Resenhas IME-USP 1993, Vol. 1, No. 2/3, 283 - 304.

Applying the Bootstrap: An Example

A. C. Davison and D. V. Hinkley

Abstract We illustrate bootstrap methods in a simple example. Among ideas discussed are: basic distributional approximations; confidence limit methods; improved calculation; tests; bootstrap sensitivity analysis; regression; and nonparametric likelihood inference.

Key words: ABC method; Balanced resampling; Bootstrap; Bootstrap hypothesis test; Bootstrap t; Empirical likelihood; Infinitesimal jackknife; Influence function; Jackknife-after-bootstrap; Linear approximation; Percentile method; Permutation test; Ratio; Regression; Saddlepoint approximation; Textile data.

1 Introduction

Bootstrap methods are simulation procedures for assessing the variability of estimators and for performing relatively model-free tests. They have developed greatly from their introduction by Efron (1979) to the recent books by Hall (1992) and Efron and Tibshirani (1993). The purpose of this paper is to illustrate some of these methods in a simple example.

Table 1 gives $n = 32$ observations on the number of faults, y, in lengths of cloth, x. The data, displayed in Figure 1, show an increase in faults with length. In this paper we focus on the mean number of faults per unit length, θ. One central conern is how to form confidence intervals for θ based on the ratio estimator, but we also address questions such as whether the ratio is a good choice for these data, and whether a model in which $\mathrm{E}(y) = \theta x$ is appropriate, or whether a nonparametric curve would be more suitable.

Section 2 describes the use of simulation in analysis of these data, and discusses how the basic sampling plan may be adapted to more complex situations. The choice of estimator from the data is discussed in §3, and §4 describes how the simulation may be made more efficient. Confidence limits and tests are discussed in §§5 and 6. Sensitivity analysis for our methods is described in §7, and in §8 we briefly outline how nonparametric likelihoods can play a role in data analysis.

2 Some Basic Ideas

2.1 Simulation

In practical problems many estimators are assumed to be approximately normally distributed. If so, inference for the parameter of interest, θ, could be based on this distribution, with estimated mean and variance. For example, if an estimator T

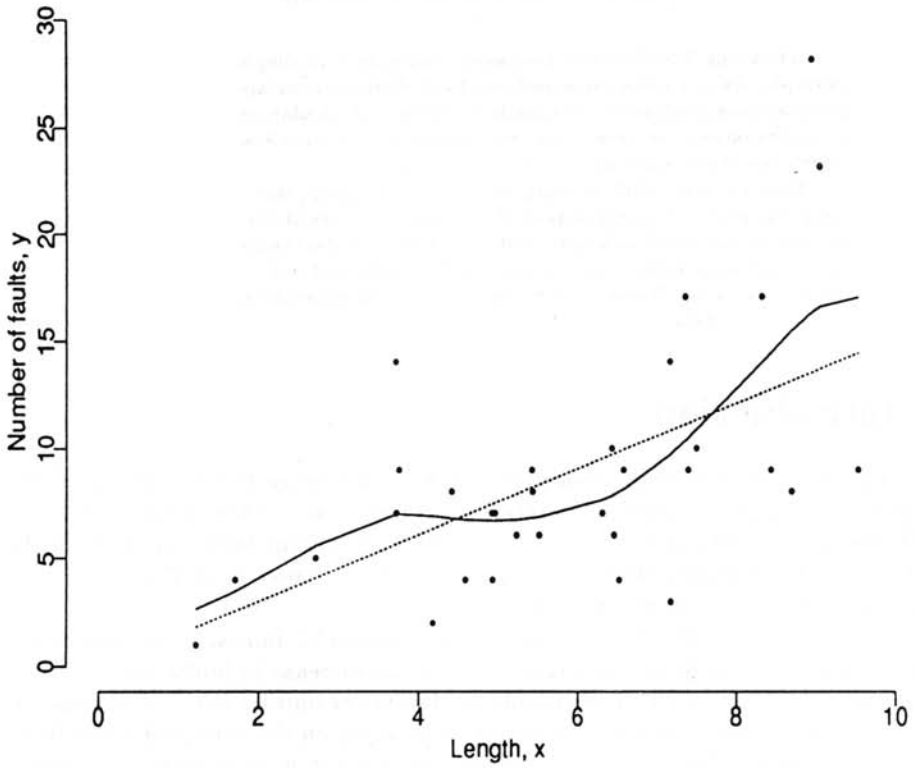

Figure 1: Numbers of faults in 32 lengths of cloth (from Bissell, 1972). *The dotted line is for a fitted Poisson model, and the solid line for a generalized additive model.*

is normally distributed with mean $\theta + B$ and variance V, we can base confidence intervals for θ on the approximate standard normal distribution of

$$Z = \frac{T - (\theta + B)}{V^{1/2}},$$

giving approximate $(1 - 2\alpha) \times 100\%$ confidence interval

$$(t - B - V^{1/2} z_{1-\alpha}, t - B - V^{1/2} z_{\alpha}), \tag{2.1}$$

where t is the observed value of T and z_α is the α quantile of the standard normal distribution. However (2.1) involves the unknown bias and variance of T and cannot be used as it is, and furthermore the normal approximation for Z may be poor.

One approach to estimation of B and V is to use the data themselves as a working model, estimating F by the empirical distribution function $\hat{F}$, where $\hat{F}$

puts mass n^{-1} on each observed pair (x_j, y_j). The true parameter and the bias and variance of T,

$$\theta = \frac{\int y dF(x,y)}{\int x dF(x,y)}, \quad B = \mathrm{E}(T \mid F) - \theta, \quad V = \mathrm{var}(T \mid F),$$

would then be estimated by

$$t = \frac{\int y d\hat{F}(x,y)}{\int x d\hat{F}(x,y)} = \frac{n^{-1}\sum y_j}{n^{-1}\sum x_j}, \quad \hat{B} = \mathrm{E}(T \mid \hat{F}) - t, \quad \hat{V} = \mathrm{var}(T \mid \hat{F}).$$

The estimates $\hat{B}$ and $\hat{V}$ would be obtained analytically if possible, and otherwise by simulation.

The jackknife can provide simple analytic estimates of bias and variance for T, using the influence function or infinitesimal jackknife values for the statistic. These are

$$L_j = \lim_{\epsilon \to 0} \frac{t\{(1-\epsilon)\hat{F} + \epsilon H_j\} - t(\hat{F})}{\epsilon},$$

where H_j represents the distribution function putting unit mass on (x_j, y_j). Once the L_j have been obtained, by numerical differentiation if need be, simple bias and variance formulae are available: $\hat{B}_L = 0$, and $\hat{V}_L = n^{-2}\sum L_j^2$.

In order to approximate $\hat{B}$ and $\hat{V}$ by simulation, we generate R replicate synthetic samples, by sampling uniformly at random with replacement from the data and calculating the corresponding values of T, labelled T^*. Since our model is that the data are a random sample of pairs, an observation from the working model is a pair $(y_j^*, x_j^*) = (y_{j^*}, x_{j^*})$, where j^* is randomly selected from $\{1, \ldots, n\}$, and the rth replicate sample and simulated value of T^* are $(x_{1r}^*, y_{1r}^*), \ldots, (x_{nr}^*, y_{nr}^*)$ and $t_r^* = \sum_j y_{jr}^* / \sum_j x_{jr}^*$. Monte Carlo approximations to $\hat{B}$ and $\hat{V}$ based on $t_1^*, \ldots, t_R^*$ are

$$B^* = R^{-1}\sum_{r=1}^{R} t_r^* - t = t^* - t, \quad V^* = (R-1)^{-1}\sum_{r=1}^{R}(t_r^* - \bar{t}^*)^2. \tag{2.2}$$

Table 2 shows the calculations for a very small simulation with $R = 9$. Each row gives the frequencies with which the original pairs occur in a simulated sample. Here $B^* = -0.002$ and $V^* = 0.0268$, so the estimated 95% confidence interval from supposed normality of $Z^* = \{T^* - (t + \hat{B})\}/\hat{V}^{1/2}$ is obtained by replacing $\hat{B}$ and $\hat{V}$ with estimates B^* and V^*, to give

$$(t - B^* - 1.96V^{*1/2}, t - B^* + 1.96V^{*1/2}) = (1.19, 1.83).$$

The accuracy of B^* and V^* as approximations to $\hat{B}$ and $\hat{V}$ depends on R, and in practice we would of course choose $R \gg 9$.

j	1	2	3	4	5	6	7	8	9	10	11
x	1.22	1.70	2.71	3.71	3.72	3.75	4.17	4.41	4.58	4.91	4.92
y	1	4	5	14	7	9	2	8	4	7	4
j	12	13	14	15	16	17	18	19	20	21	22
x	4.95	5.22	5.42	5.43	5.51	6.30	6.42	6.45	6.51	6.57	7.15
y	7	6	9	8	6	7	10	6	4	9	14
j	23	24	25	26	27	28	29	30	31	32	
x	7.16	7.35	7.38	7.49	8.32	8.42	8.68	8.95	9.05	9.52	
y	3	17	9	10	17	9	8	28	23	9	

Table 1: The numbers of faults (y) in lengths (x) (metres $\times 10^2$) of cloth. From Bissell (1972).

j	1	2	3	4	$\cdots$	29	30	31	32	
x	1.22	1.70	2.71	3.71	$\cdots$	8.68	8.95	9.05	9.52	
y	1	4	5	14	$\cdots$	8	28	23	9	
	Numbers of times each pair sampled									Statistic
	1	1	1	1	$\cdots$	1	1	1	1	$t = 1.510$
Replicate r										
1	1	2	1	0	$\cdots$	0	2	0	1	$t_1 = 1.358$
2	3	0	2	0	$\cdots$	1	0	1	3	$t_2 = 1.372$
3	1	0	1	0	$\cdots$	0	0	0	0	$t_3 = 1.394$
4	0	1	2	0	$\cdots$	3	1	0	2	$t_4 = 1.449$
5	2	1	1	0	$\cdots$	1	2	1	3	$t_5 = 1.460$
6	1	1	1	0	$\cdots$	2	2	0	0	$t_6 = 1.521$
7	1	1	1	1	$\cdots$	0	3	4	1	$t_7 = 1.812$
8	1	3	1	1	$\cdots$	1	0	0	3	$t_8 = 1.456$
9	0	0	2	3	$\cdots$	2	2	2	1	$t_9 = 1.752$

Table 2: $R = 9$ replicate samples for the cloth data.

If we were unhappy with the normal approximation to T^*, the empirical quantiles of the t^* can be used to estimate the true quantiles of T, as described in §5.

An advantage of this approach over purely analytical calculations is that we can inspect the simulation output to diagnose difficulties with an analysis, and to suggest alternatives. For example, the top left panel of Figure 2 shows a histogram of 1000 simulated ratios, overlaid with a density for t^* obtained by a saddlepoint approximation: the distribution seems very close to normal. This is confirmed by the rankit plot in the middle left panel, which overlies the null line except in the extreme tails. The top right panel shows the simulated t^* plotted against the corresponding simulated averages $\bar{x}^*$, while the vertical dotted line shows the value of $\bar{x}$ for the data. The other panels are discussed below.

2.2 General comments

Among questions arising from the approach to data analysis used in §2.1 are:

(a) what statistic T should be used to estimate θ?

(b) which estimate $\hat{F}$ of F should be used?

(c) how do we calculate properties of T^* given $\hat{F}$?

In choosing T, we should consider its bias and efficiency under plausible models, its robustness to departures from those models, its resistance to small changes in the data, ease of calculation, and its pivotality under plausible model(s), where a pivot is a function $Q(T;\theta)$ whose distribution does not depend (much) on F. Some of these can be checked from the replicate samples themselves.

The commonest possibility for (b) is a parametric model, often with parameters estimated by maximum likelihood. Another possibility is a nonparametric model like the one in Section 2.1, based on the empirical distribution function $\hat{F}$. However other estimates such as smoothed versions of $\hat{F}$ might be considered for certain problems (Silverman and Young, 1987; De Angelis and Young, 1992).

Once an estimator T and estimate of F have been chosen, the final choice to be made is how to calculate the properties of interest. We have already mentioned simulation from $\hat{F}$, known as the *nonparametric bootstrap*, but we could also simulate from a fitted parametric model, a procedure sometimes called a *parametric bootstrap*. If the statistic is sufficiently simple, analytic calculation of some of its properties may be feasible, but this is rare, and usually simulation must be used.

2.3 Extensions

The general idea of bootstrapping is to approximate the required property of t by the empirical property of t^*'s calculated from simulated data sets. Here "property of t" can mean things such as bias, variance or cumulative probabilities, and hence percentiles, all of which can be expressed in terms of averages of a function h of

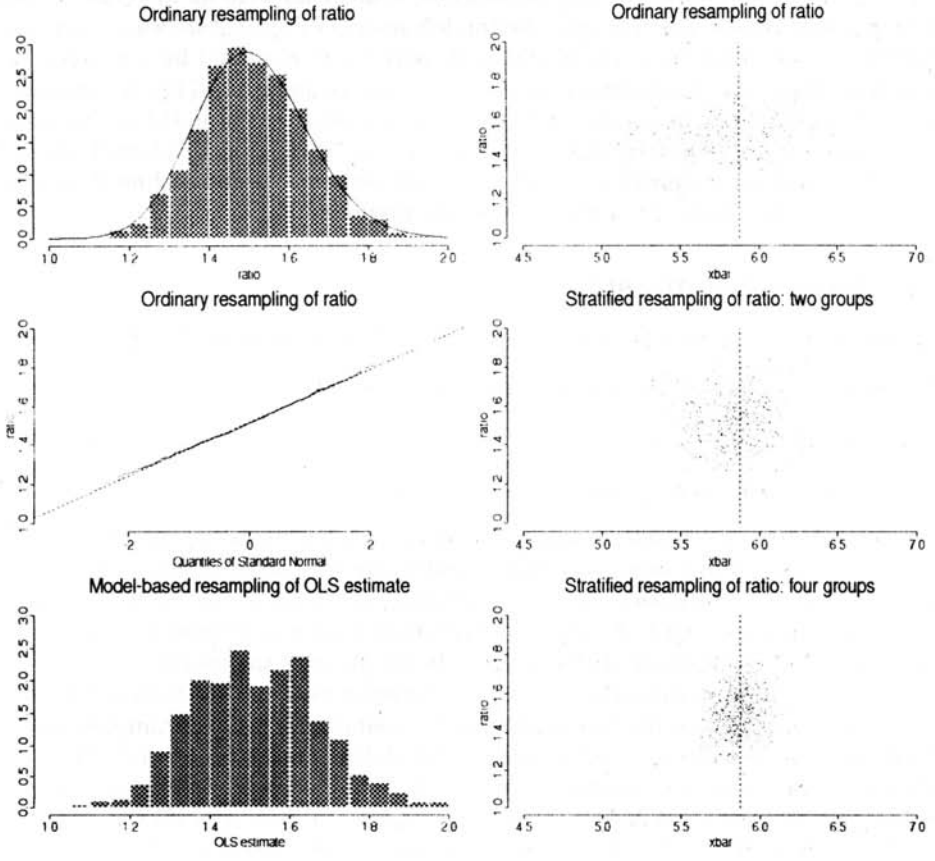

Figure 2: Bootstrap results for cloth data. *The top left panel shows a histogram of 1000 simulated ratios, t^*, overlaid with a saddlepoint approximation to their density. The middle left panel is a rankit plot of the resampled ratios t^*. The bottom right panel is a histogram of resampled ordinary least squares estimates. The right panels show resampled ratios t^* plotted against $\bar{x}^*$ for the original sampling scheme and two in which the pairs are stratified by the values of x_j.*

$T - \theta$. For bias and variance we have (2.1), and for the cumulative probability $\text{pr}(T - \theta \leq c)$ we take h to be the indicator of the event $T - \theta \leq c$.

The general approach that underlies the simulation described in §2.1 is :

1. identify the underlying distribution(s) F for data;
2. define the quantity of interest as a function of F, $\theta \equiv t(F)$;
3. define an estimate $\hat{F}$ for F based on data $\mathcal{D}$;
4. choose $t \equiv t(\hat{F})$ as an estimate of $\theta \equiv t(F)$;
5. identify t as the value of θ for $\hat{F}$, the simulation model;
6. define T^* as the estimator under sampling from $\hat{F}$;
7. identify $T^* - t$ as an approximation to $T - \theta$;
8. repeat R times:
   - generate a sample $\mathcal{D}^*$ by sampling with replacement from $\hat{F}$;
   - calculate the value of t^* by applying the t-algorithm to $\mathcal{D}^*$;
9. approximate $\mathrm{E}\{h(T - \theta)\}$ by $R^{-1} \sum_{r=1}^{R} h(t_r^* - t)$.

Embellishments on this procedure deal with non-iid structure of data, as with several samples, regression or time series; correction for error in Step 7 (by bootstrapping the bootstrapping); replacement of actual simulation of samples (by small-sample asymptotics); more efficient versions of Steps 8 and 9 (using Monte Carlo tricks); and refinements for special objectives, such as confidence sets and significance tests

To illustrate how the algorithm might be changed, suppose that it was desirable to construct resamples with roughly the same value of $\bar{x}^*$ as the original data. This would be sensible if it was thought necessary to condition the analysis on the x values. One procedure for generating simulations in which $\bar{x}^*$ was less variable would be to divide the data into strata depending on the values of the x_j, and to resample within the strata. The bottom right panels of Figure 2 show the effect of this for two and four strata.

One model for the data is that $\text{var}(y_j) = \sigma^2 \theta x_j$, that is, a model with a linear variance function but response overdispersed relative to the Poisson distribution if $\sigma^2 > 1$, as the data seem to indicate. If this model was judged sensible on the basis of residual plots, it would be better to accomodate the changing variance of y_j and the conditioning by using a different resampling plan. In the regression formulation

$$y_j = \theta x_j + \sigma \theta^{1/2} x_j^{1/2} \varepsilon_j,$$

where the ε_j are independent errors with mean zero and unit variance, $\hat{F}$ would represent the empirical distribution function of the residuals $e_j = (y_j - t x_j)/x_j^{1/2}$.

This would be bootstrapped by resampling $\varepsilon_1^*, \ldots, \varepsilon_n^*$ at random from $e_1, \ldots, e_n$, forming new responses $y_j^* = tx_j + x_j^{1/2}\varepsilon_j^*$, and constructing the weighted least squares estimate $b^* = \sum y_j^* / \sum x_j$. Notice that the statistic is still the ratio, which is the appropriate weighted least squares estimate under this model, but the resampling scheme is different. Possible refinements include making a leverage adjustment to the residuals, and rounding the y^* to the nearest non-negative integer, because the y_j are counts.

3 Choice of estimator from the data

In any practical situation there may be uncertainty about which estimator to use, especially if the analysis is to be robust or model-free. To take a simple example, if a robust alternative the average is to be used, then one might want to use a trimmed mean but not know what percentage of the data to trim. In such problems a bootstrap analysis can be very effective, by providing estimates of variation for each of the possible estimators — this is illustrated for the trimmed mean by Efron (1992).

In the regression context, we can use bootstrap methods to help choose among a variety of estimates. For example, with the cloth data we can easily compare an ordinary least squares estimate of slope $b = \sum x_i y_i / \sum x_i^2$ with the ratio estimate discussed earlier. The same simulated data sets used to estimate variation of the ratio can be used to estimate variation of b: we just add computation of bootstrap sample estimate b^* to the previous simulation algorithm. The bottom left panel of Figure 2 shows the histogram from 1000 such simulations, when random sampling of data pairs was used. Evidently the ratio is less variable. The respective bootstrap standard errors for t and b are 0.136 and 0.161.

In the context of robust estimation, i.e. estimation which is resistant to outliers, the influence of individual cases on any particular estimator can be studied by the "jackknife-after-bootstrap" method outlined in §7. In this way choice of estimator can be combined with data-screening.

Sometimes the choice of estimator is based on some form of test, as with regression curves. Here again, as we illustrate in §6, one can use bootstrap methods to advantage when more traditional methods are either unsafe or unavailable.

A danger with data-driven selection of estimator is that a bias is induced in the estimated variation of the chosen estimator: the minimum of several estimated standard errors is biased downward. The danger is usually not serious. For discussion see Leger and Romano (1990) and Efron (1992).

4 Improved calculation

Numerous methods have been suggested to improve the efficiency of the nonparametric simulation scheme outlined in §2.1. Unfortunately the easiest to implement

also usually give the smallest gains in efficiency, and the more powerful general-purpose methods must be used with care lest they make matters worse.

A powerful approach for statistics that can be written as smooth functions of averages or solutions to estimating equations rests on saddlepoint techniques (Reid, 1988). These remove all Monte Carlo error and can give highly accurate approximations to densities and probabilities (Davison and Hinkley, 1988; Daniels and Young, 1991; DiCiccio, Martin and Young, 1992). The saddlepoint approximation to the density of $\bar{Y}$ when $Y_1, \ldots, Y_n$ is a random sample from a distribution function F with cumulant generating function $K(u) = \log \mathrm{E}(e^{uY})$ is

$$f_s(\bar{y}) = \left\{ \frac{n}{2\pi K''(u)} \right\}^{1/2} \exp\left[n\{K(u) - u\bar{y}\} \right],$$

where u solves the saddlepoint equation $K'(u) = \bar{y}$. There is a related expression, $F_s(\bar{y})$, for the cumulative distribution function of $\bar{Y}$. The key to using this in nonparametric simulation is to note that if $Y_1^*, \ldots, Y_n^*$ is a random sample from the empirical distribution function F, the cumulant generating function of Y_j^* is

$$K(u) = \log\left(n^{-1} \sum_{j=1}^{n} e^{uy_j} \right),$$

to which we may apply the formulae for $f_s(\bar{y})$ and $F_s(\bar{y})$. The top left panel of Figure 2 shows the saddlepoint approximation to the density of T^* for the cloth data, which is obtained by noting that $T^* \le t$ if and only if $\sum(y_{j\bullet} - t x_{j\bullet}) \le 0$, and then applying the ideas above to this sum (Daniels, 1983). These methods can be factors of 50–100 times faster than simulation in certain problems (DiCiccio, Martin, and Young, 1992; Hinkley and Shi, 1989), but they can be complicated to implement.

A number of methods are based on linear approximations to the statistic of interest. Suppose that T^* admits an expansion

$$T^* = t + n^{-1} \sum_{j=1}^{n} f_j^* L_j + \cdots, \tag{4.1}$$

where f_j^* is the frequency with which the jth observation appears in a simulated sample. The linear approximation to T^*, T_L^*, is the first two terms on the right of (4.1). Control variate methods rest on the correlation between T^* and T_L^*, aiming to use T_L^* as a proxy for T^* in such a way that only the difference between the two statistics needs to be estimated by simulation. If T^* and T_L^* are highly correlated, the result can be highly accurate estimates of moments and quantiles of T^* (Davison, Hinkley and Schechtman, 1986; Efron, 1990; Do and Hall, 1992). For our problem the correlation between the ratio and its linear approximation is 0.99, so we would expect these methods to work very well.

Other approaches to efficient simulation can be based on importance sampling (Johns, 1988; Booth, Hall and Wood, 1993; Do and Hall, 1991), balanced resampling (Davison *et al.*, 1986; Graham, Hinkley, John and Shi, 1990; Gleason, 1988) or antithetic resampling (Hall, 1989). Appendix II of Hall (1992) contains a theoretical comparison of efficient simulation methods, which are also discussed in Efron and Tibshirani (1993, Chapter 23).

5 Confidence limits

As bootstrap methodology has developed, a variety of confidence limit techniques has surfaced. Here we review the main techniques, and apply them to our ratio estimation problem.

5.1 Basic confidence limit method

The simplest method is essentially to put an interval or region around the point estimate. So for a scalar parameter θ, estimated by T, define a_p to be the p quantile for $T-\theta$: $\mathrm{pr}(T-\theta \le a_p) = p$. Then the exact upper confidence limit with coefficient $1-\alpha$ is $t - a_\alpha$, which is based on the formula

$$\mathrm{pr}(T-\theta \ge a_\alpha) = 1-\alpha.$$

The corresponding equi-tailed $1-2\alpha$ interval is $(t-a_{1-\alpha}, t-a_\alpha)$ Usually the exact a_p's are unknown and must be estimated. While often a normal approximation can be used to do this, the bootstrap offers a more accurate approximation.

A bootstrap estimate for the $1-2\alpha$ interval is obtained from the resampling distribution of T^*-t. If the R t^* values have already been ordered, then $t^*_{pR} - t$ estimates a_p. So the bootstrap confidence interval is

$$t - (t^*_{(1-\alpha)R} - t) = 2t - t^*_{(1-\alpha)R}, \quad t - (t^*_{\alpha R} - t) = 2t - t^*_{\alpha R},$$

with approximate error rates α at each end.

This interval will not have exact coverage $1-2\alpha$, because the distributions of $T-\theta$ and T^*-t are not exactly the same. There are four ways to correct the basic formula for greater accuracy: use of transformation (i.e. calculation of limits on a transformed scale known to produce correct coverage, followed by untransformation); use of Edgeworth series, explicit or implicit; use of studentization; and bootstrap estimate of coverage bias.

In our ratio example, with $R = 1000$ simulations, the following ordered values of the bootstrap ratio were obtained:

r	5	25	50	950	975	995
t^*_r	1.178	1.252	1.285	1.736	1.779	1.893

The sample ratio was 1.510. So, for example, the 95% bootstrap interval for the population ratio has limits $2 \times 1.510 - 1.779 = 1.24$ and $2 \times 1.510 - 1.252 = 1.77$.

5.2 Bootstrap t

If we follow the analogy with Student's t-statistic, then a more stable quantity should be the studentized version of $t - \theta$, i.e.

$$Z = (T - \theta)/SE(T)$$

where $SE(T)$ is the estimated standard error of T, i.e. calculated assuming that the CDF is F. One simple standard error formula is the *infinitesimal jackknife* formula $SE^2 = n^{-2} \sum_{j=1}^{n} L_j^2$, with L_j the empirical influence function for case j. The α quantile for Z is estimated by the αRth ordered value of $z^* = (t^* - t)/SE^*$ in the same R simulations producing the t^*'s. Then, assuming the z^* values are already ordered, the $1 - 2\alpha$ confidence interval has limits

$$t - z^*_{(1-\alpha)R} \times SE, \quad t - z^*_{\alpha R} \times SE;$$

these are analogous to the Student-t limits for the Normal mean in classical statistics.

If the infinitesimal jackknife is used, we can approximate the bootstrapped version SE^{*2} by $n^{-2} \sum_{j=1}^{n} L_{j*}^2$.

Although in principle the bootstrap t method is more accurate than the basic method, the denominator may be somewhat unstable: variance estimates often are. For further theoretical discussion see Hall (1992).

For our data, the bootstrap distribution of Z from $R = 1000$ samples is quite skewed. Figure 3 shows a normal quantile plot of the z^* values, and a scatter plot of t^* versus SE^*, the latter suggesting possible instability of the method.

The following ordered values of z^* were obtained:

r	5	25	50	950	975	995
z^*_r	−3.669	−2.715	−2.157	1.429	1.719	2.364

The infinitesimal jackknife standard error for the ratio is $SE = 0.1398$. So, for example, the 95% confidence interval has limits $1.510 - 0.1398 \times 1.719 = 1.13$ and $1.510 - 0.1398 \times (-2.715) = 1.75$. These differ considerably from those computed with the basic method.

5.3 Percentile method

In some cases, a transformation of the estimate T may be nearly or exactly symmetric in distribution, perhaps even Normal with constant variance. Then the

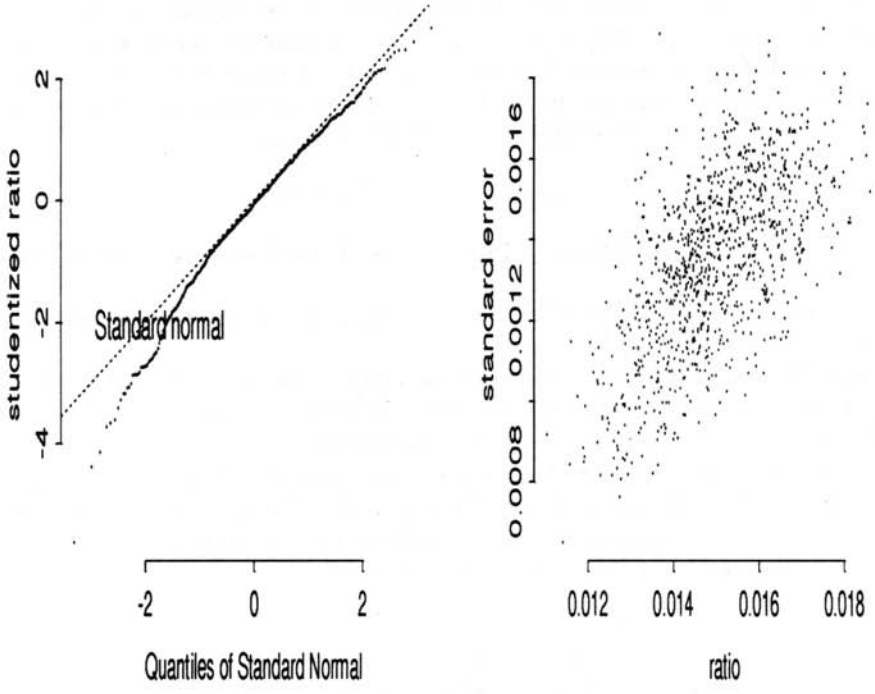

Figure 3: Normal quantile plot of studentized ratio estimates and scatter plot of ratio versus estimated standard error from R=1000 bootstrap samples.

percentile method is applicable, even if the transformation itself is unknown. For if the transformation $g(T)$ of T has symmetric distribution, then on the transformed scale $a_\alpha = -a_{1-\alpha}$. Therefore the upper $1-\alpha$ limit for θ in the basic method can be re-expressed as

$$g^{-1}\{g(t) - a_\alpha\} = g^{-1}\{g(t) + a_{1-\alpha}\},$$

which is estimated by

$$g^{-1}\left\{g(t) + g(t^*_{(1-\alpha)R}) - g(t)\right\} = g^{-1}\{g(t^*_{(1-\alpha)R}\} = t^*_{(1-\alpha)R},$$

which does not depend upon $g(.)$ at all. This upper limit is called the percentile upper limit. The corresponding $1-2\alpha$ confidence interval has limits $(t^*_{\alpha R}, t^*_{(1-\alpha)R})$.

The percentile method is slightly simpler than the basic method, and in a limited number of cases will work better. But the assumed symmetry ignores bias. And as with the basic method its accuracy requires that $T^* - t$ and $T - \theta$ have the same distribution, in particular the same variance. So in general practice the percentile method is not good. However a reliable corrected version of the method is available, as described next.

For the ratio data, some relevant ordered values were given above. From these we calculate, for example, that the 95% interval has limits 1.25 and 1.78. These are very similar to the limits calculated from the basic method.

5.4 Adjusted percentile method

The basis for the corrected method is a refined Normal approximation which incorporates bias and nonconstant variance. The derivation is described by Efron and Tibshirani (1993, Chapter 22). Here we outline the method for the nonparametric case.

Two constants are used for the adjustment, defined as

$$a = \tfrac{1}{6}\frac{\sum_{j=1}^n L_j^3}{\left(\sum_{j=1}^n L_j^2\right)^{3/2}}$$

and $b = \Phi^{-1}\{\hat{G}(t)\}$, where $\hat{G}$ is the empirical distribution of bootstrap statistic values t^* and Φ is the standard normal integral. These constants correspond to heteroscedasticity and bias adjustment factors. The upper α percentile limit $t^*_{\alpha R}$ is then replaced as follows: calculate $z = \Phi^{-1}(\alpha)$, next

$$z_{adj} = b + \frac{(b+z)}{1 - a(b+z)},$$

and then $\alpha_{adj} = \Phi(z_{adj})$. The adjusted upper α limit is $t^*_{\alpha_{adj}R}$.

For the ratio data, we calculate $a = 0.0326$ and $b = \Phi^{-1}(0.532) = 0.080$, the latter since 532 of the 1000 t^* values are below $t = 1.51$. So for the 95% interval,

which combines the 2.5% and 97.5% limits, we first calculate the adjusted values α_{adj} =0.0454 and 0.9883. Then the adjusted confidence limits are $t^*_{454} = 1.28$ and $t^*_{988} = 1.82$. These are similar to earlier limits, but adjusted slightly to the right, which seems unnecessary.

A recent theoretical development has been the approximation of the adjusted percentile limits using numerical differencing in place of resampling. The resulting method, called the ABC method, is fully discussed by Efron and Tibshirani (1993, Chapter 22).

5.5 Comments

There is often little difference among the various confidence limits, and as yet there is no definitive method of choice, despite the theoretically higher accuracy of the bootstrap t and adjusted percentile methods.

Iterated resampling is another device for improving on accuracy, essentially using bootstrapping a bootstrap confidence limit algorithm to estimate and correct for bias in coverage.

When θ is a vector, confidence regions require a shape. One possibility is to work with a quadratic from, such as $Q = (T-\theta)^T V^{-1}(T-\theta)$ where V is an estimated variance matrix. Using quantiles of the bootstrap distribution of Q, one would then obtain as an elliptical confidence region the set of θ satisfying $Q \leq q^*_{(1-\alpha)R}$ assuming the q^*'s were ordered. Methods such as this might be capable of giving good coverage, meaning close to nominal level $1-\alpha$, but the ellipsoidal shape could be very misleading — as the corresponding regions often are in nonlinear regression problems. One possible improvement would be to find a transformation which normalizes the distribution of T^*, and use it with the preceding method, followed by untransformation — assuming this can be done. Ideally one would wish to have something like a likelihood-based confidence region. This is to some extent possible using one of the nonparametric likelihood methods mentioned in §8. Further research is needed in this important area.

6 Tests

One tool in the statistician's toolkit is the permutation or randomization test, which gives a model-free assessment of significance for a test statistic. Bootstrap tests have a similar but wider domain. For example they can be applied to fairly standard problems, such testing equality of means, lack of correlation and zero regression effects. Or they can be applied to test unimodality of a density, equality of curves, and so forth where no explicit model is available.

In most applications we choose a test statistic T, observe its value t and then want to calculate the significance probability $P = \mathrm{pr}(T \geq t \mid H_0)$ where H_0 is the null hypothesis of interest. (We assume that T is carefully chosen so as to give large values under likely alternative hypotheses.) Ideally P is uniquely specified

by H_0 but this is not always possible. So we must choose a null-hypothesis data distribution F_0 which satisfies H_0, and calculate

$$P = \text{pr}(T \geq t \mid \hat{F}_0).$$

In parametric analysis $\hat{F}_0$ would be a maximum likelihood estimate, and T would be chosen so that its null hypothesis distribution depends as little as possible on the exact values of unspecified parameters, e.g. by conditioning. Our interest, however, is in nonparametric or semiparametric models where the stochastic element (e.g. error distribution) is unknown.

6.1 Permutation tests

Permutation tests, and related randomization tests, are well-known and widely used. For recent accounts see the book by Manly (1991), and the article by Romano (1989).

As an example, suppose we wish to test the null hypothesis of independence of X and Y using the sample correlation $R(\mathbf{x}, \mathbf{y})$ whose observed value is r. Then the permutation test significance is

$$P = \text{pr}[R\{\mathbf{x}, perm(\mathbf{y})\} \geq r]$$

which can be approximated by Monte Carlo — for example generating 1000 random permutations of $\mathbf{y}$ and counting the proportion of them which lead to $R\{\mathbf{x}, perm(\mathbf{y})\} \geq r$.

The permutation distribution, which is justified by a conditioning argument, is very similar to that obtained under a null hypothesis estimate of data distribution.

Note that adaptive permutation testing, is possible with data-based selection of statistic, as described by Donegani (1991).

6.2 Nonparametric bootstrap test

We define the null hypothesis sampling model by bending the EDF $\hat{F}$ to $\hat{F}_0$ which satisfies H_0. The calculation of P will usually be by bootstrap sampling from $\hat{F}_0$ to obtain R simulated samples and their test statistic values t^*, and then

$$P = \frac{1}{R}(\text{ number of times } t^* \geq t).$$

For example, in the problem of testing a correlation we can take $\hat{F}_0 = \hat{G}\hat{H}$, the product of the marginal EDF's of x and y. The resulting test is very similar to the permutation test.

One general approach is to choose $\hat{F}_0$ to minimize some appropriate distance $dist(\hat{F}, \hat{F}_0)$ subject to H_0. This may lead to such things as empirical exponential families (Efron and Tibshirani, Chapter 21), or may be infeasible. Often it is simpler to impose conditions slightly stronger than H_0, but this should be done

with care. So, for example, to compare two sample means we could assume that under H_0 the corresponding two distributions are identical, in which case $\hat{F}_0$ is a common distribution equal to the pooled EDF.

The general recommendation for R when calculating probabilities is to use $R = 1000$ or bigger if possible, but this applies to estimating the whole distribution of T. In fact we could determine R sequentially: if $R = 20$ and $P \geq 0.5$ then stop and if $R = 100$ and $P \leq 0.01$ then stop, etc. For discussion see Besag and Clifford (1991).

For our ratio data, one model we have referred to is $E(Y \mid x) = \theta x$. Suppose that we had fitted a model-free curve to the (x, y) data by one of the new computer-intensive exploratory methods. To compare such a curve to the fitted linear model is one way to test for linearity. Figure 1 shows an example, the solid curve being a generalized additive model fit.

Thinking of the linear model as a generalized linear model, we do not know the error distribution: it is too dispersed to be Poisson. So how can we do a test? The first step is to choose a test statistic S, such as sum of squared differences between the line and the curve. Next we must choose the null hypothesis model to generate bootstrap samples. One way to do this is as follows: First obtain the weighted least squares fit $\hat{y} = tx$, where t is the ratio estimate if we use Poisson weights. Then simulate data according to the null hypothesis model

$$y_i^* = tx_i + \sqrt{x_i}\delta_i^*, \quad \delta_i^* \sim U(\hat{\delta}_1, \ldots, \hat{\delta}_n)$$

with $\hat{\delta}_j = \text{round}(y_j - tx_j)^+ / \sqrt{x_j}$. For each sample generated in this way, refit the line by weighted least squares and refit the nonparametric curve, then calculate the measure of difference S^*. Such an analysis for these data was carried out by Firth, Glosup and Hinkley (1991) who found a nonsignificant result ($P \gg 0.05$) as one would expect *a priori*. This contrasted with an analysis assuming Poisson error distribution, which misleadingly found a very significant result.

It seems very likely that bootstrapping will occupy a key place in the assessment of nonparametric curve fitting.

7 Sensitivity Analysis

In order to understand the implications of a statistical calculation, it is important to assess its sensitivity to changes in the data. If a parametric model has been fitted, there is wide range of diagnostics for detecting outliers and influential cases, particularly in regression analysis, and careful scrutiny of these is part of a bootstrap analysis, just it is part of any other analysis. But if a nonparametric bootstrap has been used, then the empirical distribution function $\hat{F}$ in effect is the model, and there is no baseline against which to test for outliers. Sensitivity analysis will then concern the effect of individual observations on bootstrap calculations, to answer questions such as "would the confidence interval differ greatly if

this point was removed?", and "what happens to the significance level when this observation is deleted?"

A direct answer to these questions is possible from the array of bootstrap frequencies, simply by restricting attention to those replicates in which observations do not appear. That is, for each observation, we examine the distribution of the t_r^* for samples where it did not appear. This is called the jackknife-after-bootstrap, because it involves seeing the effect of deleting each observation in turn on a bootstrap calculation. See Efron (1992). Figure 4 shows the quantiles of these jackknifed distributions plotted against the infinitesimal jackknife values for the observations. If observation 30, which has the largest L_j, was deleted, the distribution would become slightly more concentrated, with the largest change at its upper end, and its quantiles would decrease by about 0.05. Deletion of each of the other observations seems not to change the concentration of the simulated distribution, though the quantiles move in the way one would expect.

8 Empirical Likelihoods

Likelihood methods are widely used in parametric inference because they allow formal use of prior information via Bayes' theorem, they allow amalgamation of information from different samples, they are often transformation-invariant, and they can lead to confidence sets for multivariate parameters. There is a growing literature on nonparametric analogues of the likelihood function, the most prominent of which is Owen's (1988, 1990, 1991) empirical likelihood.

Suppose that given a random sample $y_1, \ldots, y_n$, we restrict interest to distributions supported on the data, that is, we focus on multinomial distributions that put probabilities p_j on the observations y_j. A parameter θ determined nonparametrically by $t(F) = \theta$ will be a function of the p_j, so we can write $t(F) \equiv t(p_1, \ldots, p_n)$. The observed value of θ is $t_0 = t(n^{-1}, \ldots, n^{-1})$. Owen proposed that inference about θ be based on the profile likelihood

$$L_E(\theta) = \sup_{p:t(p_1,\ldots,p_n)=\theta} \prod_{j=1}^{n} p_j,$$

which he called the empirical likelihood for θ. In a situation like ours, where θ is determined by the estimating equation $\int u(\theta; y)\, dF(y) = 0$, a straightforward application of Lagrange multipliers shows that

$$L_E(\theta) = \prod_j \frac{1}{1 + \eta_\theta u(\theta; y_j)},$$

where the Lagrange multiplier η_θ is the root of the equation

$$\sum_j \frac{u(\theta; y_j)}{1 + \eta_\theta u(\theta; y_j)} = 0.$$

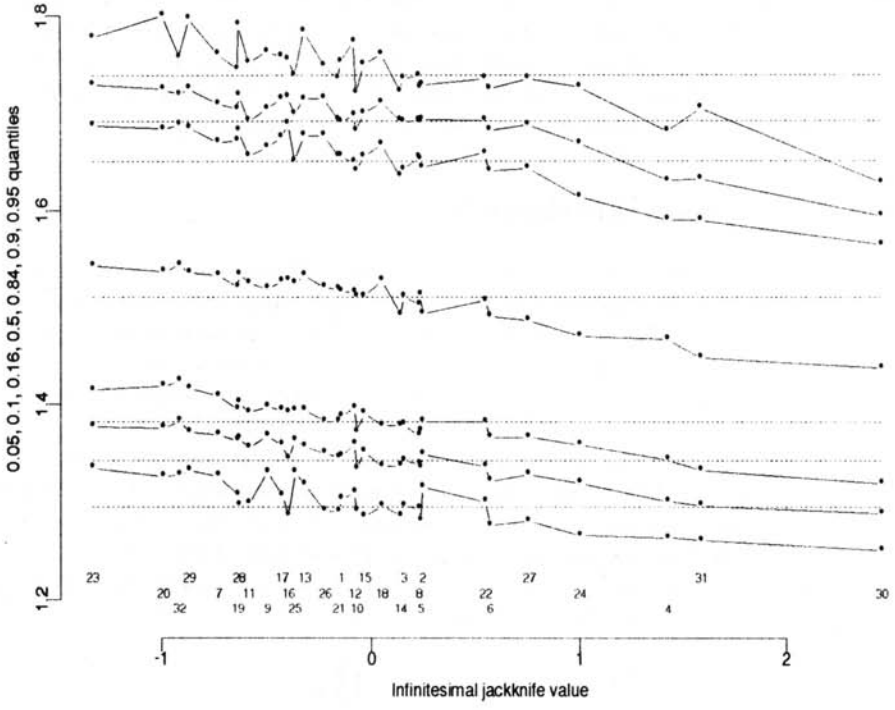

Figure 4: Jackknife after bootstrap plot for cloth data. *The solid lines show the effect on the* 0.05, 0.1, 0.16, 0.5, 0.84, 0.9, *and* 0.95 *quantiles of dropping each observation in turn, plotted against the infinitesimal jackknife value for the ratio. The dotted lines are the quantiles based on the entire dataset. The observation number is shown at the foot of the plot.*

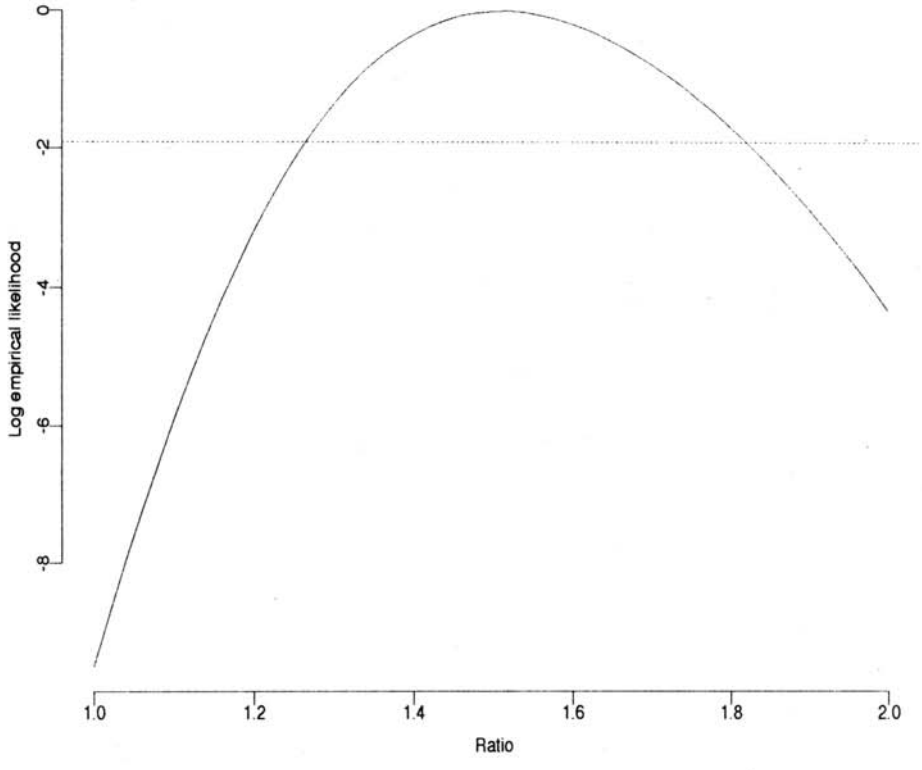

Figure 5: Log empirical likelihood for cloth data.

For the ratio, $u(\theta; y_j) = y_j - \theta x_j$, and given a value of θ, η_θ is easily obtained numerically. The log empirical likelihood for the cloth data, $\log L_E(\theta)$, is shown in Figure 4.

Remarkably, standard chi-squared approximations apply to the empirical likelihood ratio statistic, and a 95% confidence interval for the true ratio can be read off as the set of θ such that $2\{\log L_E(t_0) - \log L_E(\theta)\} \leq \chi^2_1(0.95)$. The interval is (1.26, 1.82), which compares well with the more sophisticated bootstrap intervals in §5.

The detailed properties of L_E have been described by DiCiccio, Hall, and Romano (1991), Hall and La Scala (1990), and in unpublished work by Corcoran and Spady. Other nonparametric likelihood analogues have been investigated by Davison, Hinkley, and Worton (1992), Hall (1987) and Efron (1993). See Efron and Tibshirani (1993, Chapter 24) for more discussion.

Acknowledgements

We thank the organizers of CLAPEM V for the invitation to present a short course on bootstrap methods, SERC for supporting this work through grants and an Advanced Research Fellowship to ACD, and Professor Antonio Galves for his forbearance in waiting for this paper to arrive.

References

Besag, J.E. and Clifford, P. (1991) Sequential Monte Carlo p-values. *Biometrika*, **78**, 301-304.

Bissell, A.F. (1972). A negative binomial model with varying element size. *Biometrika*, **59**, 435–41.

Booth, J.G., Hall, P. and Wood, A.T. (1993) Balanced importance resampling for the bootstrap. *Ann. Statist.*, **21**, 286–298.

Daniels, H.E. (1983). Saddlepoint approximations for estimating equations. *Biometrika*, **70**, 89–96.

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

Davison, A.C. and Hinkley, D.V. (1988) Saddlepoint appoximations in resampling methods. *Biometrika*, **75**, 417–431.

Davison, A.C., Hinkley, D.V. and Schechtman, E. (1986) Efficient bootstrap simulation. *Biometrika*, **73**, 555–566.

Davison, A.C., Hinkley, D.V., and Worton, B.J. (1992) Bootstrap likelihoods. *Biometrika*, **79**, 113–130.

De Angelis, D. and Young, G.A. (1992) Smoothing the bootstrap. *Int. Statist. Rev.*, **60**, 45–56.

DiCiccio, T.J., Hall, P. and Romano, J.P. (1991) Empirical likelihood is Bartlett-correctable. *Ann. Statist.*, **19**, 1053–1061.

DiCiccio, T.J., Martin, M.A. and Young, G.A. (1992) Fast and accurate double bootstrap confidence intervals. *Biometrika*, **79**, 285–295.

Do, K.-A. and Hall, P. (1991) On importance resampling for the bootstrap. *Biometrika*, **78**, 161–167.

Do, K.-A. and Hall, P (1992) Distribution estimation using concomitants of order statistics, with application to Monte Carlo simulation for the bootstrap. *J. R. Statist. Soc. B* **54**, 595–607.

Donegani, M. (1991) An adaptive and powerful randomization test. *Biometrika*, **78**, 930–933.

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

Efron, B. (1990) More efficient bootstrap computations. *J. Amer. Statist. Assoc.*, **55**, 79–89.

Efron, B. (1992) Jackknife-after-bootstrap standard errors and influence functions (with Discussion). *J. R. Statist. Soc. B*, **54**, 83–127.

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

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

Firth, D., Glosup, J. and Hinkley, D.V. (1991) Model checking with nonparametric curves. *Biometrika*, **78**, 245–252.

Gleason, J.R. (1988) Algorithms for balanced bootstrap simulations. *American Statistician*, **42**, 263–266.

Graham, R.L., Hinkley, D.V., John, P.W.M. and Shi, S. (1990) Balanced design of bootstrap simulations. *J. R. Statist. Soc. B*, **52**, 185–202.

Hall, P. (1987) On the bootstrap and likelihood-based confidence regions. *Biometrika*, **74**, 481–93.

Hall, P.G. (1989) Antithetic resampling for the bootstrap. *Biometrika*, **76**, 713–724.

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

Hall, P. and La Scala, B. (1990) Methodology and algorithms of empirical likelihood. *Int. Statist. Rev.*, **58**, 109–28.

Hinkley, D.V. and Shi, S. (1989) Importance sampling and the nested bootstrap. *Biometrika*, **76**, 435–446.

Johns, M.V. (1988) Importance sampling for bootstrap confidence intervals. *J. Amer. Statist. Soc.*, **83**, 709–714.

Leger, C. and Romano, J.P. (1990) Bootstrap adaptive estimation: The trimmed mean example. *Canad. J. Statist.*, **18**, 297–314.

Manly, B.F.J. (1991) *Randomization and Monte Carlo Methods in Biology.* London: Chapman and Hall.

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

Owen, A.B. (1990) Empirical likelihood ratio confidence regions. *Ann. Statist.*, **18**, 90–120.

Owen, A.B. (1991) Empirical likelihood for linear models. *Ann. Statist.*, **19**, 1725–1747.

Reid, N. (1988) Saddlepoint methods and statistical inference (with Discussion). Statistical Science, **3**, 213–238.

Romano, J.P. (1989) Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Statist.*, **17**, 141–149.

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

A.C. Davison and D.V. Hinkley
Department of Statistics
University of Oxford
1 South Parks Road
Oxford OX1 3TG
UK