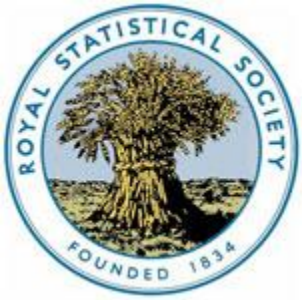

Bootstrap Methods
Author(s): David V. Hinkley
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 50, No. 3 (1988), pp. 321-337
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2345698
Accessed: 02-03-2017 22:59 UTC

J. R. Statist. Soc. B (1988)
50, *No.* 3, *pp.* 321–337

Bootstrap Methods

By DAVID V. HINKLEY†

The University of Texas at Austin, USA

[*Read before the* Royal Statistical Society *on Wednesday, March 16th, 1988, at a meeting organized by the* Birmingham Group, *Professor* J. B. Copas *in the Chair*]

SUMMARY

A survey of some developments in bootstrap methodology is given. Topics include confidence limits, significance tests, empirical likelihoods, conditioning, double bootstrapping, and numerical techniques. Special attention is given to regression problems. There are brief remarks about more complex problems, including variance component problems, time series and nonparametric regression.

Keywords: BALANCED SAMPLES; SADDLEPOINT METHODS; PIVOTS; CONFIDENCE LIMITS; SIGNIFICANCE TESTS; CONDITIONAL INFERENCE; MONTE CARLO METHODS; JACKKNIFE; LIKELIHOOD; PERMUTATION TEST; REGRESSION; VARIANCE COMPONENTS; TIME SERIES; SAMPLE SURVEYS; NONPARAMETRIC METHODS

1. INTRODUCTION

The essence of bootstrap methods is the simulation of relevant properties of a statistical procedure with minimal model assumptions. The word 'simulation' here is used in the widest possible sense, from simple substitution of an estimated distribution in a formula to complex Monte Carlo simulation of representative samples and their analysis. In any given context bootstrap methods may be similar variously to simulation methods, permutation methods, jackknife methods or other familiar 'resampling' methods. One major focus of research has been the search for reliable, automatic, empirical methods for calculating confidence limits. Because most bootstrap methods involve numerical approximation, potentially powerful techniques of theoretical and Monte Carlo approximation have been and continue to be studied. As to potential applications, considerable effort has been devoted to classical problems involving means, correlations and regression. But increasingly attention is directed to more complex problems such as those associated with variance components, time series, sample surveys and nonparametric curve fitting.

The aim of the present paper is to review and illustrate many of the developments in bootstrap methodology, so as to highlight key ideas and potential usefulness. The choice of material inevitably reflects personal interests, however, so that the paper is in no way a comprehensive review. The first sections deal with the relatively simple context of homogeneous samples; Sections 2–5 respectively discuss the basic bootstrap method, numerical techniques, confidence limit methods and significance test methods. Regression problems are considered in Section 6, and the idea of a conditional bootstrap introduced there is further discussed in Section 7. Section 8 looks at some recent suggestions for empirical likelihoods. Some more complex applications are outlined in Section 9. Finally, Section 10 contains some general discussion.

† *Address for correspondence*: Department of Mathematics, The University of Texas at Austin, Austin, TX 78712, USA.

 0035–9246/88/50321 $2.00

2. BASIC BOOTSTRAP METHOD

To begin with a very simple example, consider the sample of $n = 10$ measurements $x_1, \ldots, x_{10}$ in the first row of Table 1, whose average and standard deviation are $\bar{x} = 17.87$ and $s = 7.19$. Suppose that we wish to make statistical statements about the accuracy of the sample average $\bar{x}$ as an estimate of μ, the mean of X in the population from which the sample was drawn. For the sake of definiteness, suppose that we wish to know (a) the variance of $\bar{X}$, (b) $\Pr\{c \leqslant \bar{X} - \mu \leqslant d\}$ for specified c and d, and (c) 95% confidence limits for μ on either side of $\bar{x}$.

One classical approach would be to describe random variation in sampled X values by a distribution function $F(x \mid \theta) = \Pr\{X \leqslant x\}$, with θ an unknown parameter (vector or scalar) which includes μ. Possible answers to problems (a)–(c) are found by theoretical calculation based on F with an estimate $\hat{\theta}$ in place of θ. For example, if F is the cdf of the $N(\mu, \sigma^2)$ distribution, so that $\theta = (\mu, \sigma^2)$, then the variance of $\bar{X}$ is σ^2/n, which we usually calculate with $s^2 = (n-1)^{-1}\Sigma(x_i - \bar{x})^2$ in place of σ^2. In bootstrap terminology, this is a *parametric bootstrap* calculation.

The *nonparametric bootstrap*, more usually called simply *bootstrap*, approach is to not assume anything about the form of F, only that it exists. Then in place of $F(x \mid \hat{\theta})$ one might use the empirical cdf

$$\tilde{F}(x) = n^{-1} \sum hv(x - x_i),$$

TABLE 1

A random sample and small bootstrap analyses of its mean†

Bootstrap sample	*Frequencies of datum values for the following data*										$\bar{x}^*$
	9.6	*10.4*	*13.0*	*15.0*	*16.6*	*17.2*	*17.3*	*21.8*	*24.0*	*33.8*	
Simple bootstrap											
1	1	0	0	1	3	1	1	0	2	1	19.07
2	1	0	1	1	1	1	0	3	2	0	18.48
3	0	0	2	1	2	0	2	0	3	0	18.08
4	1	1	1	2	0	1	1	1	0	2	18.69
5	1	0	1	1	3	1	1	1	1	0	16.77
6	1	1	2	0	0	1	1	2	1	1	18.19
7	0	1	3	1	0	1	3	0	1	0	15.75
8	2	1	0	0	2	1	0	0	2	2	19.56
9	1	1	1	2	0	0	1	1	1	2	19.37
10	0	1	2	0	2	1	0	3	1	0	17.62
Sample average of $\bar{x}^*$s = 18.16, sample variance of $\bar{x}^*$s = 1.41											
Randomized block bootstrap											
1	0	0	1	1	3	1	0	0	2	2	21.06
2	1	3	1	0	1	0	0	1	2	1	17.40
3	2	0	0	0	1	1	0	2	3	1	20.24
4	1	0	1	0	0	3	3	1	0	1	18.17
5	1	2	1	0	2	0	2	2	0	0	15.48
6	0	2	2	0	1	0	1	2	1	1	18.21
7	1	0	0	3	1	3	0	0	1	1	18.06
8	2	1	1	2	0	0	2	0	1	1	16.50
9	2	1	1	1	0	0	2	1	0	2	18.16
10	0	1	2	3	1	2	0	1	0	0	15.42
Sample average of $\bar{x}^*$s = 17.87, sample variance of $\bar{x}^*$s = 3.33											

†Average $\bar{x} = 17.87$.

where $hv(u) = 0$ $(u < 0)$, $1(u \geqslant 0)$; possibly one would consider a smoothed version of $\tilde{F}$ (Efron, 1982, ch. 5; Silverman and Young, 1987). For problem (a), σ^2 in the formula $\text{var}(\bar{X}) = \sigma^2/n$ would now be calculated with $\tilde{F}$ in place of F as $\tilde{\sigma}^2 = \int x^2 d\tilde{F}(x) - (\int x d\tilde{F}(x))^2 = n^{-1}\Sigma(x_i - \bar{x})^2$, perhaps modified to its unbiased form s^2. There is nothing novel about this, of course, but there is about using $\tilde{F}$ to do the probability calculations for problems (b) and (c).

Consider problem (b) in detail, and rewrite the required probability in the more suggestive form

$$P = \Pr\{c \leqslant \text{mean}(data) - \text{mean}(F) \leqslant d\}. \tag{1}$$

If this is calculated with $\tilde{F}$ substituted for F *everywhere*, the result is the estimate

$$\tilde{P} = \Pr\{c \leqslant \text{mean}(data^*) - \text{mean}(\tilde{F}) \leqslant d\}, \tag{2}$$

where $data^*$ is a random sample of size n drawn from $\tilde{F}$. Because theoretical evaluation of $\tilde{P}$ appears impossible, one might well adopt the strategy of numerical simulation: draw repeated samples $data^*(1), \ldots, data^*(B)$ from $\tilde{F}$, and calculate

$$\tilde{P}_{\text{sim}} = \frac{\text{number of times } c \leqslant \text{mean}(data^*(i)) - \text{mean}(\tilde{F}) \leqslant d}{B}. \tag{3}$$

Table 1 illustrates this for $B = 10$. Each bootstrap sample $data^*(i)$ is recorded in the form of frequencies of original data values. For $c = -1$ and $d = +1$ we get $\tilde{P}_{\text{sim}} = 0.50$, a not very accurate approximation to $\tilde{P} = 0.37$ (see Section 3) resulting from the ridiculously small value of B: it would be customary to have B well in excess of 100.

Note that in the simulation, drawing a random sample from $\tilde{F}$ means simply sampling n values from *data* randomly with replacement. But is this a good numerical strategy? It would not be if we required only $\text{var}(\bar{X})$, because the simpler technique known as the *jackknife* (Miller, 1974; Efron, 1982) uses n systematic samples from *data* and gives the correct answer—here meaning $\tilde{\sigma}^2/n$. Can $\tilde{P}$ itself be calculated without numerical simulation? Such questions are addressed in the next section.

A very different question concerns the accuracy of $\tilde{P}$ as an approximation to, or estimate of, P. If $\tilde{P}$ is very inaccurate, then choosing data-dependent values $c = \tilde{c}$ and $d = \tilde{d}$ to make $\tilde{P} = 0.95$, for example, would make the natural 0.95 bootstrap confidence limit formula

$$\text{mean}(data) - \tilde{d} \leqslant \text{mean}(F) \leqslant \text{mean}(data) - \tilde{c} \tag{4}$$

unreliable. We know from experience that this is likely to happen for small samples of, say, normal or gamma data: the reliable confidence limit methods are based on probabilities for $(\bar{x} - \mu)/s$ and $\bar{x}/\mu$ respectively, not $\bar{x} - \mu$. Is there some way of finding out that $\tilde{P}$ is inaccurate? Is there a general, reliable way to calculate confidence limits for μ? To these questions we return in Section 4.

The example of the average illustrates a general type of problem to which considerable theoretical effort has been directed. Given a statistical estimate $T = t(\tilde{F})$ of population characteristic $\theta = t(F)$, we wish to calculate $Q = E\{R_t(F, \tilde{F}) \mid F\}$; here $E(\cdot \mid F)$ denotes the expectation with respect to F. The quantity $R_t(F, \tilde{F})$ might be simple, e.g. $(\bar{X} - \mu)^2$, or complicated, e.g. the indicator of whether or not $(\bar{X} - \mu)/S \leqslant a$. The nonparametric bootstrap approximation of Q is $\tilde{Q} = E\{R_t(\tilde{F}, \tilde{F}^*) \mid \tilde{F}\}$, where $\tilde{F}^*$ is the empirical cdf of the bootstrap sample $X_1^*, \ldots, X_n^*$ which is drawn randomly

from $\tilde{F}$. While the consistency of $\tilde{Q}$ for Q flows from the consistency of $\tilde{F}$ for F, a more detailed assessment of $\tilde{Q} - Q$ is often useful, especially if one is trying to compare confidence limit procedures or if alternative approximations to Q are being considered. The majority of theoretical results (see Beran (1982, 1984) and Hall (1987a), and references therein) deal either with estimates T which are functions of vector averages, so that standard expansion techniques apply, or with estimates representable by Volterra series,

$$T = \theta + n^{-1} \sum a_1(X_j; F) + n^{-2} \sum\sum a_2(X_i, X_j; F) + \ldots, \tag{5}$$

in which a_1 is the influence function of T. Some of the relevant results for confidence limit methods are reviewed by DiCiccio and Romano (1988).

In what follows, the discussion focuses first on some of the questions raised in this section, and then reviews a variety of bootstrap methods, in a rather non-technical way. Throughout we shall denote bootstrap samples of data by $X_1^*, \ldots, X_n^*$ and corresponding statistics by T^*.

3. NUMERICAL TECHNIQUES

The exact calculation of property $Q = E\{R_t(\tilde{F}, \tilde{F}^*) \mid \tilde{F}\}$ is ordinarily not possible. There are essentially two ways to proceed: theoretical approximation and purely numerical approximation.

The simplest type of theoretical approximation would be to replace $T = t(\tilde{F})$ by its linear approximation

$$T_L = t(F) + n^{-1} \sum a_1(X_j; F), \tag{6}$$

i.e. the first two terms on the right of (5). From T_L is derived the $N(0, \tilde{V})$ approximation for the distribution of $n^{1/2}(T - \theta)$, with $\tilde{V} = n^{-1} \Sigma\{a_1(X_j; \tilde{F})\}^2$. This is the (infinitesimal) jackknife method, which may often be adequate, but which negates a potential advantage of bootstrap methods, namely high order or small sample accuracy.

The simplest example of numerical approximation, illustrated by (3), is the generation of B samples x_b^*, $b = 1, \ldots, B$, from $\tilde{F}$ followed by calculation of

$$\tilde{P}_{\text{sim}} = B^{-1} \sum_{b=1}^{B} R_t(\tilde{F}, \tilde{F}_b^*).$$

The required magnitude of B will depend on the form of R_t, but will often be at least 100.

A general discussion of improvement in numerical techniques by Thernau (1983) suggests several approaches, including the importance sampling and control methods familiar in Monte Carlo methodology. The different approach of balanced sampling has been studied in more detail (Obgonmwan and Wynn, 1986; Davison *et al.*, 1987; Graham *et al.*, 1987). The central idea here can be expressed in two ways, the more profitable of which is as follows. Write a simulated sample from $\tilde{F}$ as $(x_{\xi(1)}, \ldots, x_{\xi(n)})$, and $\xi = (\xi(1), \ldots, \xi(n))$. Then the B vectors $\xi_1, \ldots, \xi_B$ which define the bootstrap simulation should cover the n-dimensional lattice cube $\{1, 2, \ldots, n\}^n$ in as uniform a manner as possible. Exact uniformity on one- and two-dimensional margins is achievable by use of classical experimental designs. For example, one-dimensional balance is achieved if the $B \times n$ matrix with (b, i)th element $\xi_b(i)$ defines a randomized block design with columns as blocks, entries as treatment labels. The second half of

Table 1 illustrates this with $B = 10$, corresponding to a single randomized block. Note that the average of the 10 $\bar{x}^*$s is necessarily equal to $\bar{x}$, thereby yielding a correct estimate of zero bias for $\bar{X}$:

$$\text{estimated bias} = B^{-1} \sum (\bar{x}_b^* - \bar{x}) = 0.$$

Also the variance of the $\bar{x}^*$s is closer to the correct value $n^{-1}\tilde{\sigma}^2$ for the variance of $\bar{X}$.

Two-dimensional balance can be achieved using orthogonal Latin squares, and a somewhat weaker form of balance, suitable for homogeneous data, is achievable using balanced incomplete block designs (Graham *et al.*, 1987). What two-dimensional balance gives is error-free approximation of bias and variance for the linear part of a statistic, which for large samples is adequate.

What do such designs achieve in practical terms? Probably a fourfold or fivefold reduction in B for any given level of simulation error, if we are approximating moments of T. But for estimating the $100p$th percentile, say, of $T - \theta$ by the $(B+1)p$th ordered value of $T^* - T$, balanced designs are not so effective, especially for $p < 0.05$ or $p > 0.95$. It seems quite likely that a more effective strategy is to select among one-dimensional balanced designs using a rejection technique along the lines suggested by Ogbonmwan and Wynn (1986). Further research is needed in this area.

Switching now to theoretical approximation, particularly for the probability distribution of T^*, one elementary approach is to modify normal approximations with Edgeworth corrections. More interesting, and usually more effective, is the use of saddlepoint approximations (Davison and Hinkley, 1988). For example, consider again $T = \bar{X}$, and write the empirical cumulant generating function of X as

$$\tilde{K}(\lambda) = \log \int e^{\lambda x} d\tilde{F}(x) = \log\left(n^{-1} \sum e^{\lambda x_i} \right).$$

Then a direct application of equation (4.9) of Daniels (1987) gives

$$\tilde{P} = \Pr(T^* - t \leqslant y \mid \tilde{F}) \doteq \Phi(w_y) + \phi(w_y)(w_y^{-1} - z_y^{-1}),$$

where

$$w_y = [2n\{\lambda_{t+y}(t+y) - \tilde{K}(\lambda_{t+y})\}]^{1/2} \operatorname{sgn}(\lambda_{t+y}),$$

$$z_y = \lambda_{t+y}\{n\tilde{K}''(\lambda_{t+y})\}^{1/2}$$

with λ_{t+y} the unique solution of $\tilde{K}'(\lambda) = t + y$. Table 2 gives a brief summary of numerical results so obtained for the data of Table 1, in the form of percentile approximations. Comparison is made to exact results (simple numerical simulation with $B = 50\,000$) and normal approximation results. The saddlepoint approximation is excellent.

There are two difficulties with the saddlepoint approximation method in this context. First is a technical difficulty associated with the discreteness of $\tilde{F}$; this makes formal proofs of asymptotic expansions complicated, but not impossible. More important is the limited range of problems to which known saddlepoint approximations apply, essentially those for which T solves a linear estimating equation of the form $\Sigma\psi(X_j, T) = 0$ with $\psi(x, t)$ monotone in t. An *ad hoc* approximation can be obtained via series expansions of $T^* - T$, but the result does not have the degree of accuracy typical for saddlepoint methods. A key unsolved problem is to derive saddlepoint

TABLE 2

Approximations to bootstrap percentage points for $\bar{X} - \mu$; data in Table 1

	p							
	0.001	*0.01*	*0.05*	*0.10*	*0.90*	*0.95*	*0.99*	*0.999*
Exact percentile†	−6.34	−5.55	−3.34	−2.69	2.87	3.73	5.47	7.52
Saddlepoint percentile	−6.31	−5.52	−3.33	−2.69	2.85	3.75	5.48	7.46
Normal percentile	−8.46	−7.03	−3.74	−2.91	2.91	3.74	5.29	7.03
Fisher–Cornish	−6.51	−5.74	−3.48	−2.81	3.00	3.97	5.89	8.19

† From 50 000 random samples.

approximations for non-linear statistics such as $T = n^{-1}\Sigma a(X_j) + n^{-2}\Sigma\Sigma b(X_i, X_j)$: such approximations would give accurate results for statistics with expansion (5).

What of the other possible numerical techniques? The Monte Carlo control method can be applied to approximate moments of a statistic, for example using T_L in (6) as control, since T_L has known moments under sampling from $\tilde{F}$. Use of the Monte Carlo method of importance sampling is currently under investigation by Dr A. C. Davison. For approximation of probabilities, such as (2), one obvious approach is to apply smoothing techniques to the empirical distribution of simulated values of the relevant statistical quantities, such as $\bar{X}^* - \bar{X}$.

4. CONFIDENCE LIMIT METHODS

The most studied problem in (nonparametric) bootstrap methodology is the determination of reliable confidence limit procedures. This is the subject of the companion paper by DiCiccio and Romano (1988), so an exhaustive survey will not be attempted here.

The basic problem arises from the discrepancy between (1) and (2). In principle a confidence interval procedure for parameter θ based on estimate T would be solved by finding a_P such that $\Pr(T - \theta \leqslant a_P) = P$, for given P. Then, for example, equitailed $1 - \alpha$ limits for θ would be $T - a_{1-\alpha/2}$ and $T - a_{\alpha/2}$, cf. (4). Bootstrap estimates $\tilde{a}_P$ are usually not satisfactory, in essence because $T^* - T$ is not pivotal for $\tilde{F}$s within probable range of F. A useful analogy is the problem of setting confidence limits for a normal mean, where the $N(0, \tilde{\sigma}^2/n)$ approximation for $\bar{x} - \mu$ would not give a satisfactory confidence distribution for μ if n were very small. Actually the solution to the latter problem suggests at least one of several possible approximate solutions for the nonparametric bootstrap problem.

One way to construct a reliable confidence limit procedure is to construct an invertible pivot, say $Q(T, \theta, \mathbf{S})$ with $\mathbf{S}$ containing relevant ancillary features. Familiar examples in classical statistics are Student's t statistic for a normal mean, and $\bar{X}/\mu$ for an exponential mean. In the bootstrap context we would require that $Q^* = Q(T^*, T, \mathbf{S}^*)$ be very close to pivotal under sampling from $\tilde{F}$s within probable range of F. Analogy with the normal mean problem suggests trying $Q = (T^* - \theta)/S^*$ with S^* a nonparametric estimate of standard error such as is provided by a jackknife method (Miller, 1974; Efron, 1982, ch. 6). In his detailed theoretical comparison of confidence limit procedures, Hall (1988) shows that this Studentized form leads to one-sided confidence limits whose coverage is correct to $O(n^{-1/2})$.

A different pivotal construction is offered by Beran (1987), who mimics the probability integral transform approach. Thus if $Q_0(T, \theta)$ has cdf $\tilde{G}_0$ under sampling from $\tilde{F}$, then $Q = \tilde{G}_0(Q_0(T, \theta))$ is very nearly pivotal. If $\tilde{G}$ is the distribution function of Q under sampling from $\tilde{F}$, and if Q is monotone in θ, then solutions to

$$\tilde{G}(Q(T, \theta)) = \tfrac{1}{2}\alpha \text{ and } 1 - \tfrac{1}{2}\alpha$$

define approximate equitailed $1 - \alpha$ limits for θ. The difficulty is that $\tilde{G}$ is based on second-level bootstrapping, i.e. sampling from samples from $\tilde{F}$: see below. On the surface this suggests the need for a rather extravagant numerical simulation, perhaps using 10^5 or 10^6 samples. The theoretical and numerical properties are comparable to those for the Studentized estimate approach outlined above. The method seems worthy of further study.

It may be appropriate here to say a little more about second-level bootstrapping, a process which has several potential uses. Suppose that one wants to check whether or not $T - \theta$ is pivotal, considering this in the first instance as the limited question as to whether or not $\text{var}(T \mid F) = \sigma^2(\theta)$ is in fact constant. (Note that this ignores the possibility of another parameter ϕ affecting the distribution of T.) An empirical strategy is to simulate several samples from each of several populations, each of which has a different value of θ. For each population, then, one obtains an estimate of $\sigma^2(\theta)$: these estimates are compared to assess possible dependence on θ. In the nonparametric bootstrap context, a population and its θ value are equated to a simulated sample $(x_1^*, \ldots, x_n^*)$ and its θ estimate t^*. Therefore $\sigma^2(t^*)$ is estimated by taking samples $(x_1^{**}, \ldots, x_n^{**})$ from $(x_1^*, \ldots, x_n^*)$ and computing the empirical variance $\tilde{\sigma}^2(t^*)$ of the t^{**} values which are the θ estimates calculated from $(x_1^{**}, \ldots, x_n^{**})$. One might take 50 t^{**}s for each of 20 t^*s. This idea appears to be due to P. L. Chapman; see Chapman and Hinkley (1986).

By way of illustration, Fig. 1(a) shows estimated 5th and 95th percentiles of the error in sample correlation coefficient r for 20 values of population correlation ρ, all obtained from two-level bootstrapping of one sample of $n = 20$ bivariate normal pairs. Fig. 1(b) gives corresponding results for Fisher's z transform, $z = \tanh^{-1} r$. Note that in the first plot, the estimated percentiles of $r - \rho$ mimic the normal theory trend: the fitted curves are close to $\pm 1.645(1 - \rho^2)/\sqrt{n}$. The plot suggests strongly that $r - \rho$ is not pivotal. On the other hand, the near-horizontal trends of percentiles in the second plot suggest that error in z is very nearly pivotal. This would imply that bootstrap results for z are reliable approximations to theoretical properties of z.

Beran's pivotal construction is not the only confidence limit method based on second-level bootstrapping. More recently Tibshirani (1987) has considered explicit use of smoothed versions of $\tilde{\sigma}^2(t^*)$ to obtain a variance-stabilized estimate

$$U = h(T) = \int^T \{\tilde{\sigma}^2(t^*)\}^{-1/2} \mathrm{d}t^*,$$

to which is then applied a confidence limit procedure for the invertible function $h(\theta)$. Initial results show the method to be competitive with the best known methods in many problems.

The final method to be mentioned is the accelerated bias-corrected percentile method of Efron (1987), which attempts implicit rather than explicit use of variance stabilization, while at the same time recognizing that the variance-stabilized estimate

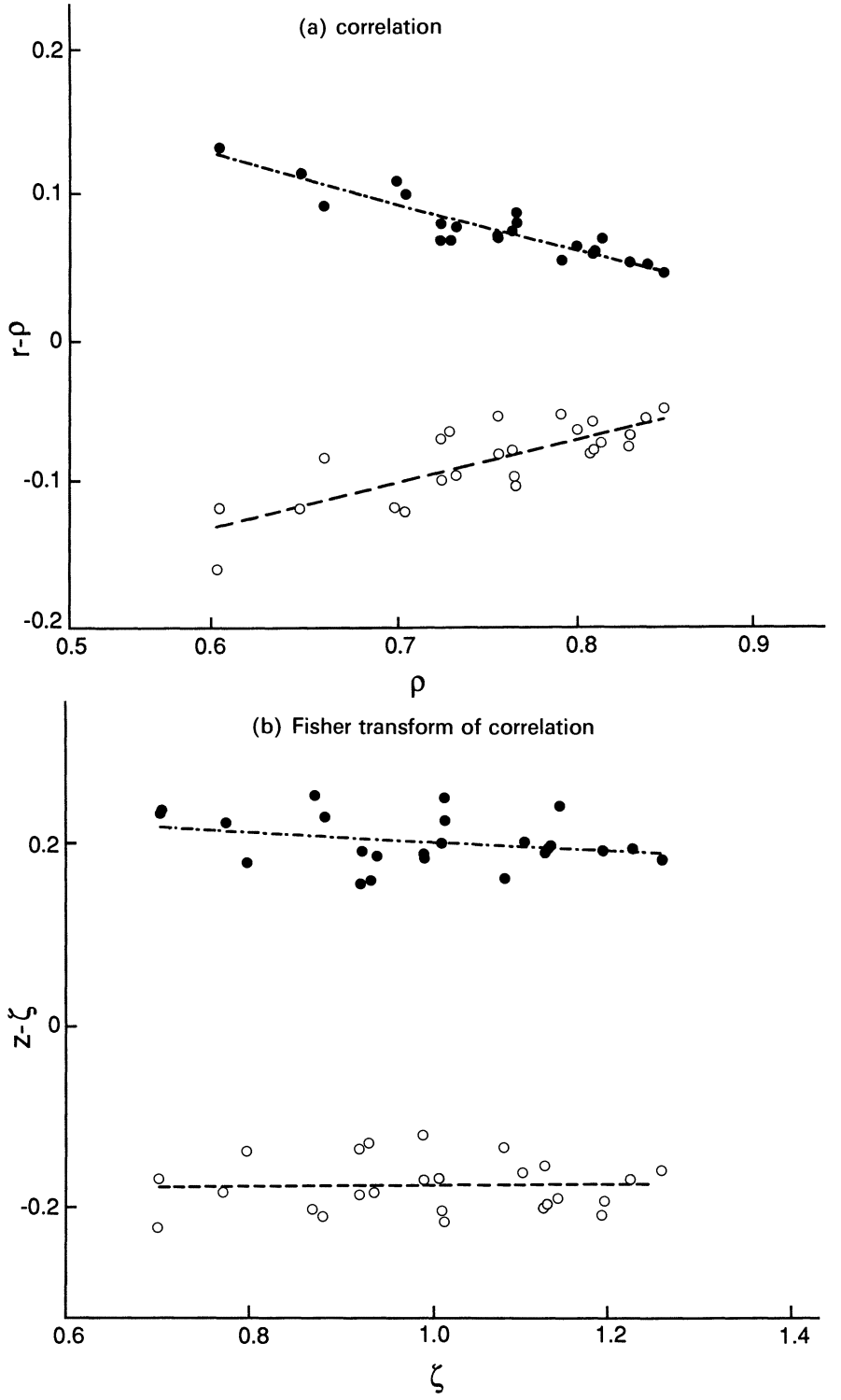

Fig. 1. Bootstrap estimates of 5% (○) and 95% (●) quantiles of (a) $r-\rho$ and (b) $z-\zeta=\tanh^{-1}r-\tanh^{-1}\rho$ obtained from analysis of one sample of $n=20$ pseudo-normal pairs: values of ρ are $B=25$ r^*s; estimated quantiles of r are quantiles of empirical cdf of r^{**} from 100 second-level bootstrap samples

$h(T)$ may have bias of order n^{-1} and standardized skewness of order $n^{-1/2}$. This leads to the working assumption that for appropriate $h(\cdot)$ and constants τ, α and β_0,

$$Q = \frac{\tau\{h(T) - h(\theta)\}}{1 + \alpha\tau h(\theta)} + \beta$$

has a standard normal distribution. Efron's use of this assumption in the bootstrap context does not involve knowing $h(\cdot)$, τ, α or β. The reliability of the resulting confidence limit method is rather uneven, albeit often very good. One obvious defect is that for large enough α, Q may not be monotone over an appropriately wide range for θ. DiCiccio and Romano (1988) discuss the method in detail.

There are many empirical studies of the performances of bootstrap confidence limits, and the results shown in Table 3 seem quite representative. Here T is the mean $\bar{X}$ of samples of size $n = 20$, artificially generated from the χ_1^2 distribution. For each sample, bootstrap simulation with $B = 1000$ was used. Table 3, taken from Owen (1987) shows empirical error rates of nominal 90% equitailed intervals for mean θ, based on 1000 data sets.

A different approach to bootstrap assessment of parameter uncertainty is via a nonparametric likelihood. This is discussed separately in Section 8.

5. SIGNIFICANCE TESTS

The connexion between confidence limits and significance tests (Cox and Hinkley, 1974) may be exploited to test certain kinds of hypotheses about parameters. But a direct approach is also possible using bootstrap techniques, particularly for 'pure significance tests' (Cox and Hinkley, 1974, ch. 3). There are, of course, connexions to other nonparametric methods of testing.

Suppose that T is a statistic proposed for testing hypothesis H, large values of T being evidence against H. We have indicated in Section 2 how the simple bootstrap approximates a probability such as $\Pr(T \leqslant d \mid F)$ by $\Pr(T^* \leqslant d \mid \tilde{F})$. Now a different sampling distribution is required, because the test P value is calculated under the restriction imposed by H. If $\delta(\cdot,\cdot)$ is a distance measure between distributions, and if $\mathscr{F}_H$ is the set of all distributions satisfying H, then the bootstrap data distribution might be taken as

$$\tilde{F}_H \text{ minimizing } \delta(F, \tilde{F}) \text{ for } F \in \mathscr{F}_H.$$

TABLE 3

Error rates of bootstrap 90% confidence intervals for mean θ of χ_1^2, samples of size $n = 20$ (Owen, 1987)

Method	*Proportion of times $\theta <$ lower limit*	*Proportion of times $\theta >$ upper limit*	*Aggregate error rate*
Exact parametric	0.051	0.056	0.107
Bootstrap percentile	0.023	0.150	0.173
Efron's accelerated, bias-corrected bootstrap	0.050	0.105	0.155
Bootstrap Student t	0.038	0.072	0.112

The bootstrap test P value corresponding to observed statistic t_{obs} would be

$$\tilde{P}_H = \Pr\{T_H^* \geqslant t_{\text{obs}} \mid \tilde{F}_H\}, \tag{7}$$

where T_H^* is the test statistic calculated under random sampling from $\tilde{F}_H$.

There are basically two ways to obtain $\tilde{F}_H$ from $\tilde{F}$, one being to change the probabilities at $x_1, \ldots, x_n$ from n^{-1} to $w_1, \ldots, w_n$; the other being to redistribute the probabilities n^{-1} to a wider support than $x_1, \ldots, x_n$. Efron (1982, ch. 10) discusses applications of the former, specifically embedding $\tilde{F}$ in an exponential family; see also Owen (1987).

Uses of modified support are described by Ducharme *et al.* (1985) and by Young (1986). For example, in one of Young's applications, the hypothesis H asserts independence of the two components of $X = (Y, Z)$, and $\tilde{F}_H$ is naturally taken to be the product of the empirical marginal cdfs $\tilde{G}$ and $\tilde{H}$ of Y and Z respectively. The resulting test is therefore very similar to a randomization test, the difference being only that between sampling with and without replacement. The same phenomenon would occur in a two-sample comparison, where a common aggregate distribution would be defined by $\tilde{F}_H$.

A rather striking application of the bootstrap is Silverman's (1981) test for unimodality of a distribution, which uses smooth density estimates as the particular form of probability redistribution. This nicely illustrates the usefulness of bootstrap methods when classical theoretical approaches to calculation of the P value are intractable. Another example is outlined in Section 6.

6. REGRESSION PROBLEMS

Application of bootstrap methods in regression is of potential importance because of the ever-increasing generality of regression methods, for which the classical methods of assessment used in textbook linear regression are inappropriate. Efron (1986) gives a useful introduction, and Wu (1986) with its accompanying discussion refers to much of what is known about properties of the bootstrap in regression analysis. There are two general types of problem, one the assessment of accuracy of regression coefficients or fitted values of mean response, the other being selection of variables or choice of model on the basis of some measure of model fit.

Suppose that we have a particular form of model $y_i = \mu(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i$ connecting continuous responses y_i to explanatory variables $\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})$, with ε_is as random errors. Given some method of fitting the relationship, such as least squares or M estimation, we obtain coefficient estimate $\hat{\boldsymbol{\beta}}$ and fitted values $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$. Inspection of the residuals $\hat{e}_i = y_i - \hat{\mu}_i$, or prior evidence, may suggest that the errors ε_i are homogeneous, with distribution F estimable by the empirical distribution $\tilde{F}$ of residuals. If so, the bootstrap methods discussed earlier extend straightforwardly, simulated data sets *data** taking the form $\{(x_i, y_i^*), i = 1, \ldots, n\}$ with $y_i^* = \hat{\mu}_i + \varepsilon_i^*$, where ε_i^* is randomly sampled from $(\hat{e}_1, \ldots, \hat{e}_n)$. Fitting the model to *data** gives simulated estimate $\hat{\boldsymbol{\beta}}^*$ and fitted values $\hat{\mu}_i^*$. Repeated simulation then leads to required assessments of uncertainty as in earlier sections.

As an example, consider the following significance testing problem. The mean relationship $\mu(x)$ is either linear (hypothesis H) or piecewise linear with two linear segments intersecting at $x = \gamma$. Statistic T is the normal theory likelihood ratio test statistic, whose exact null distribution is intractable even if errors ε are normal.

In the terminology of Section 5, the empirical distribution of residuals $\hat{e}_i$ from the linear regression is $\tilde{F}_H$, and significance probability $\tilde{P}_H$ in (7) is calculated using samples $y_i^* = \hat{\mu}_i + \varepsilon_i^*$ as described above with $\hat{\mu}_i$ the fitted linear regression values. This method was applied to a small set of data from a noise signal experiment in which the $n = 9$ values of x were natural logarithms of 10, 20, 30, 50, 100, 150, 200, 300 and 500 with corresponding values of y being 87.83, 86.50, 84.83, 83.50, 80.17, 79.50, 79.17, 78.67 and 78.67. The estimated point of intersection in the two-segment model is $\hat{\gamma} = 5.1$ and the test statistic is $t = 14.7$. From $B = 1000$ bootstrap samples, $\tilde{P}_H$ was calculated to be approximately 0.02. The null distribution of T, as estimated by the empirical distribution of T_H^*, is not at all close to the χ_2^2 distribution which an (invalid) appeal to classical theory might suggest; see Feder (1975).

One might argue that raw residuals $\hat{e}_i$ should be modified prior to use as simulated errors, e.g. by standardizing to remove the effects of leverage and by adjusting to zero mean. Unpublished numerical evidence supports such modifications. Whether or not one need use complicated modifications such as those described by Cook and Tsai (1985) for non-linear models is unclear.

A more interesting context is that in which errors are not homogeneous, so that a single empirical error distribution is inappropriate. One simple approach is then to consider $(\mathbf{x}_i, y_i)$ as sampled from a joint distribution F, the implication being to sample vectors $(\mathbf{x}_i^*, y_i^*)$ from the data vectors in the bootstrap simulation. There are two drawbacks with this approach. First, it would often be the case that $\text{var}(\varepsilon_i)$ changes smoothly with $\mathbf{x}_i$ or μ_i, and use might be made of this. Secondly, on the general grounds of requiring inference to be conditional on the design $D = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, one should not risk having simulated data sets whose designs $D^* = (\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*)$ are very different from D.

The last point could be dealt with separately either by pre-stratification or post-stratification of the sampling of data vectors, in either case forcing D^* and D to be close in a meaningful sense.

The design difficulty may be moot, of course, if some form of modelling for the errors is used. An example of this in nonparametric regression is given by Efron (1986). A local smoothing algorithm is used first to fit $\hat{\mu}(x)$, and is then applied to squared residuals $\hat{e}_i^2$ to fit a smooth relationship between $\sigma^2 = \text{var}(\varepsilon)$ and x, say $\hat{\sigma}^2(x)$. This permits calculation of homogeneous standardized residuals $r_i = \hat{e}_i / \hat{\sigma}(x_i)$, and thence defines a bootstrap model

$$y_i^* = \hat{\mu}(x_i) + \hat{\sigma}(x_i) r_i^*$$

with the r_i^* randomly sampled from $(r_1, \ldots, r_n)$. Bootstrap samples are then used to obtain confidence bands for $\mu(x)$.

So far we have assumed that responses y are continuous and that errors are additive. How might one apply bootstrap methods to responses which are counts, i.e. non-negative integers, say? One approach is to use the local linearization which GLIM uses for its iterative weighted least squares fitting of generalized linear models. But such an approach offers little more than jackknife methods. If count data are thought of as extended Poisson, that is with variance function $\phi(\mathbf{x})\mu$, then a locally smooth estimate of $\phi(\mathbf{x})$ could be produced and the data could be analysed appropriately in GLIM. More needs to be learned about the possible role of bootstrap methods in such situations.

Special mention should be made of cases where replication exists at every design point. In such cases it would be possible to estimate response distributions F_i at each $\mathbf{x}_i$, and thence bootstrap by sampling from $\tilde{F}_i$ at each $\mathbf{x}_i$. This approach of course applies to multisample problems, unless separate variance components are involved (Section 9). An open question is how well the bootstrap will perform when each of very many $\tilde{F}_i$s is based on few responses. The results of Bickel and Freedman (1982) are probably relevant. There is a very useful series of papers by Freedman and Peters (1984 and references therein) on the performance of bootstrap methods in econometric regression models.

The rather different types of problems typified by model selection, variable selection and prediction assessment are problems to which cross-validation techniques (Stone, 1974) are often applied. A detailed analysis by Efron (1983) shows that cross-validation techniques may be inferior to bootstrap assessments in many cases; see also Bunke and Droge (1984). This important problem will not be discussed here.

7. CONDITIONAL BOOTSTRAP METHODS

In the preceding section the idea of conditioning was mentioned briefly. Conditioning on ancillary statistics is an important general component of statistical inference. As to whether or not relevant conditioning is generally possible in bootstrap methods, the situation is unclear.

A crucial issue may be the nature of the conditioning variable, or ancillary statistic. For example, suppose that $E(T-\theta \mid a,\ F) = b(a-\alpha,\ F)$ with $\alpha = E(A \mid F)$ or with $a = a(\tilde{F})$ and $\alpha = a(F)$. A bootstrap simulation can estimate $b(\cdot, F)$ by $b(\cdot, \tilde{F})$, but this cannot be used without knowing α, at least with error negligible compared to $a-\alpha$. This difficulty seems to preclude conditional bootstrap analysis of the sample mean, for example. The regression application suggested in Section 6 is different in the sense that the effect of an ancillary measure a of the design D does not involve the mean of A.

There is also the difficulty of choosing a in the absence of a model, accompanied by the difficulty of estimating properties conditional on a. For example, in a regression problem with non-homogeneous errors, the precise form of effect of the design D on the variances of coefficients will usually be unknown. However, if the regression fit is approximately linear with weight w_i attached to $(\mathbf{x}_i, y_i)$, and if $\mathrm{var}(y_i \mid \mathbf{x}_i)$ is estimated by $\hat{\sigma}_i^2$, then it would seem appropriate to define a in terms of the elements of $\Sigma w_i^2 \hat{\sigma}_i^2 \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$, by analogy with weighted least squares linear regression. Once a is chosen, the required conditional property would be estimated using discrete partitions of the bootstrap simulation. For example, $\mathrm{var}(\hat{\beta} \mid a)$ could be approximated by a smoothed version of $\mathrm{var}(\hat{\beta}^* \mid a^*)$ evaluated at $a^* = a$.

In some, possibly rare, cases conditional distributions will be amenable to special numerical techniques, such as stratified simulation or conditional saddlepoint approximations. One example of the latter is given by Davison and Hinkley (1988).

It may be worth remarking that in classical statistics the likelihood function itself provides exact or approximate conditional inference (Barndorff-Nielsen, 1983; Cox and Reid, 1987). Quite possibly one might use the bootstrap likelihoods of Section 8 in the same way.

8. BOOTSTRAP PARTIAL LIKELIHOODS

Alchemy failed. But bootstrappers have produced likelihoods, or confidence

distributions. For want of something better, the term partial likelihood may be appropriate.

One direct approach by Hall (1987) is to derive a smooth density estimate from the bootstrap simulation values of the Studentized pivot $Q = (T - \theta)/S$ mentioned in Section 4. Such a partial likelihood has good properties when used to calculate confidence sets, and may show interesting features which standard normal approximations do not. A second approach is via the second-level bootstrap of Section 4, with likelihood evaluations at $\theta = t^*$ being calculated as approximate densities of T^{**} at t^*.

A more classical analogy is pursued by Ogbonmwan and Wynn (1988) for problems involving contrast parameters. Suppose that data $\mathbf{y} = y_1, \ldots, y_n$ are such that, for the correct value of θ, the transformed vector $g(\mathbf{y}, \theta) = g_1(\theta), \ldots, g_n(\theta)$ may be assumed to be a random sample from a fixed distribution function F_0. If $T = t(\mathbf{y})$ is the estimating function for θ, define $T_\theta = t(g(\mathbf{y}, \theta))$ with observed value t_θ. Then a partial likelihood for θ is the density of T_θ at t_θ. The bootstrap version of this definition involves replacing F_0 by the empirical distribution function $\tilde{F}_\theta$ defined by data values $g(\mathbf{y}, \theta)$, and approximating the density of statistic T_θ^* obtained from samples generated by $\tilde{F}_\theta$. In some cases numerical simulation can be avoided, as in the following example, taken from Davison and Hinkley (1988).

Suppose that θ is the difference between means for two populations from which the following two samples were drawn

sample 1:	37.5	34.8	38.9	38.6	37.0	37.4	36.5	38.4	38.0	30.7
sample 2:	37.7	36.3	38.0	37.0	37.6	33.2	36.7	27.4	37.1	37.4

Denote general samples by $(x_1, \ldots, x_m)$ and $(x_{m+1}, \ldots, x_{m+n})$, and suppose that we choose to estimate θ by $t = m^{-1}\Sigma_{m+1}^{m+n} x_i - m^{-1}\Sigma_1^m x_i$. Since the two sample variances are nearly equal, it seems reasonable to take $g(\mathbf{x}, \theta) = (x_1, \ldots, x_m, x_{m+1} - \theta, \ldots, x_{m+n} - \theta)$. If $g_1^*, \ldots, g_{m+n}^*$ denotes a random sample from the elements of $g(\mathbf{x}, \theta)$, then $T_\theta^* = n^{-1}\Sigma_{m+1}^{m+n} g_i^* - m^{-1}\Sigma_1^m g_i^*$. A saddlepoint density approximation can be obtained for T_θ^*, and its evaluation at $t_\theta = t - \theta$ defines the bootstrap partial likelihood. The result is graphed in Fig. 2, together with the normal theory modified profile likelihood.

Note that this type of bootstrap partial likelihood could just as easily be based on any estimate T, although the saddlepoint simplification requires that T be defined by linear estimating equations. The method is very similar to the use of randomization distributions.

A more direct approach is taken by Owen (1987), who considers $\tilde{F}$ to be embedded in a class of distributions $\mathcal{F}_\mathbf{x}$ whose support is $x_1, \ldots, x_n$ in the simple case of homogeneous data. Then if $\theta = t(F)$, the bootstrap likelihood of θ is the profile likelihood under the 'model' $\mathcal{F}_\mathbf{x}$. More concretely, consider $F_\mathbf{w}$ to attach probabilities $w_1, \ldots, w_n$ at points $x_1, \ldots, x_n$; $\tilde{F}$ is the maximum likelihood estimate with $w_i \equiv n^{-1}$, $i = 1, \ldots, n$. Then define the bootstrap likelihood to be

$$BL(\theta) = \sup_{\mathbf{w}: t(F_\mathbf{w}) = \theta} \prod_{i=1}^{n} w_i.$$

Owen (1987) outlines and applies an algorithm for calculating $BL(\theta)$. He also demonstrates that, at least in simple cases, conventional chi-squared asymptotics apply to the log-likelihood ratio.

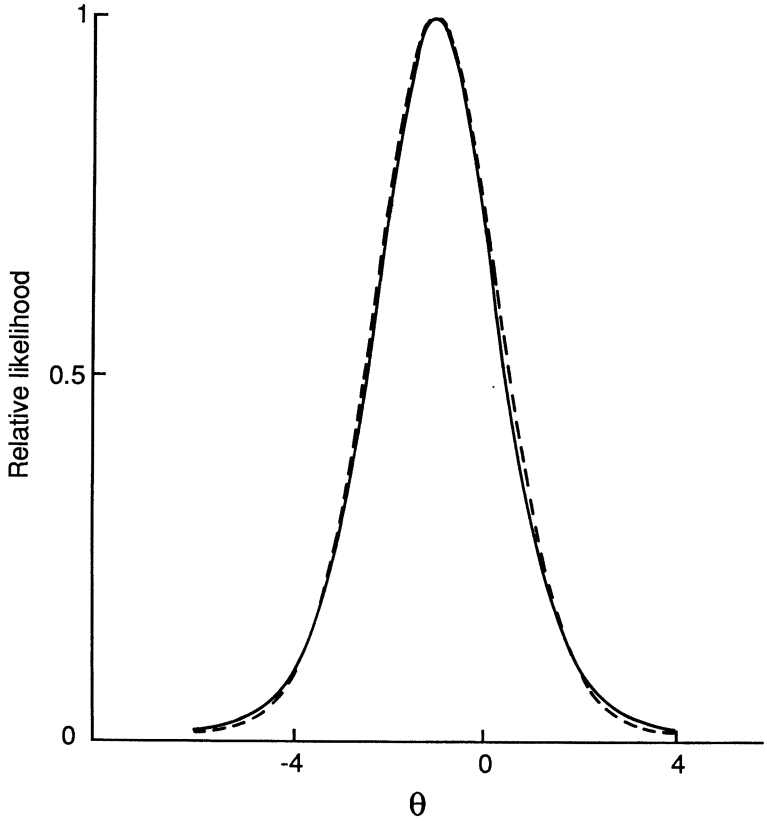

Fig. 2. Relative likelihoods for two-sample contrast parameter θ: the full curve is the saddlepoint approximation to bootstrap likelihood; the broken curve is the normal theory modified profile likelihood

One unusually simple model where an empirical likelihood and resulting conditional analysis are possible is the change-point model. The basic theory and one application are described by Hinkley and Schechtman (1986). Another application is to the mean shift analysis of the series of UK coal-mining disasters (Andrews and Herzberg, 1985, p. 51). The model for count x_i in the ith period of length one year is that $\Pr(X_i = r) = f_0(r)$, $i \leqslant \theta$, and $\Pr(X_i = r) = f_1(r)$, $i > \theta$, successve counts being independent. A nonparametric profile likelihood for θ is therefore

$$PL(t) = \prod_{i=1}^{t} \hat{f}_0(x_i \mid t) \prod_{i=t+1}^{n} \hat{f}_1(x_i \mid t),$$

where

$$\hat{f}_0(r \mid t) = t^{-1} \sum_{i=1}^{t} \delta(x_i - r), \quad \hat{f}_1(r \mid t) = (n-t)^{-1} \sum_{i=t+1}^{n} \delta(x_i - r).$$

Table 4 shows the crucial part of the data series and corresponding values of $PL(t)$ after normalizing to unit sum: the result is an approximate conditional distribution, in that if $\hat{\theta}$ maximizes PL then

$$\Pr(\hat{\theta} - \theta = d \mid a) \propto PL(\hat{\theta} - d).$$

TABLE 4

Part of the annual UK coal-mining disaster frequencies x_t and corresponding normalized bootstrap likelihood PL (t), t = calendar year — 1850

Year t	34 (1884)	35	36	37	38	39	40	41	42	43	44	45	46	47
Frequency x_t	2	3	4	2	1	3	2	2	1	1	1	1	3	0
Normalized $PL(t)$	0.002	.003	.189	.199	.048	.100	.130	.220	.061	.021	.008	.004	.008	.001

The ancillary a here is the set of likelihood ratio increments $PL(\hat{\theta}+k)/PL(\hat{\theta}+k-1)$, most influential being those for small $|k|$. These same increments could be used to partition a bootstrap simulation if a non-likelihood analysis were performed (Hinkley and Schechtman, 1987). Note that bootstrap simulation extends easily to more complicated models, such as first-order Markov processes.

9. OTHER APPLICATIONS

The types of applications mentioned thus far are mostly elementary, save for regression. There is a growing literature on other, more complex applications, some of which are mentioned in this section; see also the general remarks in Section 10.

One traditional area of application for subsampling techniques is the analysis of complex sample surveys. In the usual case where data sampling is without replacement from finite populations, ordinary bootstrapping (done with replacement) may produce inadmissible simulated samples. Partly for this reason, a series of special bootstrap techniques has been proposed in the sample survey literature. Some of the techniques are appraised by McCarthy and Snowden (1985), who give preliminary endorsement to the simple modification of increasing bootstrap sample size from n to $n/(1-f)$, where f is the data sampling fraction.

Problems involving time series, or more generally a stochastic process, raise the difficulty of the single realization. What plays the role of $\tilde{F}$? There are two possible elementary strategies: (i) split the realization into several pieces, and sample from these, or (ii) fit a model with independent innovations, and simulate realizations by adding sampled residuals to fitted values. More sophisticated versions of these strategies will be required for fairly general application.

Perhaps more conventional are problems involving variance components, such as occur in empirical Bayes models. The essential point here is that bootstrap simulation should, implicitly or explicitly, simulate each component of variability. Precisely how will depend on the application. Suppose, for example, that a notional model for the data matrix x_{ij} of p samples is $x_{ij}=\mu_i+\varepsilon_{ij}$, where the μs and εs respectively have distributions G and F. If we are interested in a statistic symmetric in the samples, such as $\bar{x}..$ or $\max_i \bar{x}_{i\cdot}$, then we can simulate data x_{ij}^* by $x_{ij}^*=\mu_i^*+\varepsilon_{ij}^*$, where μ_i^* is randomly sampled from estimates $(\hat{\mu}_1, \ldots, \hat{\mu}_p)$ and ε_{ij}^* are randomly sampled from residuals $\{x_{ij}-\bar{x}_{i\cdot}\}$. The estimates $\hat{\mu}_i$ would be of empirical Bayes type but corrected to have appropriate mean and variance, e.g. $\bar{x}..$ and the unbiased estimate of $\mathrm{var}(\mu)$. Such a simulation would not be appropriate if, say, we were interested in mean μ_1 *a priori*, for then one simulated sample should use $\mu^*=\hat{\mu}_1$. Further discussion of these kinds of applications will be found in Hill (1986) and Laird and Louis (1987).

10. GENERAL REMARKS

One might observe that bootstrap methods essentially embrace, or enlarge upon, familiar methods of simulation, subsampling and permutation. What is new is the generality of approach, the range of potential applications and the massive use of computer power.

It would be presumptive to dismiss the many simple applications because of existing classical methods: such applications are mere scale exercises, which help to tune the instruments and their players in the bootstrap orchestra so that they will perform better in the complex pieces of modern data analysis. Thus, for example, bootstrap methods may prove to be uniquely reliable tools for analysing nonparametric curve fits, complex pure significance test problems and nonstationary time series models. At the very least bootstrap methods provide a simple approach to assessment of the sensitivity of traditional methods to model assumptions. This thought also suggests the possible use of simulated samples to generate diagnostics, akin to the more usual case deletion diagnostics.

Because bootstrap methods also apply in the arena of model assessment, they are pertinent to the larger, often neglected area of decision analysis under model uncertainty.

In the previous sections we have not commented on the non-negligible tendency for misapplication of bootstrap methods, in particular the misuse of simple random sampling from data sets. There is a very clear need to bring classical statistical theory to bear in the development of reliable methodology, as evidenced by the importance of pivots in confidence limit methods. In this context one should also consider Bayesian approaches to bootstrapping, which involve Dirichlet models; see Rubin (1981) and Banks (1987).

REFERENCES

Andrews, D. F. and Herzberg, A. M. (1985) *Data: a Collection of Many Problems from Many Fields for the Student and Research Worker*. New York: Springer.

Banks, D. L. (1987) Improving the Bayesian bootstrap. Unpublished.

Barndorff-Nielsen, O. (1983) On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 345–365.

Beran, R. J. (1982) Estimated sampling distributions: the bootstrap and its competitors. *Ann. Statist.*, **10**, 212–225.

———(1984) Bootstrap methods in statistics. *Jber. Dtsch. Math-Ver.*, **86**, 14–30.

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

Bickel, P. J. and Freedman, D. A. (1982) Bootstrapping regression models with many parameters. Unpublished, University of California at Berkeley.

Bunke, O. and Droge, B. (1984) Bootstrap and cross-validation estimates of the prediction error for linear regression models. *Ann. Statist.*, **12**, 1400–1424.

Chapman, P. L. and Hinkley, D. V. (1986) The double bootstrap, pivots and confidence limits. *Report 26*. Center for Statistical Sciences, University of Texas at Austin.

Cook, R. D. and Tsai, C. L. (1985) Residuals in nonlinear regression. *Biometrika*, **72**, 23–29.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*. London: Chapman and Hall.

Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional likelihood. *J. R. Statist. Soc.* B, **49**, 1–39.

Daniels, H. E. (1987) Tail probability approximations. *Int. Statist. Rev.*, **55**, 37–48.

Davison, A. C. and Hinkley, D. V. (1988) Saddlepoint approximations in resampling methods. *Biometrika*, **75** (to appear).

Davison, A. C., Hinkley, D. V. and Schechtman, E. (1987) Efficient bootstrap simulation. *Biometrika*, **74**, 555–566.

DiCiccio, T. J. and Romano, J. P. (1988) A review of bootstrap confidence intervals. *J. R. Statist. Soc.* B, **50**, 338–354.

Ducharme, G. R., Jhun, M., Romano, J. P. and Truong, K. N. (1985) Bootstrap confidence cones for directional data. *Biometrika*, **72**, 637–645.

Efron, B. (1982) The jackknife, the bootstrap and other resampling plans. In *Regional Conference Series in Applied Mathematics*, No. 38. Philadephia: SIAM.

———(1983) Estimating the error rate of a prediction rule: improvements in cross-validation. *J. Amer. Statist. Ass.*, **78**, 316–331.
———(1986) Computer-intensive methods in statistical regression. Unpublished, Department of Statistics, Stanford University.
———(1987) Better bootstrap confidence intervals. *J. Amer. Statist. Ass.*, **82**, 171–200.
Feder, P. I. (1975) On asymptotic distribution theory in segmented regression problems: identified case. *Ann. Statist.*, **3**, 49–83.
Freedman, D. A. and Peters, S. C. (1983) Bootstrapping a regression equation: some empirical results.
Graham, R. L., Hinkley, D. V., John, P. W. M. and Shi, S. (1987) Balanced design of bootstrap simulations. *Report 48*. Center for Statistical Sciences, University of Texas at Austin.
Hall, P. (1987) On the bootstrap and likelihood-based confidence regions. *Biometrika*, **74**, 481–493.
———(1988) Theoretical comparison of bootstrap confidence intervals. *Ann. Statist.*, to be published.
Hill, J. R. (1986) Empirical Bayes statistics: a comprehensive theory for data analysis. *PhD Thesis*. Department of Mathematics, University of Texas at Austin.
Hinkley, D. V. and Schechtman, E. (1987) Conditional bootstrap methods in the mean-shift model. *Biometrika*, **74**, 85–93.
Laird, N. M. and Louis, T. A. (1987) Empirical Bayes confidence intervals based on bootstrap samples. *J. Amer. Statist. Ass.*, **82**, 739–750.
McCarthy, P. J. and Snowden, C. B. (1985) The bootstrap and finite population sampling. In *Vital and Health Statistics*, Series 2, No. 95. Washington DC: Public Health Service.
Miller, R. G., Jr (1974) The jackknife: a review. *Biometrika*, **61**, 1–17.
Ogbonmwan, S. M. and Wynn, H. P. (1986) Accelerated resampling codes with low discrepancy. Unpublished, Department of Statistics, Imperial College.
———(1988) Resampling generated likelihoods. In *Statistical Decision Theory and Related Topics IV* (eds S. S. Gupta and J. O. Berger), vol. 1, pp. 133–147. New York: Springer.
Owen, A. B. (1987) Empirical likelihood ratio confidence intervals for a single functional. Unpublished, Department of Statistics, Stanford University.
Rubin, D. B. (1981) The Bayesian bootstrap. *Ann. Statist.*, **9**, 130–134.
Silverman, B. W. (1981) Using kernel density estimates to investigate multimodality: *J. R. Statist. Soc.* B, **43**, 97–99.
Silverman, B. W. and Young, A. (1987) The bootstrap: to smooth or not to smooth? *Biometrika*, **74**, 469–479.
Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions. *J. R. Statist. Soc.* B, **36**, 111–147.
Thernau, T. (1983) Variance reduction techniques for the bootstrap. *PhD Thesis*. Department of Statistics, Stanford University.
Tibshirani, R. (1987) Variance stabilization and the bootstrap. Unpublished, Department of Statistics, University of Toronto.
Wu, C. F. J. (1986) Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, **14**, 1261–1350.
Young, A. (1986) Conditioned data-based simulations: some examples from geometrical statistics. *Int. Statist. Rev.*, **54**, 1–13.