Seber, G. A. F. (1977), *Linear Regression Analysis*, New York: Wiley.

Seely, J. (1970), "Linear Spaces and Unbiased Estimation," *Annals of Mathematical Statistics*, 41, 1725–1734.

Stein, M. (1999), *Interpolation for Spatial Data: Some Theory for Kriging*, New York: Springer-Verlag.

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, New York: Wiley.

Zeger, S., Liang, K-Y., and Albert, P. S. (1988), "Models for Longitudinal Data: A Generalized Estimating Equation Approach," *Biometrics*, 44, 1049–1060.

Zhang, Q. (1995), "Wavelets and Regression Analysis," in *Wavelets and Statistics*, eds. A. Antoniadis and G. Oppenheim, New York: Springer-Verlag, pp. 397–407.

The Bootstrap and Modern Statistics

Bradley EFRON

I once began an address to a mathematics conference with the following preposterous question: Suppose that 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, and Fermat's last theorem (this was a while ago), and still have time for breakfast. Would this be the end of mathematics?

My question was not 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 standpoint, our current computational abilities *are* effectively infinite, at least in terms of answering many common questions that arise in statistical practice. And no, this has not 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 by DiCiccio and Efron (1996) and Efron (1998). Twenty AIDS patients received an experimental antiviral drug. 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 θ?

An immense amount of prewar effort, much of the best by Fisher himself, was devoted to answering this 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:

$$\widehat{\mathrm{SE}} = \frac{\hat{\theta}}{\sqrt{n}} \left\{ \frac{\hat{\mu}_{22}}{\hat{\mu}_{11}^2} + \frac{1}{4} \left(\frac{\hat{\mu}_{40}}{\hat{\mu}_{20}^2} + \frac{\hat{\mu}_{04}}{\hat{\mu}_{02}^2} + \frac{2\hat{\mu}_{22}}{\hat{\mu}_{20}\hat{\mu}_{02}} \right) - \left(\frac{\hat{\mu}_{31}}{\hat{\mu}_{11}\hat{\mu}_{20}} + \frac{\hat{\mu}_{13}}{\hat{\mu}_{11}\hat{\mu}_{02}} \right) \right\}^{1/2}. \quad (1)$$

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 Figure 1(b), which shows the histogram of 4,000 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 had wished to bootstrap the Gaussian model.) In total, about 1,000,000 elementary numerical calculations were required. This is less than 1 second of effort on a modern computer, even a small one, or perhaps 1 minute if, like me, you prefer to trade some speed for the programming ease of a high-level language like S-PLUS. The same computation on Fisher's "millionaire" mechanical calculator would have taken years of grinding human effort. Calling today's computers "infinite" is not hyperbole from this standpoint.

The sample standard deviation of the $4,000\hat{\theta}^*$ values was .0921, which is the bootstrap estimate of standard error for $\hat{\theta}$. (Here 4,000 is at least 10 times too many for a standard error, but not excessive for the confidence interval discussion to come.) This compares with .0795 from (1). An immense amount of effort has been spent justifying the theoretical basis of the bootstrap (more than 1,000 papers since Efron 1979), but the basic principle is simple, amounting in this case to an application of nonparametric maximum likelihood estimation:

Bradley Efron is Professor of Statistics and Biostatistics, Department of Statistics, Stanford University, Stanford, CA 94305-4065 (E-mail: *brad@stat.stanford.edu*).

© 2000 American Statistical Association
Journal of the American Statistical Association
December 2000, Vol. 95, No. 452, Vignettes

This content downloaded from
155.246.151.38 on Wed, 12 Oct 2022 01:09:24 UTC
All use subject to https://about.jstor.org/terms

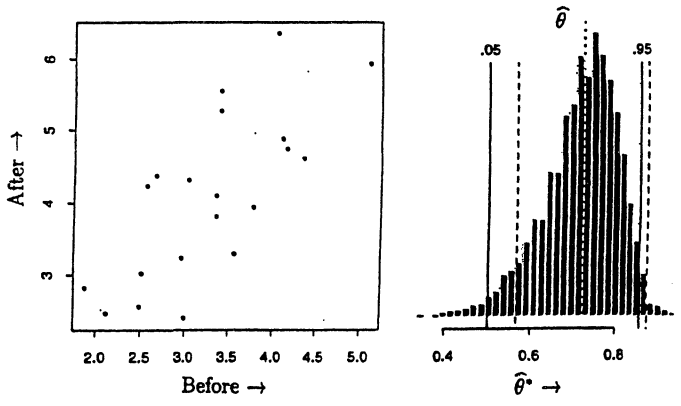

Figure 1. Study of 20 AIDS Patients Receiving an Experimental Antiviral Drug. Left Panel shows Their CD4 counts in hundreds before and after treatment; Pearson correlation coefficient $\hat{\theta} = .723$. Right Panel Histogram of 4,000 nonparametric bootstrap replications $\hat{\theta}^*$, with solid lines indicating the central 90% BC_a interval for true correlation θ and the dashed lines indicating standard interval endpoints. The bootstrap standard error is .0921, compared to the nonparametric delta method standard error .0795.

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 θ with some statistic $\hat{\theta}$.
3. We wish to know σ_F, the standard error of $\hat{\theta}$ when sampling from F.
4. We approximate σ_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 prewar 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 muscularized big brother to the Quenouille–Tukey jackknife (see Efron and Tibshirani 1993, chap. 11), with the same principal tasks in mind: routine calculation of biases and standard errors. A more ambitious goal soon pushed itself forward: automatic computation of bootstrap confidence intervals. A theory of confidence intervals more useful than the "standard intervals" $\hat{\theta} \pm 1.645\widehat{\mathrm{SE}}$ must operate at an increased level of theoretical accuracy. Some of the deepest parts of the 1,000-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 θ 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, meaning that the actual noncoverage 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 4,000 bootstrap replications. In this case they are the 4.6th and 94.7th percentiles, though in other examples the disparity could be much larger. 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 θ.

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 these 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 (see, e.g., Kass and Wasserman 1996), connects back to the theoretical basis of bootstrap confidence intervals. Efron (1998, secs. 6–8) speculated about these connections.

Today's computers may indeed seem infinitely fast when carrying out traditional statistical calculations. Not so

This content downloaded from
155.246.151.38 on Wed, 12 Oct 2022 01:09:24 UTC
All use subject to https://about.jstor.org/terms

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, create a demand for still better and inevitably more computationally intensive methodology. Bootstrap confidence intervals, usually an improvement over the traditional $\hat{\theta} \pm 1.645\widehat{\mathrm{SE}}$, may still not give very accurate coverage in a small-sample nonparametric situation like that illustrated in Figure 1. Getting up to "third-order accuracy" seems to require bootstrapping the bootstrap, as in Beran (1987) and Loh (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 postwar 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 appears in most introductory statistics texts. As far as what we are 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 in 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 nonroutine analyses, where classical methods do not exist, though I still find it easy to forget that today we can answer questions that once were 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 antiviral drug. There were 49 patients in the study, each measured on 43 predictor variables and a response, altogether creating a 49×44 data matrix $\mathbf{X}$. 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 used.

Our answer was based on a fundamentally straightforward bootstrap analysis:

1. Construct a bootstrap data matrix $\mathbf{X}^*, 49 \times 44$, by resampling the rows of $\mathbf{X}$ (i.e., by resampling the patients).
2. Rerun the step-up/step-down regression selection programs on $\mathbf{X}^*$, 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 ΔR^{2^*}, the difference in predictive ability for the bootstrapped rule on its own bootstrap data set $\mathbf{X}^*$ minus its predictive ability on the original data $\mathbf{X}$.

The average value of ΔR^{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) discussed a more elaborate example of predictive bias correction.

The prehistory of the bootstrap is heavily involved with the jackknife. Rupert Miller's influential article "A Trustworthy Jackknife" (Miller 1964) was a successful early effort at demystifying what had seemed to be an almost magical device. Miller and I shared a sabbatical year at Imperial College in 1972–1973, and after one of Rupert's lectures

patient	x30	x46	x48	x54	x82	x84	x90	numo	r41	r67	r69	r70	r74	r75	r151	r184
1	0	0	0	0	0	0	0	7	1	1	0	0	0	0	0	1
2	0	0	0	0	0	1	1	7	1	0	0	0	0	0	0	1
3	0	1	0	1	1	0	1	7	1	1	0	1	0	0	0	1
4	1	0	0	0	0	0	0	4	0	0	1	1	0	1	1	1
5	0	0	1	1	1	0	0	2	0	1	1	1	0	0	0	1

	r210	r215	r219	numto	age	sex	race	aids	sqv	rtv	ind	nfv	pin	dur	tc3
1	1	1	1	10	33	0	4	3	0	0	1	0	1	28	1
2	0	1	0	4	47	0	4	3	1	0	1	1	3	48	1
3	1	1	1	7	35	0	4	3	0	1	1	0	2	60	1
4	0	0	0	5	36	0	4	1	0	0	0	1	1	12	1
5	0	0	1	9	48	0	4	2	1	0	1	0	3	88	1

	azt	d46	ddc	ddi	duron	revdur	cdb	rnb	cd4	cd12	rn4	rn12	response
1	1	1	0	1	0	82	20	6	80	120	-3	-3	1.695
2	1	1	0	1	0	76	320	5	40	40	-1	0	2.070
3	1	1	0	0	0	108	170	6	-10	-10	0	0	0.640
4	1	1	0	1	0	312	310	5	-50	0	-2	0	2.630
5	1	1	0	1	0	465	150	4	10	10	0	0	0.645

Figure 2. Data for the First Five of 49 AIDS Patients, Each Measured on 43 Predictors and Response to an Experimental Antiviral Drug. An extensive data-mining effort produced a three-variable prediction model with $R^2 = .73$. How optimistic was the R^2 value?

This content downloaded from
155.246.151.38 on Wed, 12 Oct 2022 01:09:24 UTC
All use subject to https://about.jstor.org/terms

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 a technical report by Jaeckel (1972), unfortunately unpublished, which suggested the kind of σ_F explanation given earlier.

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 did not want to spoil a perfect effort. The name "bootstrap," suggested by Muenchausen'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 article.

Books by Davison and Hinkley (1997), Efron and Tibshirani (1993), Hall (1992) and Shao and Tu (1995) provide different views of the bootstrap, and also extensive bibliographies. Influential articles include those of Bickel and Friedman (1981), Hall (1988), Romano (1988), and Singh (1981), 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

Beran, R. (1987), "Prepivoting to Reduce Level Error of Confidence Sets," *Biometrika*, 74, 457–468.

Bickel, P., and Freedman, D. (1981), "Some Asymptotic Theory for the Bootstrap," *The Annals of Statistics*, 9, 1196–1217.

DiCiccio, T., and Efron, B. (1996), "Bootstrap Confidence Intervals" (with discussion), *Statistics of Science*, 11, 189–228.

Davison, A., and Hinkley, D. (1997), *Bootstrap Methods and Their Application*, New York: Cambridge University Press.

Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Philadelphia: SIAM.

——— (1987), "Better Bootstrap Confidence Intervals" (with discussion), *Journal of the American Statistical Association*, 82, 171–200.

——— (1995), "The Statistical Century," *RSS News*, 22(5), 1–2.

——— (1998), "R. A. Fisher in the 21st Century" (with discussion), *Statistics of Science*, 13, 95–122.

Efron, B., and Gong, G. (1983), "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *American Statistician*, 37, 36–48.

Efron, B., and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, New York: Springer.

Jaeckel, L. (1972), "The Infinitesimal Jackknife," Technical Report MM72-1215-11, Bell Laboratories.

Kass, R., and Wasserman, L. (1996), "The Selection of Prior Distributions by Formal Rules," *Journal of the American Statistical Association*, 91, 1343–1370.

Loh, W. Y. (1987), "Calibrating Confidence Coefficients," *Journal of the American Statistical Association*, 82, 155–162.

Miller, R. (1964), "A Trustworthy Jackknife," *Annals of Mathematical Statistics*, 39, 1598–1605.

Romano, J. (1988), "A Bootstrap Revival of Some Nonparametric Distance Tests," *Journal of the American Statistical Association*, 83, 698–708.

Shao, J., and Tu, D. (1995), *The Jackknife and the Bootstrap*, New York: Springer.

Singh, K. (1981), "On the Asymptotic Accuracy of Efron's Bootstrap," *The Annals of Statistics*, 9, 1187–1195.

Young, A. (1994), "Bootstrap: More Than a Stab in the Dark?" (with discussion), *Statistical Science*, 9, 382–415.

Prospects of Nonparametric Modeling

Jianqing FAN

1. INTRODUCTION

Modern computing facilities allow statisticians to explore fine data structures that were unimaginable two decades ago. Driven by many sophisticated applications, demanded by the need of nonlinear modeling and fueled by modern computing power, many computationally intensive data-analytic modeling techniques have been invented to exploit possible hidden structures and to reduce modeling biases of traditional parametric methods. These data-analytic approaches are also referred to as nonparametric techniques. For an introduction to these nonparametric techniques, see the books by Bosq (1998), Bowman and Azzalini (1997), Devroye and Györfi (1985), Efromovich (1999), Eubank (1988), Fan and Gijbels (1996), Green and Silverman (1994), Györfi, Härdle, Sarda, and View (1989), Hart (1997), Hastie and Tibshirani (1990), Müller (1988), Ogden (1997), Ramsay and Silverman (1997), Scott (1992), Silverman (1986), Simonoff (1996), Vidakovic (1999), Wahba (1990), and Wand and Jones (1995), among others.

An aim of nonparametric techniques is to reduce possible modeling biases of parametric models. Nonparametric techniques intend to fit a much larger class of models to reduce modeling biases. They allow data to search appropriate nonlinear forms that best describe the available data, and also provide useful tools for parametric nonlinear modeling and for model diagnostics.

Over the past three decades, intensive efforts have been devoted to nonparametric function estimation. Many new

Jianqing Fan is Professor, Department of Statistics, University of California, Los Angeles, CA 90095 (E-mail: *jfan@stat.ucla.edu*) and Professor of Statistics, Chinese University of Hong Kong. Fan's research was partially supported by National Science Foundation grant DMS-9804414 and a grant from the University of California at Los Angeles.

© 2000 American Statistical Association
Journal of the American Statistical Association
December 2000, Vol. 95, No. 452, Vignettes

This content downloaded from
155.246.151.38 on Wed, 12 Oct 2022 01:09:24 UTC
All use subject to https://about.jstor.org/terms