Second Thoughts on the Bootstrap

Bradley Efron

Abstract

This brief review article will appear in the issue of *Statistical Science* marking the 25th anniversary of the bootstrap. It concerns some of the theoretical and methodological aspects of the bootstrap, and how they might influence future work in statistics.

Second Thoughts on the Bootstrap

Bradley Efron

My first thoughts on the bootstrap centered around variance and bias estimation. This was natural enough given the bootstrap's roots in the jackknife literature, with Quenouille on bias and Tukey on variance setting the agenda. The oldest note I can find says simply "what is the jackknife an approximation to?" poor English but a good question that resulted in the 1977 Rietz lecture "Bootstrap methods: Another look at the jackknife". Louis Jaeckel's 1972 Bell Labs memorandum on the infinitesimal jackknife was particularly helpful in answering the approximation question.

Now its 25 years later and the bootstrap baby is old enough to be in grad school. I've had some second thoughts about the bootstrap, its strengths and weaknesses, its foundations, what it can and cannot do, what it might do in the future, and these second thoughts are what I'll talk about, briefly, here. This volume is full of excellent essays discussing and sometimes answering many of these questions in the context of authentic applications. So with apologies to the authors and the readers for any redundancy,here are a few comments and concerns.

The Plug-In Principle The diagram in Figure 1 describes a typical bootstrap application: An unknown probability model P, for example a logistic regression that depends on an unknown vector of coefficients, has yielded an observed data vector $\mathbf{x}$; from $\mathbf{x}$ we calculate a statistic $\widehat{\theta} = s(\mathbf{x})$ intended to estimate a parameter $\theta = t(P)$ of particular importance, perhaps one of the unknown coefficients. We are interested in $\widehat{\theta}$'s accuracy for estimating θ, with accuracy defined in terms of bias, variance, confidence intervals, prediction error, or some other such measure.

The right half of the diagram describes the "bootstrap world" (in David Freedman's picturesque terminology): $\widehat{P}$ is a point estimate of P, in the logistic regression example obtained perhaps by substituting maximum likelihood estimates for the unknown coefficients. $\widehat{P}$ yields bootstrap data vectors $\mathbf{x}^*$ and then bootstrap replications $\widehat{\theta}^* = s(\mathbf{x}^*)$. Since $\widehat{P}$ is completely known we can generate as many $\widehat{\theta}^*$'s as we want, or have time for, and use their observed variability to assess the accuracy of $\widehat{\theta}$. During the past 25 years an enormous amount of statistical research has investigated the validity of the bootstrap approach. For most models P and most statistics $\widehat{\theta}$ we know that the bootstrap standard deviation $sd_*\{\widehat{\theta}^*\}$ is a good estimator for the true standard deviation $sd\{\widehat{\theta}\}$, and likewise for other accuracy measures.

$$\begin{tikzcd}
& \textit{Real World} & & \textit{Bootstrap World} \\
P \arrow[r] & \mathbf{x} \arrow[d] \arrow[r, Rightarrow] & \widehat{P} \arrow[r] & \mathbf{x}^* \arrow[d] \\
& \widehat{\theta} & & \widehat{\theta}^*
\end{tikzcd}$$

Figure 1 Typical Bootstrap diagram; unknown probability model P gives observed data $\mathbf{x}$; we wish to know the accuracy of statistic $\widehat{\theta} = s(\mathbf{x})$ for estimating parameter of interest $\theta = t(P)$. Point estimate $\widehat{P}$ for P yields bootstrap data sets $\mathbf{x}^*$; accuracy is inferred from observed variability of bootstrap replications $\widehat{\theta}^* = s(\mathbf{x}^*)$.

The double arrow in Figure 1 indicates the estimation of P from $\mathbf{x}$. The utility of the bootstrap depends on the double arrow process being easy to execute. It is particularly easy in the one-sample nonparametric case, where a completely unknown probability distribution gives $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ by random sampling, in which case we can take $\widehat{P}$ to be the empirical distribution putting probability $1/n$ on each x_i. Simply stated, the bootstrap is a device for upgrading a point estimate for P to an accuracy estimate for θ. Point estimates $\widehat{P}$ are so ubiquitous it comes as a shock when, as in some versions of the proportional hazards model, point estimates don't exist.

Figure 1 exemplifies the *plug-in principle*: We travel from the real world to the bootstrap world simply by plugging in a point estimate $\widehat{P}$ for P.This is the only inference step. All other arrows on the right are exact analogs of those on the left. Plug-in methods are familiar friends in classical statistics, when for instance we estimate the standard deviation $[p(1-p)/n]^{\frac{1}{2}}$ of a binomial proportion $\widehat{p}$ by $[\widehat{p}(1-\widehat{p})/n]^{\frac{1}{2}}$. Fisher extended the same tactic to information calculations for MLE's, substituting $\mathcal{I}(\widehat{\theta})^{-\frac{1}{2}}$ for $\mathcal{I}(\theta)^{-\frac{1}{2}}$. Our advantage is that modern computers allow us to carry out the plug-in principle with impunity, calculating $\widehat{P} \rightarrow \mathbf{x}^* \rightarrow \widehat{\theta}^*$ by brute force.

How far can the plug-in principle be trusted? "Pretty far" is a reasonable summary of current bootstrap research. Simple bootstrap ideas, like resampling from the empirical distribution, work surprisingly well in a surprisingly large catalog of cases. Yet there are situations where plugging in starts to get worrisome.

Figure 2 concerns a genomics example. 1391 HIV viral genomes were collected from AIDS patients who were taking various protease inhibitor (PI) drugs. The data for each genome comprised 74 numbers, the amino acid present at each of 74 positions on the viral protease gene, recorded as 0 or 1 if the amino acid was or was not the usual one present at that position in wild-type HIV: 1's indicate mutations caused by the drug treatment. The investigators wondered which of the six different PI drugs were associated with which mutations. Complicating matters, a majority of the 1391 patients took more than one PI, the average being 2.05, with a few even taking all six.

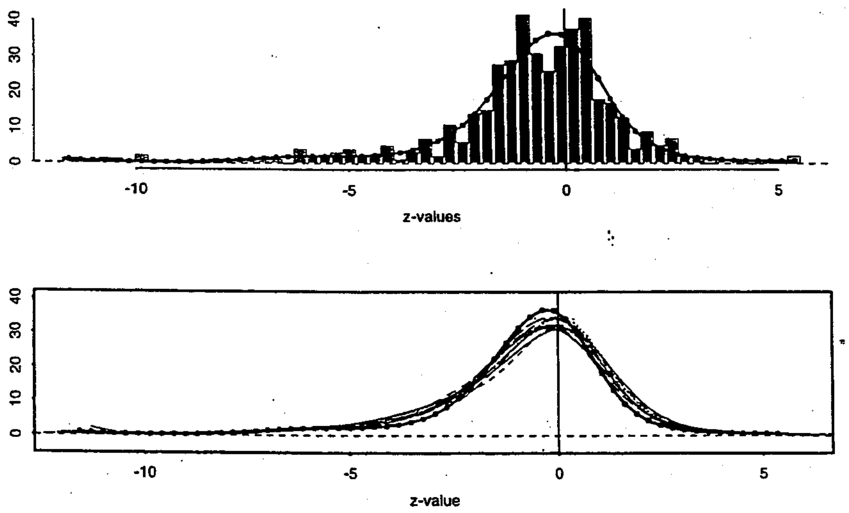

Figure 2 *Top panel*: histogram of z-values for 444 main effects, genome data logistic regression; beaded curve is spline fit to histogram counts. *Bottom panel*: First 10 of 50 bootstrap replications of spline fit; the replications tend to be wider than the original beaded curve.

A logistic regression having $444 = 6 \times 74$ main effects, one for each drug at each genome position, was fit to the 1391×74 0-1 amino acid responses. This gave the 444 z-values (coefficient estimate divided by standard error) whose histogram appears in the top panel of Figure 2. The beaded curve is a smooth Poisson GLM fit to the histogram counts, using a natural spline with seven degrees of freedom. The central peak is normal shaped with mean and standard deviation

$$\widehat{\mu} = -.38 \quad \text{and} \quad \widehat{\sigma} = 1.20,$$

$\widehat{\sigma}$ being computed from the curvature of the spline fit at $\widehat{\mu}$.

How accurately determined is the spline fit? The usual Poisson GLM standard errors are inappropriate since the 444 z-values, and therefore the histogram counts, are mutually correlated. Instead I applied the nonparametric one-sample bootstrap, with the 1391 genomes (each with its 74 numbers intact) as the resampling units. Each bootstrap data set gave bootstrap z-values, histogram, and natural spline fit. The bottom panel shows the first 10 of 50 bootstrap spline fits.

The 50 bootstrap estimates $\widehat{\sigma}^*$, each computed in the same way as the original $\widehat{\sigma} = 1.20$, had empirical mean 1.37 and empirical standard deviation 0.12. The value 0.12 is a reasonable estimate for the standard error of $\widehat{\sigma} = 1.20$, but in this case there is some cause for concern about the plug-in principle: 43 of the 50 $\widehat{\sigma}^*$'s exceeded $\widehat{\sigma}$; in the bottom panel we can see that the bootstrapped curves systematically exceed the width of the original curve.

It's easy to understand what's happening here: If the ith bootstrap z-value z_i^* has bootstrap mean and variance (z_i, v_i) (nearly true, with the v_i's roughly 1, except for z_i near 0 where they are smaller) then the empirical variance of the bootstrap histogram will be inflated by about $\bar{v}$. We could correct the $\widehat{\sigma}^{*2}$ values by subtracting $\bar{v}$ but this takes us beyond the realm of the plug-in principle.

The "dilation phenomena" in the bottom panel of Figure 2 occurs in classical situations, as with Stein estimation or the Neyman-Scott example. It points to a limitation of the plug-in principle, and the bootstrap, that I wish I understood better. In this case we get a warning from the bootstrap analysis, from the miscentering of the $\widehat{\sigma}^*$ values, but I'm not certain that other plug-in pathologies aren't possible, especially in situations involving a great many parameters. We seem to be living in a "great many parameters" era, which makes a critical examination of the plug-in principle especially timely.

Bootstrap Confidence Intervals A pleasant surprise of the early bootstrap literature was second order accuracy, developed originally in the 1981 papers of Singh and of Bickel and Freedman. Second order accuracy suggested that the bootstrap could provide good approximate confidence intervals, better than the usual "standard intervals" $\widehat{\theta} \pm z_\alpha \widehat{\sigma}$. The actual construction of such intervals looked like a formidable task: similar attempts employing the jackknife had failed.

In fact two classes of second-order accurate bootstrap confidence intervals have been developed, under the name "bootstrap t" and "BCA". These look different than each other, and can behave differently, but in fact are closely related as mentioned below. Neither method seems to be widely applied. People, even experienced statisticians, seem all too happy with the standard intervals $\widehat{\theta} \pm z_\alpha \widehat{\sigma}$, though they may use the bootstrap to get $\widehat{\sigma}$.

There's more at stake here than the term "second order" suggests. Table 1 concerns a simple example: 15 pairs of points $x_i = (y_i, z_i)$ were drawn from a bivariate normal distribution, giving sample correlation coefficient $\widehat{\theta} = 0.562$. Four types of confidence intervals were computed, (1) exact 90% intervals (non-coverage probability .05 in each tail) based on bivariate normal theory for correlation coefficients; (2) parametric ABC intervals, DiCiccio and Efron (1992), an analytic version of BCA; (3) nonparametric ABC, which assumes only that the points x_i are i.i.d. from an unknown bivariate distribution; and (4) standard intervals, using the normal-theory delta-method estimate of σ.

	LIMITS		TAIL AREAS	
	.05	.95	.05	.95
exact	0.155	0.790	0.050	0.950
parametric abc	0.158	0.788	0.051	0.948
nonparametric abc	0.188	0.775	0.063	0.935
standard	0.271	0.830	0.112	0.980

Table 1 Exact and approximate confidence intervals for correlation coefficient of a bivariate normal sample, $n = 15$, with sample correlation coefficient 0.562. Tail Area is actual probability of exceeding 0.562 when the parameter value is the corresponding interval endpoint.

The parametric ABC interval is almost exactly right in this case, while the nonparametric ABC interval is a little short in both directions. The standard interval is terrible, much too short to the right of $\widehat{\theta}$ and too long to the left. We can fix the standard intervals with Fisher's $\tanh^{-1}$ transformation, but in situations less familiar than the normal correlation coefficient neither a fix nor an exact solution will be available. Bootstrap intervals are always available, offering second-order accuracy on a routine basis.

Second-order accuracy isn't perfection. Bootstrap intervals are not exact and can be far from perfect in small-sample situations. Nonparametric intervals seem particularly vulnerable, the shortness seen in Table 1 being a typical performance. These are third-order errors, akin to dividing by n instead of $n - 1$ in variance estimation. Third order improvements may be just what are needed to nudge bootstrap confidence intervals into the widely-used category.

That being said, current bootstrap intervals, even nonparametric ones, are usually more accurate than their standard counterparts. "Accuracy" is a word that needs careful definition when applied to confidence intervals. The worst definition (seen unfortunately often in simulation studies of competing confidence interval techniques) concentrates on overall coverage. Even the standard intervals might come reasonably close to 90% overall coverage in Table 1's situation, but they will do so in lopsided fashion, failing to cover much more often than 5% on the left and much less than 5% on the right. The purpose of a two-sided confidence interval is accurate inference in both directions.

Coverage, even appropriately defined, is not the end of the story. Stability of the intervals, in length and location, is also important. Here is an example. Suppose we are in a standard normal situation where the exact interval is student's t with 10 degrees of freedom. Method A produces the exact 90% interval except shortened by factor .90; Method B produces the exact 90% interval either shortened by factor 2/3 or lengthened by factor 3/2, with equal probability. Both methods provide about 86% coverage, but Method B's intervals will always be substantially misleading.

The combination of maximum likelihood estimation and standard intervals made a profound contribution to scientific practice. Computation and theory are now in place for a substantially improved confidence interval methodology, but we seem to be one step short of making the sale to the scientific community.

Analytics Personally my biggest bootstrap surprise involved the ABC intervals developed with Tom DiCiccio in 1992. ABC is an analytic approximation to the BCA method, intended to cut down on the 2000 or so bootstrap simulations required for BCA. In fact, ABC involves no simulation at all, which was the surprise, especially since the method gives excellent results for smoothly differentiable statistics like the correlation coefficient.

I find myself coming back to the ABC method frequently because its formulaic structure is a great aid to theoretical calculations. For instance ABC leads to a nice connection between the BCA and bootstrap-t intervals, given in Section 5 of DiCiccio and Efron (1996).

Here's a simple example of ABC formulas: We observe $y \sim \text{Poisson}(\mu)$ and wish to compute α-level endpoints, say $\alpha = .05$ and .95, for μ's confidence interval. Defining

$$a = 1/(6\sqrt{y}) \quad \text{and} \quad w = a + \Phi^{-1}(\alpha),$$

the ABC endpoint is

$$y + \frac{w\sqrt{y}}{(1-aw)^2} \,.$$

For $y = 7$ this gives 90% interval $\mu \in [3.54, 12.67]$. The corresponding tail areas (i.e. probabilities of exceeding the observed value, including one-half the atom at 7, at the interval endpoints) are .0477 and .9523, gratifyingly close to the ideal values .05 and .95. Of course we don't need approximate intervals for the one-parameter Poisson family, but the formulas, which are useful even here, keep working in the great hinterland of cases where there are no exact solutions.

The standard intervals depend on estimates of two parameters, μ and σ. In addition ABC requires estimating three more parameters, the "acceleration" a, the "bias-correction" z_o, and the "nonlinearity coefficient" c_q. (In the Poisson example $\widehat{z}_o = \widehat{a} = (6\sqrt{y})^{-1}$ and $c_q = 0$.) Each of the three parameters corrects a first-order deficiency of the standard intervals, finally resulting in second order accuracy. It is of some theoretical interest that second-order accurate intervals require exactly five parameters. An important question, unanswered I believe, is how many parameters are required for third-order accuracy. Answering this question might also connect bootstrap theory more closely with the likelihood-based intervals developed by Barndorff-Neilson, Cox, Reid, and others, see Reid (1995).

The last few decades of statistical research can be broadly summarized as an immense amplification of classical theory via the power of electronic calculation. I've gone on a bit about the ABC method because it represents the reverse process: a return from computer algorithms to the classical world of formulas. Something has been gained in the round trip

$$\text{classical confidence intervals} \rightarrow \text{BCA algorithm} \rightarrow \text{ABC formulas}\ ,$$

namely a better theoretical understanding of the vast middle ground lying between exact intervals and the standard method.

This kind of reverse engineering could be important if we hope to expand the base of statistical theory beyond its classical limits. Computers enable us to explore a high dimensional statistical universe far outside classical boundaries, but how do we report back what we have found? Numerical summaries are end-products, perfectly appropriate in statistical applications but clumsy for theoretical investigations. Analysis and formulazation, the traditional approach but now applied to computer-based methods like the bootstrap, Markov chain Monte Carlo, empirical likelihood, or generalized additive models, could lead to a new round of progress in the fundamentals of statistics.

Prediction Problems Cross-validation, like the standard intervals, is a method of such obvious virtue that criticism seems almost churlish. Moreover its workhorse status in machine learning, as seen in the recent book by Hastie, Tibshirani, and Friedman, makes it a statistical success story in the outside world. Like the standard intervals, however, cross-validation is such a handy tool that it is easy to overuse. Can it be improved upon as an estimator of prediction error, perhaps in the way that the bootstrap intervals improve upon the standard method? There is some hope here: Cross-validation connects directly to the jackknife and bootstrap, as in Efron (1983), and in fact Efron and Tibshirani's 1996 paper shows the bootstrap-based "632+ rule" bettering cross-validation over a range of prediction problems.

What we don't have is a convincing theoretical bound that says how well a given estimate of prediction error is performing. Second-order accuracy provides such a bound for confidence intervals but the asymptotics are more delicate for prediction problems. It is easy to achieve the equivalent of first-order accuracy, as cross-validation does; what isn't available is theoretical reassurance that the numerical gains of methods like 632+ will hold up in general practice.

Bayesian Connections The bootstrap looks like a poor candidate for Bayesian duty. It was developed as an extension of a pure frequentist device, the jackknife, and itself violates the likelihood principle (since it depends on evaluating the statistic of interest for data sets other than the one observed.) Nevertheless Bayesian connections have persistently emerged, as in Rubin's 1981 "Bayesian bootstrap". The connection is closer to Jeffreys' objective Bayesian tradition than the subjectivist school, but it is an encouraging sign that there is any relationship at all. Two examples follow.

Figure 3 shows a phylogenetic tree charting the evolutionary history of 11 species of malaria parasite; Pme2 attacks lizards, Pfa4 is the most deadly form of human malaria, etc. The tree was constructed by applying a standard clustering algorithm to the 11 × 221 data matrix "**x**" composed of the aligned RNA base sequences at 221 sites along the malaria genome, see Efron, Halloran, and Holmes (1996). Some interesting relationships emerge, in particular the Pcy9-Pvi10 clade, which indicates a recent connection between primate and human malaria. However the tree is a statistic, admittedly a complicated one, and it is reasonable to ask how much trust we can place in its observed features.

Joseph Felsentstein proposed a bootstrap answer to this question in 1985: the columns of **x** are resampled, bootstrap trees are constructed, and we simply count the proportion of bootstrap trees having the feature of interest. For instance 193 of 200 bootstrap trees showed the Pcy9-Pvi10 clade, giving 193/200 or 96.5% bootstrap confidence in its validity, as indicated in Figure 3.

What is the statistical interpretation of Felsentstein's confidence values? Efron, Halloran, and Holmes discuss an objectivist Bayesian interpretation as well as standard confidence

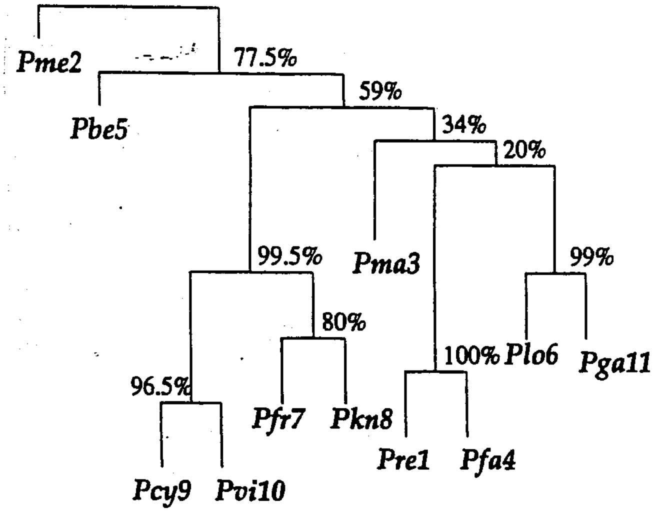

Figure 3 Phylogenetic tree for evolutionary history of malaria. Percentages indicate Felsenstein's bootstrap confidence estimates. For example there is 96.5% confidence that the Pcy9-Pvi10 clade is valid.

statements. Broadly speaking, vague prior opinions should lead to 96.5% posterior belief in the validity of the Pay9-Pvi10 clade. The problem is too complicated to actually construct an appropriate uninformative prior but the bootstrap calculations do so automatically. More sophisticated bootstrap resampling schemes can improve both the Bayesian and frequentist properties of the confidence values.

The "problem of regions", Efron and Tibshirani (1996) is a generalized statement of Felsenstein's problem. As a simple example suppose the plane is divided into checkerboard squares four units on a side, the squares being the regions, and that a single bivariate point $\mathbf{x} \sim N_2(\boldsymbol{\mu}, I)$ is observed, falling say into region $\mathbf{R}_1$. How confident should we be that $\boldsymbol{\mu}$ itself lies in $\mathbf{R}_1$? Felsenstein's tactic of resampling points $\mathbf{x}^* \sim N_2(\mathbf{x}, I)$and counting the proportion in $\mathbf{R}_1$ gives an answer something like an objective Bayesian posterior probability, which again can be improved upon with more sophisticated bootstrapping schemes.

Figure 4 illustrates another situation where the bootstrap can be used to carry through a Jeffreys'-type Bayesian analysis of a complicated problem. The data in this case consists of 5 test measurements on each of 22 students (one quarter of the data on page 4 of Mardia, Kent, and Bibby (1979).) The sample covariance matrix has ratio $\widehat{\theta} = 4.91$ of the first to second largest eigenvalue, and we wish to make inferences about the corresponding population parameter θ.

The curve in Figure 4 is an adjusted likelihood for θ computed from the output of the non-

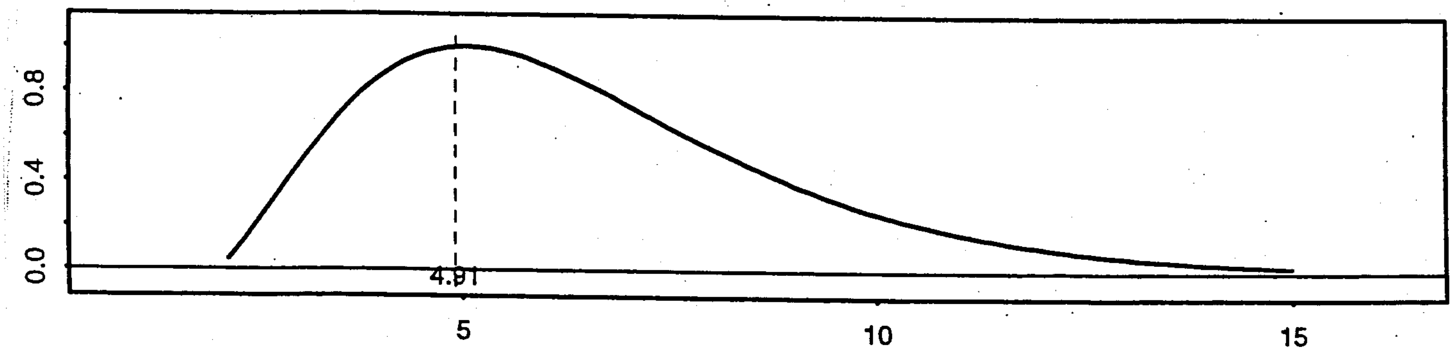

Figure 4 Relative likelihood of eigenvalue ratio parameter θ, student score data. The likelihood is longer-tailed to right of $\widehat{\theta} = 4.91$.

parametric ABC program; formula (6.12) of Efron (1993), a simple function of $(\widehat{\theta}, \widehat{\sigma}, \widehat{a}, \widehat{z}_o, \widehat{c}_q)$. We see that the likelihood function is much longer-tailed to the right of $\widehat{\theta}$. The theory behind Figure 4 equates this function with what we would obtain using a truly uninformative prior distribution (one whose *a posteriori* distributions had accurate coverage probabilities in the usual confidence interval sense.)

In fact it would be difficult to construct an uninformative prior for the eigenratio in a five-dimensional nonparametric situation. To this end the ABC likelihood is very convenient, and it can be helpful even in subjective Bayesian contexts: expert prior opinion about θ can be combined directly with the likelihood in Figure 4 using Bayes theorem, relieving the expert of the need to put a prior on the whole 5-dimensional space. Or, if we had a parallel collection of independent eigenratio problems, we could calculate the ABC likelihood for each, and combine them using empirical Bayes methods. See Efron (1996).

Final Remarks These days statisticians are being asked to analyze much more complicated problems, microarrays being the archetypal example. I believe, or maybe just hope, that a powerful combination of Bayesian and frequentist methodology will emerge to deal with this deluge of data, and that computer-intensive methods like the bootstrap will facilitate the combination. Intriguing theoretical questions are also hanging in the air. Why do the likelihood principle and the plug-in principle, which look antithetical, seem to co-exist peacefully in examples like that of Figure 4?

These remarks were based on my own frustrations and successes with the bootstrap during the past quarter century. A much broader point of view is represented in the essays that follow. It is striking how different the essays are from each other, and how different application areas have produced distinctive advances in bootstrap methodology. I am grateful to the authors, and especially to the editor George Casella, for throwing such a lively birthday party for the bootstrap.

References

Bickel, P. & Freedman, D. (1981), "Some asymptotic theory for the bootstrap", *Ann. Stat.* **9**, 1196-1217.

DiCiccio, T. & Efron, B. (1992), More accurate confidence intervals in exponential families. *Biometrika* **79**, 231-245.

DiCiccio, T. & Efron, B. (1996), "Bootstrap Confidence Internals", *Statistical Science* **11**, 189-212.

Efron, B. (1979), Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1–26.

Efron, B. (1983), Estimating the error rate of a prediction rule: improvements on cross–validation. *JASA* **78**, 316–331.

Efron, B. (1993), Bayes and likelihood calculations from confidence intervals. *Biometrika* **80**, 3-26.

Efron, B.(1996), Empirical Bayes methods for combining likelihoods. (with discussion), *JASA*, p. 538-565.

Efron, B., Halloran, E. & Holmes, S. (1996), Bootstrap confidence levels for phylogenetic trees. *Proceedings of the Nat'l Academy of Sciences* **93**, 13429-13434.

Efron, B. & Tibshirani, R. (1996), Improvements on cross-validation: the 632+ rule. *JASA*, p. 548-560.

Efron, B. & Tibshirani, R. (1998), The problem of regions, *Annal. Stat.*, 1287-1718.

Felsenstein, J. (1985), "Confidence limits on phylogenies: an approach using the bootstrap", *Evolution* **39**, 783-91.

Hastie, T., Tibshirani, R., & Friedman, J. (2001), *The Elements of Statistical Learning*, Springer, New York.

Jaeckel, L. (1972), "The Infinitesimal Jackknife," *Bell Lab. Memorandum*, MM72-1215-11.

Mardia, K., Kent, J., & Bibby, J. (1979), *Multivariate Analysis*, New York: Academic Press.

Quenouille, M. (1949), "Approximate tests of correlation in time series", *Jour. Roy. Stat. Soc. B*, **11**, 18-44.

Reid, N. (1995), "The roles of conditioning in inference", (with discussion), *Statistical Sci.* **10**, 138-199.

Rubin, D. (1981), "The Bayesian bootstrap", *Ann. Stat.* **9**, 130-134.

Singh, K. (1981), "On the asymptotic accuracy of Efron's bootstrap", *Ann. Stat.* **9**, 1187-95.

Small, C. & Murdoch, D. (1993), "Nonparametric Neyman-Scott problems: Telescoping product methods", *Biometrika* **80**, 763-79.

Tukey, J. (1958), "Bias and confidence in not quite large samples", (Abstract), *Ann. Math. Stat.* **29**, 614.