FRIEDMAN, J. H. and TIBSHIRANI, R. J. (1984). The monotone smoothing of scatter-plots. *Technometrics* **26** 243–250.
HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393.
HASTIE, T. J. and TIBSHIRANI, R. J. (1985). Discussion of Peter Huber's "Projection Pursuit." *Ann. Statist.* **13** 502–508.
HINKLEY, D. V. (1978). Improving the jackknife with special reference to correlation estimation. *Biometrika* **65**, 13–22.
HYDE, J. (1980). *Survival Analysis with Incomplete Observations. Biostatistics Casebook.* Wiley, New York.
JAECKEL, L. (1972). The infinitesimal jackknife. Memorandum MM 72-1215-11. Bell Laboratories, Murray Hill, New Jersey.
JOHNSON, N. and KOTZ, S. (1970). *Continuous Univariate Distributions.* Houghton Mifflin, Boston, Vol. 2.
KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete samples. *J. Amer. Statist. Assoc.* **53** 457–481.
KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887–906.
MALLOWS, C. (1974). On some topics in robustness. Memorandum, Bell Laboratories, Murray Hill, New Jersey.
MILLER, R. G. (1974). The jackknife—a review. *Biometrika* **61** 1–17.
MILLER, R. G. and HALPERN, J. (1982). Regression with censored data. *Biometrika* **69** 521–531.
RAO, C. R. (1973). *Linear Statistical Inference and Its Applications.* Wiley, New York.
SCHENKER, N. (1985). Qualms about bootstrap confidence intervals. *J. Amer. Statist. Assoc.* **80** 360–361.
SINGH, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9** 1187–1195.
THERNEAU, T. (1983). Variance reduction techniques for the bootstrap. Ph.D. thesis, Stanford University, Department of Statistics.
TIBSHIRANI, R. J. and HASTIE, T. J. (1984). Local likelihood estimation. Tech. Rep. Stanford Univ. Dept. Statist. **97**.
TUKEY, J. (1958). Bias and confidence in not quite large samples, abstract. *Ann. Math. Statist.* **29** 614.

Comment

J. A. Hartigan

Efron and Tibshirani are to be congratulated on a wide-ranging persuasive survey of the many uses of the boostrap technology. They are a bit cagey on what is or is not a bootstrap, but the description at the end of Section 4 seems to cover all the cases; some data y comes from an unknown probability distribution F; it is desired to estimate the distribution of some function $R(y, F)$ given F; and this is done by estimating the distribution of $R(y^*, \hat{F})$ given $\hat{F}$ where $\hat{F}$ is an estimate of F based on y, and y^* is sampled from the known $\hat{F}$.

There will be three problems in any application of the bootstrap: (1) how to choose the estimate $\hat{F}$? (2) how much sampling of y^* from $\hat{F}$? and (3) how close is the distribution of $R(y^*, \hat{F})$ given $\hat{F}$ to $R(y, F)$ given F?

Efron and Tibshirani suggest a variety of estimates $\hat{F}$ for simple random sampling, regression, and autoregression; their remarks about (3) are confined mainly to empirical demonstrations of the bootstrap in specific situations.

I have some general reservations about the bootstrap based on my experiences with subsampling techniques (Hartigan, 1969, 1975). Let $X_1, \ldots, X_n$ be a random sample from a distribution F, let F_n be the empirical distribution, and suppose that $t(F_n)$ is an estimate of some population parameter $t(F)$. The statistic $t(\hat{F}_n)$ is computed for several random subsamples (each observation appearing in the subsample with probability ½), and the set of $t(\hat{F}_n)$ values obtained is regarded as a sample from the posterior distribution of $t(F)$. For example, the standard deviation of the $t(\hat{F}_n)$ is an estimate of the standard error of $t(F_n)$ from $t(F)$; however, the procedure is not restricted to real valued t.

The procedure seems to work not too badly in getting at the first- and second-order behaviors of $t(F_n)$ when $t(F_n)$ is near normal, but it not effective in handling third-order behavior, bias, and skewness. Thus there is not much point in taking huge samples $t(\hat{F}_n)$ since the third-order behavior is not relevant; and if the procedure works only for $t(F_n)$ near normal, there are less fancy procedures for estimating standard error such as dividing the sample up into 10 subsamples of equal size and computing their standard deviation. (True, this introduces more bias than having random subsamples each containing about half the observations.) Indeed, even if $t(F_n)$ is not normal, we can obtain exact confidence intervals for the median of $t(F_{n/10})$ using the 10 subsamples. Even five subsamples will give a respectable idea of the standard error.

Transferring back to the bootstrap: (A) is the boot-

J. A. Hartigan is Eugene Higgins Professor of Statistics, Yale University, Box 2179 Yale Station, New Haven, CT 06520.

strap effective for non-normal situations? (B) in the normal case, does the bootstrap give accurate assessment of third-order terms? If not, it is scarcely justified to do many bootstrap simulations, since you will only use them to estimate a variance. The asymptotic justifications of the bootstrap such as in Bickel and Freeman (1981) or Singh (1981) do consider behavior near the normal.

To be specific, consider the case where a statistic $t(F_n)$ estimates a parameter $t(F)$. The first kind of bootstrapping might be on the quantity $t(F_n) - t(F)$; to estimate its variance $\sigma^2(F)/n$ we compute repeatedly $t(\hat{F}_n) - t(F_n)$ where $\hat{F}_n$ is the empirical distribution of a sample of size n from F_n. Thus $\sigma^2(F_n)$ will be used to estimate $\sigma^2(F)$. We might hope that

$$t(F_n) = t(F) + \xi \frac{\sigma(F)}{\sqrt{n}} + O\left(\frac{1}{n}\right)$$

where $\xi \sim N(0, 1)$. This is the case referred to above where $t(F_n)$ is normal and numerous resampling estimates are available to estimate $\sigma^2(F)$. To do better, consider the higher order terms:

$$t(F_n) = t(F) + \xi \frac{\sigma(F)}{\sqrt{n}} + \frac{s_3(F)}{n}(\xi^2 - 1) + \frac{b(F)}{n} + O(n^{-3/2}).$$

Then

$$t(\hat{F}_n) = t(F_n) + \xi \frac{\sigma(F_n)}{\sqrt{n}} + \frac{s_3(F_n)}{n}(\xi^2 - 1) + \frac{b(F_n)}{n} + O(n^{-3/2}).$$

We might expect that the sample quantities $\sigma(F_n)$, $s_3(F_n)$; $b(F_n)$ are within $O(n^{-1/2})$ of the population quantities; but since $\sigma(F_n) - \sigma(F) = O(n^{-1/2})$, the error in approximating the distribution of $t(F_n) - t(F)$ by that of $t(\hat{F}_n) - t(F_n)$ is $O(n^{-1/2})$, so that the additional skewness and bias terms are of no interest:

$$P\left[(t(F_n) - t(F) \le \frac{a}{\sqrt{n}}\right] - P\left[t(\hat{F}_n) - t(F_n) \le \frac{a}{\sqrt{n}}\right] = O(n^{-1/2}).$$

The bootstrap distribution is no better than any normal approximation using an estimate of variance accurate to $O(n^{-1/2})$!

On the other hand, if

$$R(y, F) = [t(F_n) - t(F)]/\sigma(F),$$

$$[t(F_n) - t(F)]/\sigma(F) = \frac{\xi}{\sqrt{n}} + \frac{s_3'(F)}{n}(\xi^2 - 1) + \frac{b'(F)}{n} + O(n^{-3/2})$$

$$[t(\hat{F}_n) - t(F_n)]/\sigma(F_n) = \frac{\xi}{\sqrt{n}} + \frac{s_3'(F_n)}{n}(\xi^2 - 1) + \frac{b'(F_n)}{n} + O(n^{-3/2}).$$

Now $s_3'(F_n)$ estimates $s_3'(F)$ and $b'(F_n)$ estimates $b'(F)$ to within $O(n^{-1/2})$, and the Cornish-Fisher expansion is accurate to skewness and bias terms:

$$P\left(\frac{t(F_n) - t(F)}{\sigma(F)} \le \frac{a}{\sqrt{n}}\right) - P\left(\frac{t(\hat{F}_n) - t(F_n)}{\sigma(F_n)} \le \frac{a}{\sqrt{n}}\right) = O(n^{-1}).$$

These results are given for $t(F_n) = \bar{X}$ in Singh (1981).

The conclusion is that for $t(F_n)$ near normal there is no advantage for the bootstrap over other resampling methods, unless the pivotal $[t(F_n) - t(F)]/\sigma(F)$ is used. Usually $\sigma(F)$ is not known; that's why we are resampling in the first place. We would need to estimate it by bootstrapping and use the pivotal $(t(F_n) - t(F))/\sigma(F_n)$. And the distribution of this pivotal would be determined by bootstrapping to obtain $[t(\hat{F}_n) - t(F_n)]/\sigma(\hat{F}_n)$. Note that $\sigma(\hat{F}_n)$ requires two levels of bootstrapping; this might get close to Professor Efron's objective of soaking up all the spare cycles on the West Coast!

Let us consider the modest objective of estimating the variance of $t(F_n)$. The various resampling techniques compute the variance of $t(W^1)$, $t(W^2)$, ..., $t(W^k)$ where $t(W^l)$ denotes the statistic computed on X_i repeated W_i^l times, $1 \le i \le n$. What is a good choice of $W^1, W^2, \ldots, W^k$? If in fact $X_1, \ldots, X_n$ are sampled from $N(\mu, \sigma^2)$ and $t = \bar{X}$, a minimum variance unbiased estimate of σ^2 is obtained by setting $W_i^l = 1 + \sqrt{n}\xi_i^l$ where $\xi^1, \sigma^2, \ldots, \xi^k$ are any k orthonormal vectors orthogonal to 1. The quantities $\sqrt{n}\xi_i^l$ can be obtained roughly by sampling each of them independently from $N(0, 1)$. Bootstrap resampling, for large n, has W_i^l approximately independently Poisson with expectation 1. Random subsampling, for large n has W_i^l approximately independent and approximately taking values 0 and 2 with probability ½. The Dirichlet distribution for F given F_n produces weights W_i^l that are approximately exponential with expectation 1. Any resampling scheme in which the weights are approximately independent with mean and variance 1 will give the right expected variance, but the efficiency

of the estimate (at normal means) is optimal for $W_i^l = 1 + \sqrt{n}\xi_i^l$.

For $n = 8$, obtain an efficient estimate from subsamples (1234), (1256), (1278), (1357), (1368), (1458), (1467); use as many as you need, and if $n > 8$ divide the sample into 8 groups as evenly as possible. I think it must be rare that the various approximations needed to connect the resampled computation to the computation of interest will be satisfied well enough to justify more than a few resamples. Perhaps this method might be called the *shoestring*.

ADDITIONAL REFERENCES

HARTIGAN, J. A. (1969). Using subsample values as typical values. *J. Amer. Statist. Assoc.* **64** 1303–1317.

HARTIGAN, J. A. (1975). Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values. *Ann. Statist.* **3** 573–580.

Rejoinder

B. Efron and R. Tibshirani

Professor Hartigan, who is one of the pioneers of resampling theory, raises the question of higher order accuracy. This question has bothered resamplers since the early days of the jackknife. Sections 7 and 8 of our paper show that the bootstrap can indeed achieve higher levels of accuracy, going the next step beyond simple estimates of standard error. The bootstrap confidence intervals we discuss are *not* of the crude (although useful) first-order form $\hat{\theta} \pm \hat{\sigma}z^{(\alpha)}$. They explicitly incorporate the higher order corrections about which Hartigan is legitimately concerned.

In particular the "z_0" term (7.8) is a correction for bias, and the acceleration constant "a," (7.16), is a correction for skewness. These correspond to Hartigan's $b(F)$ and $s_3(F)$, respectively. The reader who follows through Tables 5 and 7 will see these corrections in action. The fact that they produce highly accurate confidence intervals is no accident. The theory in Efron (1984a, 1984b) demonstrates higher order accuracy of the BC_a intervals in a wide class of situations. This demonstration does not yet apply to fully general problems, but current research indicates that it soon will. (The impressive higher order asymptotic results of Beran, Singh, Bickel, and Freedman, referred to in the paper, underpin these conclusions.)

It is worth mentioning that the bias and skewness corrections of the bootstrap confidence intervals are not of the simple "plug into an approximate pivotal" form suggested in Hartigan's remarks. The theory is phrased in a way which automatically corrects for arbitrary nonlinear transformations, even of the violent sort encountered in the correlation example of Table 5. In this sense the bootstrap theory does handle "non-normal situations."

Since this paper was written, research by several workers, including T. Hesterberg, R. Tibshirani, and T. DiCiccio, has substantially improved the computational outlook for bootstrap confidence intervals. It now appears possible that bootstrap sample sizes closer to $B = 100$ than $B = 1000$ may be sufficient for the task. However, these improvements are still in the process of development.

Professor Hartigan's last remarks, on the comparative efficiency of different resampling methods, need careful interpretation. There are two concepts of efficiency involved: the efficiency of the numerical algorithm in producing an estimate of variance, and the statistical efficiency of the estimate produced. There is no question that other resampling techniques, for example, the jackknife, can produce variance estimates more economically than does the bootstrap. We have argued, both by example and theory, that the bootstrap variance is generally more efficient as a statistical estimator of the unknown true variance.

This is not surprising given that methods like the jackknife are Taylor series approximations to the bootstrap (see Section 10). The simple idea in (2.3), substituting $\hat{F}$ for F, lies at the heart of all nonparametric estimates of accuracy. The bootstrap is the crudest of these methods in that it computes $\sigma(\hat{F})$ directly by Monte Carlo. For this reason it is also the method that involves the least amount of analytic approximation. It is perhaps surprising, and certainly gratifying, that a method based on such a simple form of inference is capable of producing quite accurate confidence intervals.

To say that the bootstrap is good, as we have been blatantly doing, doesn't imply that other methods are bad. Professor Hartigan's own work shows that for some problems, for example, forming a confidence interval for the center of a symmetric distribution, other methods are better. We hope that resampling methods in general will continue to be a lively research topic.