second of effort on a modern computer, even a small one, or perhaps a minute if, like me, you prefer to trade some speed for the programming ease of a high level language like Splus. The same computation on Fisher's "millionaire" mechanical calculator would have taken years of grinding human effort. Calling today's computers "infinite" isn't hyperbole from this point of view.

The sample standard deviation of the 4000 \( \hat{\theta}^* \) values was .0921, which is the bootstrap estimate of standard error for \( \hat{\theta} \). (4000 is at least 10 times too many for a standard error, but not excessive for the confidence interval discussion below.) This compares with .0795 from formula (1). An immense amount of effort has been spent justifying the theoretical basis of the bootstrap, more than 1000 papers since Efron (1979), but the basic principle is simple, amounting in this case to an application of nonparametric maximum likelihood estimation: (1) We suppose that the data have been obtained by random sampling from some unknown probability distribution \( F \) (a bivariate distribution in the AIDS example.) (2) We are estimating the parameter of interest \( \theta \) with some statistic \( \hat{\theta} \). (3) We wish to know \( \sigma_F \), the standard error of \( \hat{\theta} \) when sampling from \( F \). (4) We approximate \( \sigma_F \) with \( \sigma_{\hat{F}} \), where \( \hat{F} \) is the empirical distribution of the data (putting probability 1/20 on each of the 20 data points in the AIDS example).

The Monte Carlo routine for \( \hat{\theta}^* \) is just a way of evaluating \( \sigma_{\hat{F}} \) without going through the kind of Taylor series approximations involved in (1). In addition to being easier to use and more accurate than the Taylor Series approach, it has the great advantage of being completely general. I could just as well have bootstrapped Kendall's tau, or the largest eigenvalue of the sample covariance matrix,
or the ratio of 25% trimmed means. This generality allows the statistician to step fearlessly off the narrow path of pre-war computational feasibility, opening the door to more flexible, realistic, and powerful data analyses. Computer-based methods, including the bootstrap, have made statisticians more useful to our scientific colleagues.

The bootstrap began life as a muscle-ized big brother to the Quenouille-Tukey jackknife, with the same principal tasks in mind, the routine calculation of biases and standard errors. A more ambitious goal soon pushed itself forward: the automatic computation of bootstrap confidence intervals. A theory of confidence intervals more useful than the "standard intervals" $\hat{\theta} \pm 1.645 \hat{s}$, has to operate at an increased level of theoretical accuracy. Some of the deepest parts of the 1000-paper literature concern "second-order accuracy" and how it can be obtained via the bootstrap. Many authors participated in this work, as documented in the references.

The solid lines in Figure 1 marked .05 and .95 indicate the endpoints of the "$BC_a$" bootstrap confidence interval, Efron (1987), intended to cover the true correlation $\theta$ with 90% probability. $BC_a$ stands for bias-corrected and accelerated, enough words to suggest some difficulty in pursuing the goal of second-order accuracy. Hall (1988) verified $BC_a$'s second-order accuracy, by which is meant that the actual non-coverage probabilities, intended to be 5% at each end of the interval, approach that ideal with error proportional to $1/\text{sample size}$. This is an order of magnitude better than the $1/\sqrt{\text{sample size}}$ rate of the standard interval, indicated by the dashed lines in Figure 1.

The .05 and .95 lines in Figure 1 are not the 5th and 95th percentiles of the 4000 bootstrap replications. In this case they are the 4.6th and 94.7th percentiles, though in other examples the disparity could be much bigger. Using the "obvious" percentiles of the bootstrap distribution destroys second-order accuracy. The actual $BC_a$ percentile depend on an automatic algorithm that takes into account the bias and changing variance of $\hat{\theta}$ as an estimator of $\theta$.

I am going on a bit about the somewhat technical point because it reflects an important, and healthy, aspect of bootstrap research: the attempt to ground the bootstrap in the fundamental ideas of statistical theory, in this case coverage accuracy of confidence intervals. New statistical methodology is often applied promiscuously, more so if it is complicated, computer-based, and hard to check. The process of connecting it back to the basic principles of statistical inference comes later, but in the long run no methodology can survive if it flouts those principles. (The criticism process for bootstrap confidence intervals is still going strong, see Young (1994) and its discussion.)

The bootstrap itself was first intended as an explanation for the success of an older methodology, the jackknife.

Fisher and his colleagues were well aware that the standard intervals gave poor results for the correlation coefficient. This was the impetus for Fisher's z-transformation. But the z-transformation only fixes up the standard intervals for the Gaussian correlation coefficient while similar breakdowns, usually unrecognized, occur in many other contexts. Bootstrap confidence intervals automate the z-transform idea, bringing it to bear in a routine way on any estimation problem. The process of
grounding the bootstrap in traditional theory has worked the other way too: quite a bit more has been learned about the theory of confidence intervals through the effort of applying it outside the traditional textbook examples.

This same two-way exchange between classic statistical theory and modern computer-based methodology is going on in other areas of research. Markov Chain Monte Carlo (MCMC) offers a particularly apt example. If the bootstrap is an automatic processor for frequentist inference then MCMC is its Bayesian counterpart. The ability to compute a posteriori distributions for almost any prior, not just mathematically convenient ones, has deepened the discussion of what those priors should be. The renewed interest in "uninformative" priors, as a Kass and Wasserman (1996), which connects back to the theoretical basis of bootstrap confidence intervals. Sections 6-8 of Efron (1998) speculates about these connections.

Today's computers may indeed seem infinitely fast when carrying out traditional statistical calculations. Not so though for the more ambitious data-analytic tasks suggested by modern techniques like MCMC and the bootstrap. The possibility of improved results, and the critical appraisal of just how much improvement has been achieved, creates a demand for still better and inevitably more computationally intensive methodology. Bootstrap confidence intervals, usually an improvement over the traditional $\hat{\theta} \pm 1.645 \hat{\sigma}$, may still not give very accurate coverage in a small-sample nonparametric situation like Figure 1. Getting up to "third-order accuracy" seems to require bootstrapping the bootstrap, as in Loh (1987) and Beran (1987).

There is some sort of law working here whereby statistical methodology always expands to strain the current limits of computation. Our job is to make certain that the new methodology is genuinely more helpful to our scientific clientele, and not just more elaborate. I would give the statistics community a strong "A" in this regard. Here is a list, from Efron (1995), of a dozen post-war developments that have had a major effect on the practice of statistics: nonparametric and robust methods, Kaplan-Meier and proportional hazards, logistic regression and Generalized Linear Models, the jackknife and bootstrap, EM and MCMC, and empirical Bayes and James-Stein estimation.

These topics have something less healthy in common: none of them appear in most introductory statistics texts. As far as what we're teaching new students, statistics stopped dead in 1950. An obvious goal, but one that gets lost in an historical approach to our subject, is to insert intuitively simple and appealing topics like the bootstrap into the introductory curriculum.

My own education is applied statistics, a very good one in the hands of Lincoln Moses, Rupert Miller, and Byron Brown, was heavily classical. It has taken me a long time to get over the feeling that there is something magically powerful about formulas like (1) and to start trusting in the efficacy of computer-based methods like the bootstrap for routine calculations. It has been an easier transition for non-routine analyses, where classical methods do not exist, though I still find it easy to forget that today we can answer questions once utterly beyond reach.
Figure 2 relates to a recent consulting experience. The figure shows data for the first five patients of an efficacy study on an experimental anti-viral drug. There were 49 patients in the study, each measured on 43 predictor variables and a response, altogether comprising a 49 x 44 data matrix \( X \). The goal of the study was to predict the responses from some simple function of the 43 covariates. An extensive application of step-up and step-down regression selection programs, supplemented by the scientific intuition of the investigators, resulted in a “best” model that used just three simple linear combinations of the covariates (like the sum of the “x” measurements) while giving quite accurate predictions, \( R^2 = .73 \). A reviewer for the medical journal asked how optimistic this \( R^2 \) might be given the amount of data-mining employed.

Our answer was based on a fundamentally straightforward bootstrap analysis: (1) Construct a bootstrap data matrix \( X^* \), 49 x 44, by resampling the rows of \( X \) (that is by resampling the patients). (2) Rerun the step-up step-down regression selection programs \( X^* \), including some allowance for the guidelines of "scientific intuition", producing a bootstrapped best prediction rule, sometimes one much different than the original rule. (3) Compute \( \Delta R^2^* \), the difference in predictive ability for the bootstrapped rule on its own bootstrap data set \( X^* \) minus its predictive ability in the original data \( X \). The average value of \( \Delta R^2^* \) over 50 bootstrap replications, which turned out to be .12, then gave a believable assessment of optimism for the original \( R^2 = .73 \), leaving us with a bias-corrected estimate of \( R^2 = .61 \). Efron and Gong (1983) discuss a more elaborate example of predictive bias-correction.

The pre-history of the bootstrap is heavily involved with the jackknife. Rupert Miller’s influ-
ential 1964 paper "A trustworthy jackknife" was a successful early effort at de-mystifying what had seemed to be an almost magical device. Miller and I shared a sabbatical year at Imperial College in 1972-73, and after one of Rupert's lectures David Cox asked me, in a pointed way, if I thought there was anything to this jackknife business. I took this, correctly, as a hint, and a few years later decided to make an investigation of the jackknife the subject of the 1977 Rietz lecture.

An elaborate mechanism called "the combination distribution," was to be the basis of my lecture, but the more I worked on it the less remained of the mechanism, until I was left with what seemed at the time a disappointingly simple device. One of the most helpful references for this work was Jaeckel (1972), unfortunately unpublished, which suggested the kind of $\sigma_F$ explanation given above.

The lecture, which became the basis of Efron (1979), was given at the Seattle joint statistical meetings, accompanied by insistent construction noise from the next room. At the end of the lecture Professor J. Wolfowitz asked me if I had any theorems to back up the bootstrap, to which I could only respond that I didn't want to spoil a perfect effort. The name "bootstrap", suggested by Muenchauseh's fable, was chosen for euphony with "jackknife", and I was subsequently very happy to have given up on "combination distribution". Some alternative names are reviewed in the acknowledgment of the 1979 paper.

The four books referenced below, Davison and Hinkley (1997), Shao and Tu (1995), Efron and Tibshirani (1993), and Hall (1992), provide different views of the bootstrap, and also extensive bibliographies. Influential papers include Hall (1988), Bickel and Friedman (1981), Singh (1981), and Romano (1988), but this short list excludes so many of even my personal favorites that I can only fall back on space limitations as an apology.
References


Davison, A. and Hinkley, D. (1997), Bootstrap Methods and Their Application, Cambridge Univ. Press, N.Y.


Efron, B (1987), "Better bootstrap confidence intervals" (with discussion), JASA 82, 171-200.


Young, A. (1994), "Bootstrap: more than a stab in the dark?" (with discussion), Statistical Science 9, 382-415.