Statistical Science
1994, Vol. 9, No, 3, 382–415

Bootstrap: More than a Stab in the Dark?

G. Alastair Young

Abstract. A critical review is given of recent research activity on bootstrap and related procedures. Theoretical work has shown the bootstrap approach to be a potentially powerful addition to the statistician's toolkit. We consider its impact on statistical practice and argue that, measured against the hopes raised by theoretical advances, this has been until now fairly modest. We suggest that while this state of affairs is a consequence to be expected of the sophisticated character of the bootstrap procedures required to cope reliably in many of the settings of most interest, much theoretical work is not serving the immediate needs of statistical practice. Emerging lines of research are reviewed and important future research directions suggested. In particular, we appeal for greater focussing of research activity on practicalities.

Key words and phrases: Bootstrap, independent data, dependent data, practical implementation, theoretical development, higher-order accuracy, asymptotics, validity, resampling scheme, design parameter, Edgeworth expansion, prepivoting, iteration, jackknife, nonparametric likelihood.

1. INTRODUCTION

Among the most significant developments in statistical methodology of the 1980's were those relating to the study of bootstrap resampling procedures. Publication of Efron's seminal paper (Efron, 1979a) drew attention to the potential scope of the bootstrap as a procedure by which many statistical problems could be tackled. Since then, a great deal of work has been carried out which demonstrates the extent to which the very simple procedure of resampling from a given set of data can approximate sampling variations which produced that data in the first place. Has this work revealed a statistical success story? Was early enthusiasm (Efron, 1979b; Diaconis and Efron, 1983) for the bootstrap justified? How useful has the bootstrap proved to be? Why has the full scope of the idea been so slow to catch on among practitioners?

Early work, of the kind detailed in Efron (1982), stayed close to the notion of bootstrap as a simple extension of other resampling methods of statistical error estimation, such as the jackknife and delta methods. More recently, bootstrap has come to be viewed as a more general purpose tool which can be used to tackle a wide range of statistical problems. When we judge bootstrap against the heightened expectations suggested by recent focus of work in this area, is the bootstrap revealed to be more than just a stab in the dark? To what extent can the user of the bootstrap be sure of it hitting its target? These are key concerns for the practitioner.

Publication of *Exploring the Limits of Bootstrap*, edited by LePage and Billard (1992), offers an opportunity for reflection on what recent work on the bootstrap has achieved, and in particular on its lessons for statistical practice. With publication also of monographs by Beran and Ducharme (1991), Hall (1992a) and Mammen (1992), the time is perhaps ripe to consider too whether the great volume of theoretical work devoted to bootstrap has, until now, addressed the correct issues from the point of view of statistical practice.

Readily accessible accounts of bootstrap which stress applications are given in the review article by Efron and Tibshirani (1986) and in the recent monograph (Efron and Tibshirani, 1993) by the same authors. As well as giving a detailed introduction to the bootstrap, the latter monograph gives a brief exposition of many recent developments of key importance.

The intention of the current article is to give a critical account of perspectives and progress on bootstrap methodology that have emerged in the last few years. We will discuss reasons why, though a theoretical success, the bootstrap may be judged to have been a less spectacular practical success in recent years than many might have expected or than should be

G. Alastair Young is Lecturer, Statistical Laboratory, University of Cambridge, 16 Mill Lane, Cambridge CB2 1SB, U.K.

possible. We will suggest, too, that the bulk of theoretical work has not served the immediate needs of statistical practice.

2. THE BOOTSTRAP PARADIGM

The term "bootstrap" which Efron gave to his procedure is emotive, perhaps unfortunate, and it may in part be responsible for the suspicion with which it is sometimes viewed. As noted by Hall (1992a, Section 1.1), the name has tended to convey the notion of the statistician striving to obtain something for nothing, of hopeless endeavour. The thinking should rather be that of a procedure by which sound statistical conclusions *can* often be reached, but which, being entirely conditioned by the sample data, provides no direct guarantee of doing so.

The bootstrap paradigm is easily stated, though often less easily implemented effectively. A given set of data x is assumed to come from some unknown underlying distribution F. Letting X denote the underlying random variable, interest lies in some real-valued "root" $T(X, F)$, depending on X and possibly F. The inference being performed requires knowledge of the sampling distribution of $T(X, F)$ under F. The bootstrap estimates the probability mechanism F by some mechanism $\widehat{F}$ based on x, and hence the required sampling distribution by that of $T(Y, \widehat{F})$, when Y is generated from $\widehat{F}$. This latter sampling distribution is usually constructed via Monte Carlo simulation, by repeated sampling from $\widehat{F}$. It is in linking a simple estimation idea with computing power that the great scope of bootstrap lies.

The bootstrap may be applied parametrically or nonparametrically. In the former case, some parametric form for F is assumed, unknown parameters are estimated from x and sampling takes place from the distribution so fitted. The bulk of published work, however, concerns the nonparametric case, for this is where the fun lies and the most immediate practical gains, in terms of quick error estimates and p-values, might be expected. Here, in the simplest setting, x is assumed to be an independent, identically distributed sample [$x = (x_1, x_2, \ldots, x_n)$, say], $\widehat{F}$ is the empirical distribution function of the observed sample, which puts equal mass $1/n$ at each of the x_i, and sampling from $\widehat{F}$ amounts to resampling, with replacement, from $(x_1, x_2, \ldots, x_n)$.

To illustrate further, we consider now three related examples which will be discussed further later in the paper. In each example, suppose the given data $(x_1, x_2, \ldots, x_n)$ represent an independent sample from an underlying univariate distribution F.

EXAMPLE 1. Denote by μ the mean of F, and let $X_1, \ldots, X_n$ denote independent random variables with distribution F. Let $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$. Consider estimation of the distribution function $H(z)$ of $S = n^{1/2}(\overline{X}_n - \mu)$:

$$H(z) = P\{n^{1/2}(\overline{X}_n - \mu) \le z\}.$$

Let $\overline{x}_n = n^{-1}\sum_{i=1}^n x_i$ be the observed sample mean, and let $\{Y_1, \ldots, Y_n\}$ denote a random sample of size n drawn, with replacement, from $\{x_1, \ldots, x_n\}$. Then the bootstrap estimate $\widehat{H}(z)$ of $H(z)$ is

$$\widehat{H}(z) = P\{n^{1/2}(\overline{Y}_n - \overline{x}_n) \le z\},$$

with $\overline{Y}_n = n^{-1}\sum_{i=1}^n Y_i$ and the probability computed under the repeated drawing of "bootstrap samples" $\{Y_1, \ldots, Y_n\}$.

EXAMPLE 2. Generalizing Example 1, let $\theta = \theta(F)$ denote an unknown parameter of interest, with estimate $\widehat{\theta}$ taking the observed value $\widehat{\theta}(x) \equiv \widehat{\theta}(x_1, \ldots, x_n)$. Let

$$T = \frac{n^{1/2}\{\widehat{\theta}(X_1, \ldots, X_n) - \theta\}}{\widehat{\sigma}},$$

where $\widehat{\sigma}^2 \equiv \widehat{\sigma}^2(X_1, \ldots, X_n)$ is an estimate of the asymptotic variance σ^2 of $n^{1/2}\widehat{\theta}$. Consider estimation of the distribution function $G(z)$ of T, $G(z) = P(T \le z)$.

The bootstrap estimate of G is

$$\widehat{G}(z) = P\left[\frac{n^{1/2}\{\widehat{\theta}(Y_1, \ldots, Y_n) - \widehat{\theta}(x)\}}{\widehat{\sigma}(Y_1, \ldots, Y_n)} \le z\right],$$

where, as in Example 1, $\{Y_1, \ldots, Y_n\}$ denotes a bootstrap sample drawn from $\{x_1, \ldots, x_n\}$.

Suppose that $G(v_\alpha) = \alpha$, and $\widehat{G}(\widehat{v}_\alpha) = \alpha$. Then an exact two-sided, equal-tailed confidence interval for θ is, using the notation of Hall (1992a, Section 3.2),

$$J_2 = \left(\widehat{\theta}(x) - n^{-1/2}\widehat{\sigma}(x)v_{(1+\alpha)/2}, \widehat{\theta}(x) - n^{-1/2}\widehat{\sigma}(x)v_{(1-\alpha)/2}\right).$$

The bootstrap version of this interval, the "percentile-t bootstrap confidence interval," is

$$\widehat{J}_2 = \left(\widehat{\theta}(x) - n^{-1/2}\widehat{\sigma}(x)\widehat{v}_{(1+\alpha)/2}, \widehat{\theta}(x) - n^{-1/2}\widehat{\sigma}(x)\widehat{v}_{(1-\alpha)/2}\right).$$

EXAMPLE 3. Consider use of the root

$$U = n^{1/2}\{\widehat{\theta}(X_1, \ldots, X_n) - \theta\}$$

instead of T in Example 2. The bootstrap estimates the distribution function $G^\dagger(z)$ of U, $G^\dagger(z) = P(U \le z)$, by

$$\widehat{G}^\dagger(z) = P\left[n^{1/2}\{\widehat{\theta}(Y_1, \ldots, Y_n) - \widehat{\theta}(x)\} \le z\right].$$

Now an exact, two-sided, equal-tailed $1-\alpha$ confidence interval for θ is

$$I_2 = \left(\widehat{\theta}(x) - n^{-1/2}\sigma u_{(1+\alpha)/2}, \widehat{\theta}(x) - n^{-1/2}\sigma u_{(1-\alpha)/2}\right),$$

where

$$P\left[\frac{n^{1/2}\{\widehat{\theta}(X_1, \ldots, X_n) - \theta\}}{\sigma} \leq u_\alpha\right] = \alpha.$$

The bootstrap version of this interval, a "percentile bootstrap confidence interval," is

$$\widehat{I}_2 = \left(\widehat{\theta}(x) - n^{-1/2}\widehat{\sigma}(x)\widehat{u}_{(1+\alpha)/2}, \widehat{\theta}(x) - n^{-1/2}\widehat{\sigma}(x)\widehat{u}_{(1-\alpha)/2}\right),$$

where

$$P\left[\frac{n^{1/2}\{\widehat{\theta}(Y_1, \ldots, Y_n) - \widehat{\theta}(x)\}}{\widehat{\sigma}(x)} \leq \widehat{u}_\alpha\right] = \alpha.$$

Note that, in each example, exact analytic evaluation of the bootstrap quantities will rarely be possible. In principle, however, and here lies much of the strength of the bootstrap approach, these quantities may be computed, to any desired level of accuracy, by a Monte Carlo simulation. Beran and Ducharme (1991, pages 7–14) give a detailed account of procedures for computation of a bootstrap distribution.

We must now, however, immediately acknowledge that there is a stumbling block with the bootstrap which certainly makes it difficult to adapt to many statistical problems of interest. It is only within the narrow independent, identically distributed sample setting, and here indeed only to a partial extent, that the prescription of bootstrap is so simple and therefore that the procedure has any hope of becoming automatic. With any other data structure, the bootstrap still demands user choice of the model $\widehat{F}$ to be sampled from. A simple example concerns the heteroscedastic linear model, where it is not at all obvious how the fitted model $\widehat{F}$ should be chosen [see Wu (1986) and the associated discussion].

3. THE CURRENT STATE OF PLAY

Work on bootstrap has had its theoretical high points. Historically, fundamental work (e.g., Bickel and Freedman, 1981) establishing asymptotic consistency of the bootstrap in wide generality was followed by a stream of research establishing its higher-order accuracy properties [see, e.g., Singh (1981)]. Subsequent work [see, e.g., Efron (1987) or Beran (1987)] was largely aimed at defining refined bootstrap methods and at elucidating the comparative asymptotic properties of competing procedures. Particular focus here has been on the relatively straightforward problem of interval estimation for scalar parameters (see Hall, 1988). The *Exploring the Limits of Bootstrap* volume edited by LePage and Billard (1992) suggests that once more the focus is shifting. Currently much attention is devoted to extension of the bootstrap paradigm to dependent data problems, such as time series analysis, to establishment of consistency in general settings, such as robust inference, and to use of the bootstrap within other statistical procedures, such as bandwidth selection in nonparametric smoothing problems.

Much of the current appeal of bootstrap, without doubt, stems from the not unrealistic hope of obtaining—as much of the research effort has been geared to show—higher-order accuracy in an automatic and simple manner. By use of the bootstrap we may, in particular, gain, in considerable generality, improvements over procedures based on the normal approximation, without those improvements being at the expense of sophisticated, problem-specific theoretical analysis. The only price that needs to be paid is one of increased computation.

The bulk of work carried out on bootstrap has been devoted to these higher-order asymptotic properties. While the fruits of this labour have provided strong justification for pursuing the bootstrap idea to its limits, doubts must remain about the gulf between theory and practice. Theoretically, especially with refinement, bootstrap works! It is less clear from the applied statistical literature that practitioners are impressed enough by theoretical discoveries to grasp the bootstrap approach as fully as they might.

Although much has been achieved, we must question whether the direction theoretical work has taken has led, for the most part, to the wrong issues being addressed. Researchers have succumbed too much, perhaps, to the temptation to devote their efforts to squeezing even better performance from the bootstrap, through refinement of procedures which already provide valid inference, rather than focussing their efforts on more fundamental issues concerning basic reliability of the approach. Work which is leading, as Efron and LePage (1992) note in their introduction to LePage and Billard (1992), to practical solution of the problem of producing highly accurate confidence intervals is impressive, but global answers to more basic questions seem further off. When does bootstrap work? When does bootstrap fail to provide valid inference? Are cases of failure pathological, or practically significant? The answers here should reflect an important point.

As popularly perceived, the bootstrap has two main attributes. It provides an automatic approach to inference, by utilizing the computer, and it provides a handle on inference in circumstances where standard approaches which invoke strong assump-

tions are judged inappropriate. The danger is that the practitioner may well be attracted to use of the bootstrap for the second of these attributes, in other words in precisely those circumstances where the bootstrap is most likely to fail and where least is known. Procedures which allow empirical identification of bootstrap breakdown are much needed.

Undoubtedly, bootstrap has been successfully demonstrated to be a sensible approach for simple error assessment and confidence interval construction in many statistical problems. Published applications of the bootstrap are now numerous and include problems in reliability studies, ecology, high-energy physics, genetics, record linkage studies, meteorology and biostatistics. A number of applications papers are contained in LePage and Billard (1992), and a number of examples of data analysis via bootstrap are given by Efron and Tibshirani (1993). However, most published applications of bootstrap are limited to its use for construction of simple measures of statistical accuracy, of the kind which were the focus of early work on the bootstrap and highlighted by Efron and Tibshirani (1986). Evidence that more sophisticated procedures, such as the iterated bootstrap and bootstrap resampling schemes for dependent data, which have been the predominant focus among bootstrap researchers for some time, have been so readily adopted is rather scarce. The reasons for reticence are clear.

First, theoretical work has shown that statistical inference performed by bootstrap may depend crucially on the form of estimator used. Since the early theoretical work of Bickel and Freedman (1981) and Singh (1981), it has been recognised that the bootstrap, in its simplest and most easily applied form, as described in Section 2 above, will not always yield an asymptotically valid inference. Naive use of an invalid resampling scheme can produce trouble, even to the extent of inconsistency.

A well-documented example concerns the bootstrap distribution of the maximum of a sample. Let $X = (X_1, \ldots, X_n)$, with $X_1, \ldots, X_n$ independent, identically distributed random variables with distribution supported on the interval $(0, \theta)$. Consider the root $T(X, F) = n(\theta - X_{(n)})/\theta$, where $X_{(n)} = \max\{X_1, \ldots, X_n\}$. Let, as before, $\widehat{F}$ denote the empirical distribution function of the observed sample $x = (x_1, \ldots, x_n)$. Let $x_{(n)} = \max\{x_1, \ldots, x_n\}$. Then the bootstrap version of the root is given by $T(Y, \widehat{F}) = n(x_{(n)} - Y_{(n)})/x_{(n)}$, where $Y_{(1)} \leq \cdots \leq Y_{(n)}$ are the ordered members of a bootstrap sample $Y = (Y_1, \ldots, Y_n)$ drawn independently from $\widehat{F}$. Then it is easily seen that $P\{T(Y, \widehat{F}) = 0\} \to 1 - \exp(-1)$ as $n \to \infty$. However, if F is the uniform distribution $U(0, \theta)$, the asymptotic distribution of $T(X, F)$ is exponential of mean 1, from which it is seen that the bootstrap distribution does not have a correct asymptotic limit [for further details see Bickel and Freedman (1981)].

Singh (1981) gives a further example which warns against naive use of the bootstrap. His example involves use of the resampling scheme appropriate to independent and identically distributed data when the data are actually weakly dependent observations; we will return to this example in Section 5.

As a third example, consider the bootstrap distribution of the mean in the infinite-variance case. Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables of infinite variance, but with X_1 belonging to the domain of attraction of a stable law. In these circumstances, it is known that the appropriately normalized mean $\overline{X}_n$ converges to a stable law, but that the bootstrap version of the normalized mean has a limit which is a random distribution, implying that the simple bootstrap scheme can fail for heavy-tailed distributions. See Athreya (1987) for details.

A second inhibiting factor relates to choice of bootstrap procedure. Research has shown, not unexpectedly, that there is no specific implementation of the bootstrap paradigm which is universally superior to others. Even if the simplest bootstrap sampling scheme, which resamples from the empirical distribution function of the given data, is asymptotically valid, it may not be the only feasible approach, or the best. As well as the choice of root T to be bootstrapped being crucial, as we shall see below, other choices of $\widehat{F}$ may be worthy of consideration. For example, Silverman and Young (1987) consider use of the bootstrap in simple functional estimation problems, such as those involving estimation of moments. They compare the simple bootstrap scheme which resamples from the empirical distribution function with a "smoothed" bootstrap, which constructs the distribution $\widehat{F}$ to be sampled from by smoothing the empirical distribution function using a kernel method. Whether the smoothed procedure is preferable to the standard scheme is shown to depend crucially on both the underlying distribution F and on the quantity being estimated.

A third problem relates to the relevance, or otherwise, of asymptotics. Attractiveness of the bootstrap approach is somewhat dimmed by the realization that favourable asymptotics of a particular bootstrap estimation procedure is no guarantee of good small-sample behaviour. Schenker (1985) illustrates the poor small-sample performance of procedures, which have asymptotic justification, when constructing confidence intervals for a population variance. As further forceful illustration of this point, consider the percentile-t method, as described in Section 2. It is well known that this method has desirable theoretical properties as a means of

constructing nonparametric confidence intervals; we will review some of these properties in Section 4. In small samples, however, it may lead to unreliable intervals, due to the difficulty in many problems of constructing a stable variance estimate. For instance, Hall, Martin and Schucany (1989) consider construction of bootstrap confidence intervals for a correlation coefficient, by the kind of procedure illustrated in Example 2. In that problem, it is not possible to find an estimator $\widehat{\sigma}^2$ of the asymptotic variance σ^2 which itself has low variance. The result is confidence intervals for the correlation coefficient θ which, when constructed from small samples by the percentile-t method, are often erratic, with endpoints escaping the interval $[-1, 1]$.

Difficulties of the above kind must be set against a background where, for the most part, bootstrap only provides a valid inference where other, simpler, first-order alternatives such as the normal approximation are available.

Finally, and in this author's view crucially, 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. A bootstrap approach with time series data is to resample "blocks" of successive observations. As we shall see in Section 5, a choice regarding block size must be made. Considerable practical problems remain.

Recent volumes (Le Page and Billard, 1992; Hall, 1992a) reflect well the extent to which sophisticated theory has been put into the service of analysis of the bootstrap. The bootstrap has provided statisticians with a valuable tool with which to supplement standard statistical analyses. However, theory has not generally yet developed to the extent of providing a practically dependable higher-order accurate methodology as general in scope as the range of contexts studied would appear to herald.

In the following sections we will examine in more detail the main strands in current development of bootstrap methodology and amplify the above points.

4. INDEPENDENT DATA

4.1 What Is Known

The bulk of published work on the properties of bootstrap concerns the case of independent—and usually identically distributed—data, for the simple reason that if bootstrap is to be a success, it must work here. Does it? The answer is arguably all that might be expected: a partial "yes."

Efron's early development of the bootstrap, as summarized in Efron (1982) and Efron and Tibshirani (1986, 1993), remains faithful to the idea of the bootstrap as a development of other resampling methods, particularly the jackknife. Viewed in these terms, there can be little doubt of the success of bootstrap. The bootstrap may work, for example, in the problem of estimating the asymptotic variance of a sample quantile, where jackknife does not. Efron (1982) and Efron and Tibshirani (1986) give a number of examples where bootstrap considerably outperforms preexisting methods of error assessment, such as the jackknife and delta methods. It is this author's contention, however, that early theoretical and practical findings led to bootstrap being viewed in much more ambitious terms than as merely extensions of these methods. It is when the more ambitious view is taken that the cracks begin to appear.

A detailed general account specifically addressed to questions, both in terms of underlying distribution and resampling scheme, of when bootstrap works is given by Mammen (1992). Consideration is given to application of bootstrap in three contexts: estimation of smooth functionals; nonparametric curve estimation; and linear models. Numerous other authors have considered particular problems.

Overall, the bootstrap has been shown to be a potentially effective approach to many standard problems of statistical inference, especially error estimation and confidence interval construction, not only in the independent, identically distributed sample setting, but also the practically important regression setting. A brief summary of general conditions that together imply consistency of the bootstrap is given by Efron and LePage (1992). These conditions were first established by Bickel and Freedman (1981).

In the notation of Section 2, the simple bootstrap resampling scheme for independent data, which resamples from the empirical distribution function, will work provided that the following hold:

(i) $T\{(X_1, \ldots, X_n), G\}$ tends weakly to a limit law L_G whenever $X_1, \ldots, X_n$ are independent, identically distributed with distribution G, for all G in some neighbourhood of the true distribution F into which the empirical distribution function $\widehat{F}$ eventually falls with probability 1;

(ii) the convergence in (i) is uniform on this neighbourhood;

(iii) the function $G \to L_G$ is continuous.

Establishing validity of bootstrap in a particular setting may be a highly nontrivial exercise. Of these conditions, that involving uniformity of convergence as a function of the unknown data model is often the most taxing to verify. Bickel and Freedman (1981) provide a number of examples, including the example of bootstrapping the distribution of the maximum of a sample, where the bootstrap fails because uni-

formity of convergence does not hold on any suitable neighbourhood of the true distribution F. Efron and LePage (1992) note that the bootstrap can fail due to violation of each of the three conditions individually. We have already indicated in Section 3 a number of cases of bootstrap failure. Further examples are given by, among others, Bickel and Freedman (1981) and Mammen (1992).

Theoretically important work [see, e.g., Bretagnolle (1983)] has shown how validity of the bootstrap may be ensured in many settings by simple modifications such as introduction of smoothing and alteration of the resample size. Consider the independent, identically distributed case. The latter idea involves ensuring that the above conditions hold by estimating the distribution of $T(X, F)$ by that of $T(Y, \widehat{F})$, where $Y = (Y_1, \ldots, Y_m)$ denotes a bootstrap sample drawn from the observed data $\{x_1, \ldots, x_n\}$, but of size m, for suitable m to be specified, rather than size n as would be usual. How is m to be chosen in any given context? It is easy to question the practical value of such modified bootstrap procedures.

Many of the remarks made in Section 3 are somewhat disparaging of the tone of much of bootstrap research. However, it would be wrong to be too critical of the way bootstrap research has developed. Paradoxically, the flavour of much theoretical work in the independent data setting has been influenced precisely by bootstrap's own success, what Efron and LePage (1992) describe as "an unexpected major advance." The bootstrap, properly formulated, is not only an automatic alternative to more classical procedures, such as normal approximations and procedures based on explicit Edgeworth correction, but is actually capable of producing greater accuracy automatically, without any special theoretical analysis being necessary, and will do so in considerable generality. Efron and LePage (1992) summarize very briefly some of the results concerning the higher-order accuracy properties of bootstrap. Roughly speaking, for estimating the distribution function of a studentized statistic, the bootstrap automatically produces answers as good as those obtained by a one-term Edgeworth expansion. For constructing one-sided confidence intervals, the percentile bootstrap method provides no improvement over use of a normal approximation, giving a coverage which differs from the nominal required coverage by an error of order $O(n^{-1/2})$, where n is the sample size. The percentile-t method, however, typically gives a coverage error of order $O(n^{-1})$ and in that sense displays higher-order accuracy. For two-sided confidence intervals, both the percentile and percentile-t methods yield the same order of coverage error, $O(n^{-1})$, as procedures based on the normal approximation, though as we shall discuss below more sophisticated versions of the bootstrap can reduce this error. Further details of the theoretical properties of bootstrap and references may be found in Hall (1992a, Chapter 3).

It is particularly worth noting that Singh's (1981) paper, which described the skewness-correcting role of bootstrap methods in the case of a standardized sample mean, was based on the development of a one-term Edgeworth expansion. This paper firmly established a framework, based on Edgeworth expansion methodology, for analysis of bootstrap which has been central to bootstrap research and from which we have learned much [see Hall (1988, 1992a)]. It may be argued that by its very nature this framework has encouraged a certain preoccupation with higher-order accuracy considerations, within particular classes of models. Balanced against this viewpoint, though, must be set the argument that without such higher-order comparisons a less satisfying picture, not of what bootstrap is capable of achieving, but of how this should be achieved, would be available.

Among the key issues which have been of concern to bootstrap researchers in recent years, and about which the Edgeworth expansion methodology in particular has had much to say, are questions relating to pivoting and the related notions of prepivoting, calibration and iteration. It is worthwhile summarizing here some of the current thinking with respect to these issues.

A special role is played in bootstrap methodology by pivotal methods. Recall that in Section 2 we presented two examples (Examples 2 and 3) concerning inference for a scalar parameter θ. The first procedure, Example 2, was based on an asymptotically pivotal statistic T, whose limiting distribution does not depend on unknown quantities. In Example 3 inference was based on the nonpivotal statistic U.

As noted by Hall (1992a, Section 3.1), the advantages of pivotal applications of the bootstrap are readily explained by means of Edgeworth expansion. If T is asymptotically normally distributed, then in regular cases [see Hall (1992a, Chapter 2)] we may expand its distribution function as

$$G(z) = \Phi(z) + n^{-1/2} q(z)\phi(z) + O(n^{-1}),$$

where q is an even quadratic polynomial and Φ and ϕ are the standard normal distribution and density functions, respectively. The bootstrap estimate $\widehat{G}(z)$ admits an analogous Edgeworth expansion

$$\widehat{G}(z) = \Phi(z) + n^{-1/2} \widehat{q}(z)\phi(z) + O_p(n^{-1}),$$

where $\widehat{q}$ is obtained from q on replacing unknown population quantities by bootstrap estimates. Since

such estimates are typically $O_p(n^{-1/2})$ distant from their population values in q, $\hat{q} - q = O_p(n^{-1/2})$, so that $\hat{G}(z) - G(z) = O_p(n^{-1})$. The bootstrap approximation to G is in error by a term of size n^{-1}, while a normal approximation, which approximates $G(z)$ by $\Phi(z)$, is in error by a term of size $n^{-1/2}$.

If, however, we use the bootstrap to approximate the distribution of a nonpivotal statistic U, as in Example 3 of Section 2, we commit an error of size $n^{-1/2}$. The relevant Edgeworth expansions are now

$$G^{\dagger}(z) = \Phi\left(\frac{z}{\sigma}\right) + n^{-1/2}p\left(\frac{z}{\sigma}\right)\phi\left(\frac{z}{\sigma}\right) + O(n^{-1})$$

and

$$\hat{G}^{\dagger}(z) = \Phi\left(\frac{z}{\hat{\sigma}}\right) + n^{-1/2}\hat{p}\left(\frac{z}{\hat{\sigma}}\right)\phi\left(\frac{z}{\hat{\sigma}}\right) + O_p(n^{-1}).$$

Since $\hat{p} - p = O_p(n^{-1/2})$ and $\Phi(z/\sigma) - \Phi(z/\hat{\sigma}) = O_p(n^{-1/2})$, we see that

$$\hat{G}^{\dagger}(z) - G^{\dagger}(z) = O_p(n^{-1/2}).$$

The bootstrap approximation is in error by terms of size $n^{-1/2}$, not n^{-1}, if we bootstrap the nonpivotal statistic U.

Other recent key theoretical developments concern the notion of prepivoting (Beran, 1987) or bootstrap iteration (Hall and Martin, 1988). We provide an illustration of this notion involving Example 1 of Section 2.

An ideal, one-sided confidence interval for μ of exact coverage $1-\alpha$ is

$$\begin{aligned} C &= \{\mu \mid S \le H^{-1}(1-\alpha)\} \\ &= \{\mu \mid H(S) \le 1-\alpha\}. \end{aligned}$$

The corresponding bootstrap confidence interval may be written

$$\begin{aligned} \hat{C} &= \{\mu \mid \hat{H}(S) \le 1-\alpha\} \\ &= \{\mu \mid S \le \hat{H}^{-1}(1-\alpha)\}. \end{aligned}$$

This confidence set may have coverage probability which differs substantially from the nominal value $1-\alpha$. The idea of prepivoting is that of transforming the root S to a new root S_1, whose distribution depends less strongly on F than that of S. Such a transformed root is $H(S)$, whose distribution is, assuming continuity, $U(0, 1)$. However, H is unknown, so the bootstrap approach suggests considering $S_1 = \hat{H}(S)$, which has distribution function

$$H_1(z) = P(S_1 \le z)$$

under sampling from F. Let $\hat{H}_1$ be the bootstrap estimator of H_1. Construction of the bootstrap confidence set from S_1 gives the prepivoted bootstrap confidence set

$$\begin{aligned} \hat{C}_1 &= \{\mu \mid S_1 \le \hat{H}_1^{-1}(1-\alpha)\} \\ &= \left\{\mu \mid S \le \hat{H}^{-1}\left(\hat{H}_1^{-1}(1-\alpha)\right)\right\}. \end{aligned} \qquad (*)$$

Construction of $\hat{C}_1$ entails an iterated bootstrap computation. Bootstrap samples must be drawn from the bootstrap samples drawn from the given sample data [for details see Beran and Ducharme (1991), pages 27–36]. It is clear from (*) that, in this context of constructing a confidence set, the prepivoting or iterated bootstrap idea amounts to making an additive correction to the nominal coverage level of the original bootstrap confidence interval $\hat{C}$.

Martin (1990) discusses use of iterated bootstrap procedures in constructing confidence intervals, as illustrated above, and shows that bootstrap coverage correction produces improvements in coverage accuracy of order $n^{-1/2}$ in one-sided intervals and of order n^{-1} in two-sided intervals.

In general terms, bootstrap iteration aims to enhance the accuracy of a bootstrap method by estimating some error term and adjusting the bootstrap method so as to reduce that error. Loh (1987) introduced the notion of bootstrap calibration as a device for improving the coverage accuracy of confidence intervals. If calibration is applied to a bootstrap confidence interval, it amounts precisely to bootstrap iteration. There now exists a widely held view that, at least from a theoretical viewpoint, a significant part of the future of bootstrap methodology lies in the use of the iterated bootstrap [see Hall (1992a, Section 3.11.1)]. The disadvantage is that iteration is highly computer intensive, perhaps unattractively so for the potential user.

4.2 What Is Needed

The bootstrap will usually provide a consistent statistical procedure in the independent data setting, if not directly, at least after some modification. We have noted that there are often practical problems, however, once consistency has been ensured. Patch-ups of the basic bootstrap involving devices such as modification of resampling size, while understood theoretically, suffer still from a lack of practicality.

The smoothed bootstrap is a further case in point. Smoothing of the empirical distribution function before resampling has been shown to be necessary to obtain asymptotically valid inference in some problems [see the example of bootstrapping the mode considered by Romano (1988)]. Also, while in most problems smoothing will not affect the rate of convergence of a bootstrap estimator, it can be substantially worthwhile in small-sample contexts [see

Silverman and Young (1987); De Angelis and Young (1992)]. Also, there are problems involving estimation of quantities which depend on local properties of the underlying distribution, such as problems involving confidence intervals for quantiles, where a slow rate of convergence of the simplest bootstrap estimator can be improved by smoothing. Examples are considered by Hall, DiCiccio and Romano (1989) and De Angelis, Hall and Young (1993a, b). The introduction of smoothing takes away the automatic character of bootstrap and leaves behind the practically important and nontrivial problem of specifying a smoothing bandwidth.

The point here is that work which we may perform to ensure that the bootstrap enjoys the second of its key attributes above, general applicability, may well deprive it of the first, simplicity. The practitioner may not mind, provided some more specific guidelines than those currently available are provided.

Some of the reticence that potential users of bootstrap may feel relates to absence of direct procedures for critical thinking about the bootstrap procedures they may be using. Study of empirical procedures which enable informed decision from the data of when bootstrapping is sensible is required. 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. Efron (1992) presents a jackknife-after-bootstrap procedure for this use [see also Efron and Tibshirani (1993, Chapter 19)]. An important research direction will certainly be investigation of the practical value and scope of such procedures. How generally might they be applied? Do they provide the basis for empirical determination of breakdown of the bootstrap? Will extension of these methods be a panacea for the problems of bootstrap which have emerged from the great volume of theoretical study to which it has been subjected, and some of which we have noted here? One can remain hopeful, but since the most natural and flexible procedures for internal error assessment are based on the bootstrap itself (Efron, 1992; De Angelis and Young, 1992), some circularity of argument seems inevitable.

The vast bulk of published work on bootstrap has been concerned with asymptotics, which may reassure, but provide little direct backing for use of the bootstrap in the small sample context. It may be argued that too little has been done to study small-sample properties. Young and Daniels (1990) provide a very limited small-sample study of bias of the bootstrap in simple estimation problems, and analytic procedures by which, if extended, it may be possible to analyze small-sample behaviour more generally. Owen (1991) provides small-sample comparison between various procedures used for construction of nonparametric confidence intervals for a population mean. While limited simulation studies of this kind will often suggest that asymptotic results are reasonably true in small samples, this is not always the case. Simulation results which illustrate this point very clearly are contained in Mammen's (1992, Chapter 1, Section 5) comparison between the bootstrap and normal approximations in the problem of estimating the sampling distribution of a sample mean.

5. DEPENDENT DATA

5.1 Recent Attention

As LePage and Billard (1992) clearly reflects, attention has recently shifted away from polishing of the bootstrap in the independent data setting toward extending the procedures to dependent observations, in particular, stationary and nonstationary time series data and Markov chains. This is an important shift of focus. A very readable account of some of the developments in this area is given by Léger, Politis and Romano (1992). They make a key point which is obscured somewhat by the theoretical tone of the papers on bootstrapping dependent data in LePage and Billard (1992). In this context, the distribution of an estimator generally depends heavily on the joint distribution of the observations, and, while distribution theory of estimators is often known, its use is often difficult. By contrast with the simpler independent data setting, therefore, the bootstrap has a more fundamental enabling role to play. In the independent data setting, bootstrap offers the possibility of automatic improvement over standard procedures, such as those based on normal approximation. In the dependent data setting, practical alternatives to bootstrap are more limited, and the bootstrap therefore more appealing. The rewards may be higher, but the problems are trickier and practical methodology is still some way off.

To illustrate some of the key points relevant to bootstrapping dependent data, consider again Example 1 of Section 2, but suppose now that the X_i are m-dependent. This example is considered by Singh (1981) and Léger, Politis and Romano (1992). Recall that the sequence of random variables $\{X_n,\ n = 0, \pm 1, \pm 2, \ldots\}$ is stationary if, for all $n = 0, \pm 1, \pm 2, \ldots$, the joint distribution of $(X_k, X_{k+1}, \ldots, X_{k+n})$ does not depend on k. A stationary sequence is m-dependent if the set of random variables $\{X_n,\ n = -1, -2, \ldots\}$ is independent of $\{X_n,\ n = m, m+1, \ldots\}$.

If $X_1, \ldots, X_n$ are from a univariate m-dependent stationary sequence, $E(\overline{X}_n) = \mu$, with $\mu = E(X_1)$,

and if $m \leq n$,

$$\sigma_n^2 = \operatorname{var}(n^{1/2}\overline{X}_n) = \operatorname{var}(X_1) + 2\sum_{i=1}^{m}\left(1 - \frac{i}{n}\right)\operatorname{cov}(X_1, X_{1+i}).$$

By the central limit theorem for m-dependent processes, $n^{1/2}(\overline{X}_n - \mu)$ converges in distribution to $N(0, \sigma_\infty^2)$, where

$$\sigma_\infty^2 = \lim_{n\to\infty} \sigma_n^2 = \operatorname{var}(X_1) + 2\sum_{i=1}^{m}\operatorname{cov}(X_1, X_{1+i}).$$

This asymptotic distribution may be used to construct confidence intervals for μ, but only provided a consistent estimate of σ_∞^2 is available, and which is far from straightforward to obtain.

Politis and Romano (1994) give a number of other examples of dependent data problems where statistics of interest possess asymptotic distributions, but where use of these is difficult. The bootstrap provides a route which can bypass difficult problems associated with use of asymptotics, but at a price. It is clear that use of the bootstrap resampling scheme appropriate to independent data will fail to provide consistent approximation even in the case of weakly dependent processes. In the above example, for instance, the simple bootstrap estimate of the variance of $n^{1/2}\overline{X}_n$ converges in probability to $\operatorname{var}(X_1)$ and is therefore inconsistent for σ_∞^2. Identification of a valid resampling scheme requires knowledge of the dependence structure of the observations.

As in the independent setting, bootstrap can be applied parametrically to structured dependent data models, often with improvement over standard asymptotic procedures. Most developments to dependent data problems have considered such structured models. Papers by Athreya and Fuh (1992) and by Datta and McCormick (1992) consider Markov chain models.

However, once again most interest lies in nonparametric resampling schemes. Key developments here are quite recent. Künsch (1989) proposed a "moving blocks" resampling scheme for stationary time series data [see also the paper by Liu and Singh (1992)]. The basic idea here is to break the observed data series x up into a collection of overlapping blocks of observations. Bootstrapped data series are obtained by independent sampling, with replacement, from among these blocks.

We illustrate this procedure in the context of the example above. Let b be a given block size. Define $\xi_i = (x_i, \ldots, x_{i+b-1})$ to be the block of b consecutive observations starting from x_i, $i = 1, \ldots, n - b + 1$. The moving blocks bootstrap is based on sampling with replacement from the collection $\{\xi_1, \ldots, \xi_{n-b+1}\}$. Suppose that k is an integer such that kb is approximately n, and let $\xi_1^*, \ldots, \xi_k^*$ be sampled independently and with replacement from $\{\xi_1, \ldots, \xi_{n-b+1}\}$. Let the $l = kb$ elements of $\xi_1^*, \ldots, \xi_k^*$ be concatenated into a single vector $(Y_1, \ldots, Y_l) \equiv (\xi_1^*, \ldots, \xi_k^*)$. Then $(Y_1, \ldots, Y_l)$ is the bootstrap sample under the moving blocks bootstrap scheme and, for example, a bootstrap estimate of $P\{n^{1/2}(\overline{X}_n - \mu) \leq z\}$ is $P\{l^{1/2}(\overline{Y}_l - \overline{x}_n) \leq z\}$, where the probability is computed under the moving blocks resampling scheme, and where $\overline{Y}_l = l^{-1}\sum_{i=1}^{l} Y_i$. Consistency under the model of m-dependence is now achieved if b is allowed to grow to infinity with n (see Liu and Singh, 1992).

Alas, the spectacular and immediate higher-order theoretical success of the bootstrap in the independent case, noted by Singh (1981), is not quite as evident with the moving blocks bootstrap.

The paper by Lahiri (1992) demonstrates that the rate of approximation by Künsch's method may be worse than the rate of normal approximation: it is only with suitable modification in the definition of the bootstrapped statistic that second-order correct approximation is obtained. In the above example, such modification amounts to estimating $P\{n^{1/2}(\overline{X}_n - \mu) \leq z\}$ by $P\{l^{1/2}(\overline{Y}_l - E^*\overline{Y}_l) \leq z\}$, where $E^*\overline{Y}_l$ denotes the expectation of $\overline{Y}_l$ under the moving blocks resampling scheme. The paper by Politis and Romano (1992a) suggests a variant of the moving blocks bootstrap which automatically provides this modification. Politis and Romano (1992b) define a generalization of the moving blocks resampling scheme which can be used for inference about parameters associated with the whole, infinite-dimensional, distribution of the observations, such as those that arise in the problem of estimating the spectral density function.

It would appear, therefore, that in the dependent data context validity of the bootstrap is a more delicate matter than in the independent case. There any one of a number of procedures may be valid, even if some are judged better than others, and the basic bootstrap is valid in quite wide generality. Theoretical results, such as those developed in LePage and Billard (1992), are strongly conditioned too by the form of dependence assumed. The assumption of m-dependence and conjecture that results will hold also for more general dependence models is widespread. In practice the dependence structure of the observations may have to be estimated. Key questions arise here also in relation to the assumption of stationarity of the data-generating mechanism. Lahiri (1992) includes the reassuring result that the moving blocks bootstrap enjoys a certain degree of robustness against departure from stationarity. Somewhat worrying, however, is the observation (Léger, Politis and Romano, 1992) that, while it assumes the mech-

anism generating the data to be stationary, the moving blocks bootstrap utilizes resampled data series which lack this stationarity property. The "stationary bootstrap" of Politis and Romano (1991) solves this problem.

5.2 Immediate Needs

The appealing simplicity of Efron's original bootstrap idea for independent data is somewhat lost for dependent data. Quite sophisticated procedures may be required to cope successfully and provide valid inference in this setting. How in practice, given only sample data, are we to identify an asymptotically justified resampling procedure? A pragmatic answer to this question for confidence interval construction with independent data may now be available: use an iterated percentile interval (Hall, 1992a, page 142). No such general answer is yet available for dependent data problems. We have noted already that, even with more theory, serious practical difficulties remain. All resampling procedures that have been described for dependent data depend in some way on a "tuning constant," such as the block size b in the above description of the moving blocks bootstrap. Practical guidelines for choice of such constants, essential to use of the schemes in any setting, are scarce and lacking in generality.

To date there has been very limited empirical study of bootstrap procedures for dependent data. Most work, such as the papers in LePage and Billard (1992), is highly theoretical and little concerned with statistical practice. Once again, though, such doubts about the current utility of bootstrap in this area must be tempered by the realization that real advances of practical importance should be obtainable, especially in such areas as spectral density estimation, through use of the bootstrap. However, the methods required may be somewhat sophisticated and far from automatic.

6. PRACTICALITIES

Practicalities are paramount. If bootstrap *is* to become a standard part of the statistician's tool kit, then it must be presented both sharpened and ready for use. As we have noted above, for some jobs bootstrap is a sensible and feasible option. For other circumstances it is still pretty well on the drawing board and needs much sharpening. There are two reasons why everyday use is not widespread. First, as we have seen, in many settings 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. A major difficulty here is noted by Léger, Politis and Romano (1992). The bootstrap is a class of procedures, rather than a single method. Even with availability of routines for use within packages such as S for bootstrap sampling, such as those detailed in the appendix to Efron and Tibshirani (1993), the user is left the task of identifying the objects to be resampled. While books such as those by Noreen (1989) and Westfall and Young (1993) help popularize the approach, LePage and Billard (1992) suggests that, in this area as much as in any other, researchers are little concerned with the packaging of their theory. They have preferred to tackle practical issues of a different kind.

Bootstrap estimates can rarely be expressed in closed form, and most often Monte Carlo simulation must be used for their approximation. Recently much attention has been paid to questions relating to efficient computation for the bootstrap; some of this work is reviewed by Hall (1992b) and by Efron and Tibshirani [(1993), Chapter 23]. Inventive though this work often is, is it important? Potentially, yes. However, despite the increasing range of problems tackled by bootstrap and the increasing number of applied publications which make use of bootstrap for error assessment, there is little evidence—certainly not in LePage and Billard (1992)—of *routine* and *repeated* use of bootstrap. The computational savings which may be achieved through use of the fairly sophisticated efficient simulation schemes may not have much practical significance. Consideration of the large volume of work in this area leads to the pragmatic conclusion that simple resampling schemes, such as the "balanced" bootstrap, strike an effective balance between the competing demands of implementational simplicity and efficiency (see Hall, 1992b).

Related to work on variance reduction methods for use in a bootstrap simulation is work on analytic approximation methods which replace entirely the need for a Monte Carlo simulation [see Davison and Hinkley (1988) and the extensions given by Daniels and Young (1991) and DiCiccio, Martin and Young (1992a, b; 1994)]. The same criticisms may, however, be levelled against these methods. They are relatively complex, compared to standard bootstrap sampling, and computational savings are probably not very worthwhile in general. The notable exception here is with the iterated bootstrap, where analytic approximation has enabled (DiCiccio, Martin and Young, 1992b) routine construction of accurate confidence intervals with significant computational savings over the standard iterated resampling scheme. As iterated procedures must be viewed as likely candidates for solution of many of the outstanding problems of bootstrap methodology, such as determining suitable tuning constants, these analytic methods seem worthy of further development. Analytic meth-

ods of this kind may have potential also as diagnostic tools. It has been demonstrated (Daniels and Young, 1991) that analytic procedures applied to the empirical distribution function may break down in the presence of outliers or with data samples from "extreme" underlying distributions, and so may provide a means of identifying empirically circumstances where bootstrap is unreliable.

Analytic procedures are likely also to be of value when considering what Efron and LePage (1992) recognize as important recent developments and refer to as "double dip" bootstrap. The idea here is that of choosing between different estimators (possibly different bootstrap estimators) that with smallest bootstrap estimated error and then subsequently using bootstrap to assess the sampling properties of the adaptively chosen estimator.

7. THE FUTURE?

There seems no limit to the circumstances where bootstrap has a role. What will be the next phase of development of bootstrap? Two areas which seem likely to attract much attention are evident.

Exciting new developments, represented, for example, in the papers by Marron (1992) and by Brownstone (1992), concern use of the bootstrap *within* other statistical procedures, such as bandwidth selection problems in curve estimation and variable selection problems in linear models. Here the emphasis is on demonstrating usefulness of bootstrap as a competitor to more conventional methods, within an accepted framework or with an accepted class of estimators, rather than on replacement of these methods with resampling procedures which assume less. Such use of bootstrap is somewhat different in tone from that envisaged by Efron's introduction of bootstrap, but less controversial. As we might expect, care must be taken with the form of resampling method used, as is crucially the case with dependent data problems. For example, Hall (1992a, Chapter 4) gives a detailed account of the extensive suite of bootstrap techniques that can be used to construct confidence intervals in nonparametric curve estimation problems and discusses their comparison.

Although much of the work on use of bootstrap within other statistical procedures is still in its infancy, early results are encouraging. Bootstrap compares very favourably with other more conventional methods and, by avoiding the need for extensive analysis, enjoys the advantage of simplicity.

Another discernible thread in the development of bootstrap—and much of the content of LePage and Billard (1992)—represents a reversion to basics. The concern here is with bootstrapping in more general settings, and in particular with establishing consistency of bootstrap in M-estimation, with bootstrapping U-statistic quantiles and with elucidation of asymptotic accuracy of bootstrap in these situations. This sort of work is important. The contexts to which the results apply (robust estimation, Hodges–Lehmann estimators etc.) are highly relevant to precisely the sort of circumstances—where there is limited knowledge about the underlying distribution—for which bootstrap was designed. It is puzzling that such work has not been undertaken before. Again, one must reflect upon whether interest in comparison between higher-order properties of competing bootstrap procedures, while it has revealed important insights and seen the development of important general theoretical tools such as Edgeworth expansion, has delayed exhaustive consideration of the questions that may be more significant from the viewpoint of statistical practice. Can we effectively delimit circumstances where bootstrap works? In what circumstances does bootstrap provide valid inference, while simpler alternatives such as normal approximation are not available? Much of the material of LePage and Billard (1992) and related work such as Mammen (1992) represents a sensible return to grapple with these basic issues.

Also to be stressed, and balanced against enthusiasm for bootstrap, are circumstances where bootstrap works, but equally good rates of convergence are obtainable by simpler means. For example, De Angelis, Hall and Young (1993b) consider bootstrap approximation of the distribution of the regression parameter estimators in an L^1 setting. The simplest resampling scheme provides consistent estimation and may be improved upon by use of a smoothed bootstrap scheme, but only to an extent that may more easily be obtained by a normal approximation. A systematic identification of these circumstances will no doubt be made, but is overdue.

As reading LePage and Billard (1992) makes clear, research activity and enthusiasm for bootstrap undoubtedly remains strong. But will it remain as strong in the future as it has over the past few years? There are reasons for doubt.

Recent work [see, e.g., the papers by Wu (1990), Tu (1992) and Politis and Romano (1994)] on distribution estimation using subsample values of a statistic has served as a reminder of the fact that bootstrap is in essence a modern expression to old ideas and has seen a rekindling of interest in alternative resampling-based procedures. One wonders whether this interest admits to the notion that hard work done to establish valid bootstrap inference might well have been focussed on establishing properties of other methods as well. Sometimes these other procedures may be consistent under weaker conditions, which are easily checked (Politis and Romano,

1994), although they may not be fully automatic in that they require specification of resample size, which may be practically difficult. The paper by Tu (1992) shows how, with hard work, versions of these methods may be obtained which yield second-order accuracy. Although general resampling procedures of this kind are currently rather undeveloped by comparison with bootstrap, interest in their applicability endorses an important point. The bootstrap may have radically altered much of how statistical inference may be performed, but it has not completely displaced previous ideas of statistical inference, such as jackknife, in which it has its roots.

Recent years have seen also the emergence of very significant work on likelihood-based nonparametric inference. Some of this work is summarised by Efron and Tibshirani [(1993), Chapter 24]. Empirical likelihood and bootstrap likelihood (Owen, 1988; Davison, Hinkley and Worton, 1992) compete with bootstrap in many circumstances and may be valid under weaker conditions. Recent work by Hall and Chen (Chen, 1993; Chen and Hall, 1993) indicates the great potential scope of empirical likelihood and suggests that in accuracy terms nonparametric likelihood may be a strong competitor to bootstrap. These nonparametric likelihood methods are rooted firmly in classical statistical ideas, such as profile and partial likelihood, and have shown how classical likelihood-based procedures for construction of confidence regions, hypothesis tests and so on may be successfully adapted to the distribution-free context. 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.

8. CONCLUDING REMARKS

In this paper we have attempted to give a critical snapshot of recent research activity on bootstrap and related procedures. As summarized by Efron and LePage (1992) in their introduction to LePage and Billard (1992), we can really discern three main strands of activity within bootstrap research:

1. First, there is work devoted to exploration of the limits of applicability of bootstrap. The bootstrap idea, especially when linked to inventiveness, can be adapted to tackle most problems. These are the really fundamental consistency questions, and the answers are encouraging, even if the versions of bootstrap that emerge are in many cases still rather cumbersome and less simple than we might have hoped.
2. Second, identification of conditions under which bootstrap is valid has renewed interest in other procedures, such as generalised forms of jackknife, which may be valid under weaker conditions, or at least conditions which are easier to check. Much remains to be done to establish versions of these methods which both will be practical and will demonstrate higher-order accuracy.
3. Third, there is work which examines higher-order accuracy properties of bootstrap methods and explores how this accuracy can be harnessed to the task of producing accurate statistical inference in an automatic way. This work includes questions relating to implementation and choice of the "correct" procedure for a given problem.

Progress on 3 has been greatest, but for the independent data case only. There is still much to be done under 1 and 2. Efron and LePage (1992) note that the goal of producing highly accurate confidence intervals by bootstrap seems to be moving toward a practical solution. The implication is clear. Even for this relatively narrow goal, the bootstrap is not quite established as part of the standard statistical armoury.

So what should be the final conclusion? Efron's realization of the scope for use of resampling ideas, especially when combined with modern computing power, has certainly had a significant impact on much of statistical thought, but relatively little on statistical practice. Nonparametric bootstrap investigations are regularly used to supplement more standard statistical analyses in simple problems of error estimation, but there the story often ends. Theoretical study of bootstrap has been undertaken with great energy, for a whole range of problems. Much of this work has shown the bootstrap approach to be a potentially powerful one, but it is not one that has been much embraced in practice outside the independent sample setting. Our contention is that this is a consequence to be expected of the sophisticated nature of bootstrap procedures required to cope reliably in most settings and of the predominantly theoretical flavour of most bootstrap research. But the bootstrap *can* cope reliably and provide useful methodology in many statistical problems, sophisticated as well as simple. Provided research activity keeps sight of the need for practicality, in a way that has sometimes been absent, with further development the bootstrap ought to realize its potential as much more than a stab in the dark.

REFERENCES

ATHREYA, K. B. (1987). Bootstrap of the mean in the infinite variance case. *Ann. Statist.* **15** 724–731.

ATHREYA, K. B. and FUH, C. D. (1992). Bootstrapping Markov chains. In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, eds.) 49–64. Wiley, New York.

BERAN, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika* **74** 457–468.

BERAN, R. and DUCHARME, G. R. (1991). *Asymptotic Theory for Bootstrap Methods in Statistics*. Les Publications CRM, Montréal.

BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217.

BRETAGNOLLE, J. (1983). Lois limites du bootstrap de certaines fonctionelles. *Ann. Inst. H. Poincaré* **19** 281–296.

BROWNSTONE, D. (1992). Bootstrapping admissible linear model selection procedures. In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, eds.) 327–344. Wiley, New York.

CHEN, S. X. (1993). On the accuracy of empirical likelihood confidence regions for linear regression model. *Ann. Inst. Statist. Math.* **45** 621–637.

CHEN, S. X. and HALL, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. *Ann. Statist.* **21** 1166–1181.

DANIELS, H. E. and YOUNG, G. A. (1991). Saddlepoint approximation for the studentized mean, with an application to the bootstrap. *Biometrika* **78** 169–179.

DATTA, S. and MCCORMICK, W. P. (1992). Bootstrap for a finite state Markov chain based on i.i.d. resampling. In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, eds.) 77–97. Wiley, New York.

DAVISON, A. C. and HINKLEY, D. V. (1988). Saddlepoint approximations in resampling methods. *Biometrika* **75** 417–431.

DAVISON, A. C., HINKLEY, D. V. and WORTON, B. J. (1992). Bootstrap likelihoods. *Biometrika* **79** 113–130.

DE ANGELIS, D., HALL, P. and YOUNG, G. A. (1993a). A note on coverage error of bootstrap confidence intervals for quantiles. *Math. Proc. Cambridge. Philos. Soc.* **114** 517–531.

DE ANGELIS, D., HALL, P. and YOUNG, G. A. (1993b). Analytic and bootstrap approximations to estimator distributions in L^1 regression. *J. Amer. Statist. Assoc.* **88** 1310–1316.

DE ANGELIS, D. and YOUNG, G. A. (1992). Smoothing the bootstrap. *Internat. Statist. Rev.* **60** 45–56.

DIACONIS, P. and EFRON, B. (1983). Computer-intensive methods in statistics. *Scientific American* **248** 116–130.

DICICCIO, T. J., MARTIÑ, M. A. and YOUNG, G. A. (1992a). Analytical approximations for iterated bootstrap confidence intervals. *Statistics and Computing* **2** 161–171.

DICICCIO, T. J., MARTIN, M. A. and YOUNG, G. A. (1992b). Fast and accurate approximate double bootstrap confidence intervals. *Biometrika* **79** 285–295.

DICICCIO, T. J., MARTIN, M. A. and YOUNG, G. A. (1994). Analytic approximations to bootstrap distribution functions using saddlepoint methods. *Statist. Sinica* **4** 281–296.

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

EFRON, B. (1979b). Computers and the theory of statistics: thinking the unthinkable. *SIAM Rev.* **21** 460–480.

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

EFRON, B. (1987). Better bootstrap confidence intervals (with discussion). *J. Amer. Statist. Assoc.* **82** 171–200.

EFRON, B. (1992). Jackknife-after-bootstrap standard errors and influence functions (with discussion). *J. Roy. Statist. Soc. Ser. B* **54** 83–127.

EFRON, B. and LEPAGE, R. (1992). Introduction to bootstrap. In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, eds.) 3–10. Wiley, New York.

EFRON, B. and TIBSHIRANI, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy (with discussion). *Statist. Sci.* **1** 54–96.

EFRON, B. and TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

HALL, P. (1988). Theoretical comparison of bootstrap confidence intervals (with discussion). *Ann. Statist.* **16** 927–985.

HALL, P. (1992a). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

HALL, P. (1992b). Efficient bootstrap simulation. In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, eds.) 127–143. Wiley, New York.

HALL, P., DICICCIO, T. J. and ROMANO, J. P. (1989). On smoothing and the bootstrap. *Ann. Statist.* **17** 692–704.

HALL, P. and MARTIN, M. A. (1988). On bootstrap resampling and iteration. *Biometrika* **75** 661–672.

HALL, P., MARTIN, M. A. and SCHUCANY, W. R. (1989). Better non-parametric bootstrap confidence intervals for the correlation coefficient. *J. Statist. Comput. Simulation* **33** 161–172.

KÜNSCH, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17** 1217–1241.

LAHIRI, S. N. (1992). Edgeworth correction by 'moving block' bootstrap for stationary and nonstationary data. In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, eds.) 183–214. Wiley, New York.

LÉGER, C., POLITIS, D. N. and ROMANO, J. P. (1992). Bootstrap technology and applications. *Technometrics* **34** 378–398.

LEPAGE R. and BILLARD, L. (eds.) (1992) *Exploring the Limits of Bootstrap*. Wiley, New York.

LIU, R. Y. and SINGH, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, eds.) 225–248. Wiley, New York.

LOH, W. (1987). Calibrating confidence coefficients. *J. Amer. Statist. Assoc.* **82** 155–162.

MAMMEN, E. (1992). *When Does Bootstrap Work? Asymptotic Results and Simulations*. Springer, New York.

MARRON, J. S. (1992). Bootstrap bandwidth selection. In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, eds.) 249–262. Wiley, New York.

MARTIN, M. A. (1990). On bootstrap iteration for coverage correction in confidence intervals. *J. Amer. Statist. Assoc.* **85** 1105–1118.

NOREEN, E. W. (1989). *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley, New York.

OWEN, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249.

OWEN, A. B. (1991). Empirical likelihood and small samples. In *Computing Science and Statistics. Proceedings of the 22nd Symposium on the Interface* (C. Page and R. LePage, eds.) 79–88. Springer, New York.

POLITIS, D. N. and ROMANO, J. P. (1991). The stationary bootstrap. Technical Report 365, Dept. Statistics, Stanford Univ.

POLITIS, D. N. and ROMANO, J. P. (1992a). A circular block-resampling procedure for stationary data. In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, eds.) 263–270. Wiley, New York.

POLITIS, D. N. and ROMANO, J. P. (1992b). A general resampling scheme for triangular arrays of α-mixing random variables with application to the problem of spectral density estimation. *Ann. Statist.* **20** 1985–2007.

POLITIS, D. N. and ROMANO, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* **22** (4).

ROMANO, J. P. (1988). On weak convergence and optimality of kernel density estimates of the mode. *Ann. Statist.* **16** 629–647.

SCHENKER, N. (1985). Qualms about bootstrap confidence intervals. *J. Amer. Statist. Assoc.* **80** 360–361.

SILVERMAN, B. W. and YOUNG, G. A. (1987). The bootstrap: to smooth or not to smooth? *Biometrika* **74** 469–479.

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.