STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:1141–1164

Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians

James Carpenter[1,*] and John Bithell[2]

[1] *Medical Statistics Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, U.K.*
[2] *Department of Statistics, 1 South Parks Road, Oxford, OX1 3TG, U.K.*

SUMMARY

Since the early 1980s, a bewildering array of methods for constructing bootstrap confidence intervals have been proposed. In this article, we address the following questions. First, when should bootstrap confidence intervals be used. Secondly, which method should be chosen, and thirdly, how should it be implemented. In order to do this, we review the common algorithms for resampling and methods for constructing bootstrap confidence intervals, together with some less well known ones, highlighting their strengths and weaknesses. We then present a simulation study, a flow chart for choosing an appropriate method and a survival analysis example. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION: CONFIDENCE INTERVALS AND COVERAGE ERROR

An accurate estimate of the uncertainty associated with parameter estimates is important to avoid misleading inference. This uncertainty is usually summarized by a confidence interval or region, which is claimed to include the true parameter value with a specified probability. In this paper we shall restrict ourselves to confidence intervals. We begin with an example which illustrates why we might want bootstrap confidence intervals.

1.1. Example: Remission from acute myelogenous leukaemia

Embury *et al.* [1] conducted a clinical trial to evaluate the efficacy of maintenance chemotherapy for acute myelogenous leukaemia. Twenty-three patients who were in remission after treatment with chemotherapy were randomized into two groups. The first group continued to receive maintenance chemotherapy, while the second did not. The objective of the trial was to examine whether maintenance chemotherapy prolonged the time until relapse. The preliminary results of the study are shown in Table I. We wish to test the hypothesis that maintenance chemotheraphy does not delay relapse by constructing a confidence interval for the treatment effect.

* Correspondence to: James Carpenter, Medical Statistics Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, U.K.

Received January 1998
Accepted August 1999

Table I. Remission times (weeks) for patients with acute myelogenous leukaemia; group 1 with maintenance chemotherapy; group 2 none. An entry such as > 13 means that the only information available is that at 13 weeks the patient was still in remission.

Treat 1	Treat 2
9	5
13	5
> 13	8
18	8
12	12
23	> 16
31	23
34	27
> 45	30
48	33
> 161	43
	45

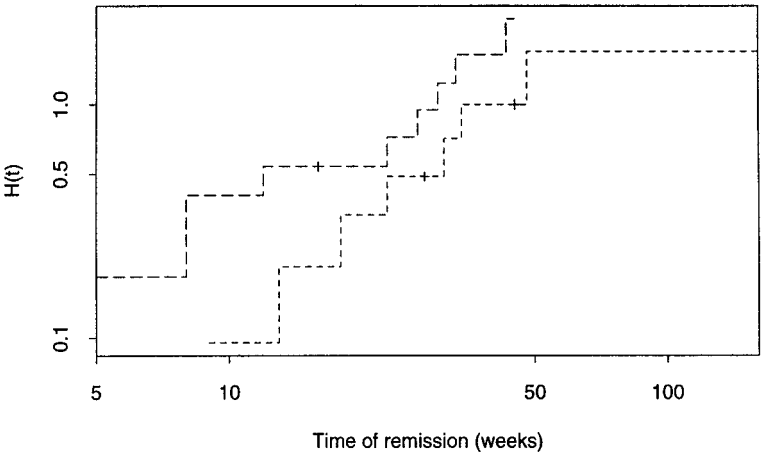

Figure 1. Plot of non-parametric estimate of cumulative hazard function for the data in Table I, on a log-log scale. The upper line corresponds to treatment group 2.

Figure 1 shows a log-log plot of the cumulative hazard function, and suggests that a proportional hazards model

$$h(t) = h_0(t) \exp(\beta x)$$

will be adequate, where $h(t)$ is the hazard at time t, $h_0(\cdot)$ is the baseline hazard and x is an indicator covariate for treatment 2. Fitting this model gives $\hat{\beta} = 0.924$ with standard error $\hat{\sigma} = 0.512$. A standard 95 per cent normal approximation confidence interval for β is therefore $\hat{\beta} \pm 1.96 \times 0.512$, that is, $(-0.080, 1.928)$.

However, the accuracy of this interval depends on the asymptotic normality of $\hat{\beta}$ and this assumption is questionable with so few observations. Accordingly, we may want to construct a confidence interval that does not depend on this assumption. Bootstrapping provides a ready, reliable way to do this. The principal questions are *which* bootstrap confidence interval method should be chosen and *what* should be done to implement it. These are the questions this article seeks to answer in a practical context, by reviewing the methods available, highlighting their motivation, strengths and weaknesses. After doing this, we return to this example in Section 7.

We begin by defining *coverage error*, which is a key concept in comparing bootstrap confidence interval methods. Suppose $(-\infty, \theta_U)$ is, for example, a normal approximation confidence interval, with nominal coverage $100(1-\alpha)$ per cent (α typically 0.05). Then it will often have a *coverage error* so that

$$\mathbf{P}(\theta < \theta_U) = (1-\alpha) + C$$

for some unknown constant C, where typically $C \to 0$ as n, the sample size, $\to \infty$.

Bootstrap confidence intervals aim to reduce this coverage error by using simulation to avoid the assumptions inherent in classical procedures. While they are often calculated for small data sets (for example, to check on the assumption of asymptotic normality), they are equally applicable to large data sets and complex models; see for example Carpenter [2] and LePage and Billard [3].

Thus, bootstrap confidence intervals will at the least validate the assumptions necessary to construct classical intervals, while they may further avoid misleading inferences being drawn. The way they go about this, and the extent on which they succeed, are described in Sections 2 and 3 and illustrated in Sections 5 and 7. First, however, we describe the *bootstrap principle* and the terms *non-parametric* and *parametric* simulation.

In many statistical problems we seek information about the value of a population parameter θ by drawing a random sample $\mathbf{Y}$ from that population and constructing an estimate $\hat{\theta}(\mathbf{Y})$ of the value of θ from that sample. The bootstrap principle is to obtain information about the relationship between θ and the random variable $\hat{\theta}(\mathbf{Y})$ by looking at the relationship between $\hat{\theta}(\mathbf{y}_{obs})$ and $\hat{\theta}(\mathbf{Y}^*)$, where $\mathbf{Y}^*$ is a resample characterized by the sample $\mathbf{y}_{obs}$. $\mathbf{Y}^*$ can either be constructed by sampling with replacement from the data vector $\mathbf{y}_{obs}$, the so-called *non-parametric* bootstrap, or by sampling from the distribution function parameterized by $\hat{\theta}(\mathbf{y}_{obs})$, the so-called *parametric* bootstrap.

Before we discuss the various methods for bootstrap confidence interval construction, we give algorithms for non-parametric and parametric simulation, and illustrate these in a regression context, where the bootstrap is frequently applied.

2. RESAMPLING PLANS

Here we give algorithms for non-parametric and semi-parametric resampling plans, and illustrate them with a linear model example. We first describe this example.

Table II. Weights of 14 babies at birth and 70–100 days. From Armitage and Berry (Reference [4], p. 148).

Case number	Birthweight (oz)	Weight at 70–100 days (oz)
1	72	121
2	112	183
3	111	184
4	107	184
5	119	181
6	92	161
7	126	222
8	80	174
9	81	178
10	84	180
11	115	148
12	118	168
13	128	189
14	128	192

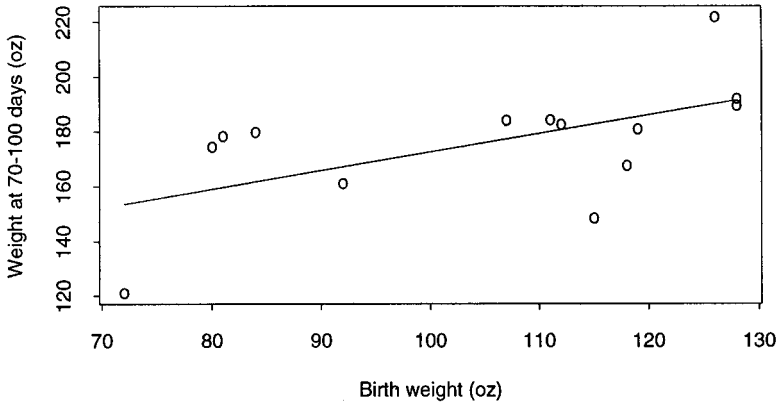

Figure 2. Plot of data in Table II, with least squares regression line 70–100 day weight $= \alpha + \beta \times$ birthweight. Estimates (standard errors) are $\hat{\beta} = 0.68$ (0.28), $\hat{\alpha} = 104.89$ (29.65).

Table II gives the weight at birth and 70–100 days of 14 babies, from Armitage and Berry (Reference [4], p. 148). The data are plotted in Figure 2, together with the least squares regression line. Parameter estimates are given in the caption. It appears that there is a borderline association between birthweight and weight at 70–100 days. However, the data set is small, and we may wish to confirm our conclusions by constructing a bootstrap confidence interval for the slope. To do this, we first need to construct bootstrap versions $\hat{\beta}^*$, of β. We now outline how to do this non-parametrically and parametrically.

2.1. Non-parametric resampling

Non-parametric resampling makes no assumptions concerning the distribution of, or model for, the data. Our data is assumed to be a vector $\mathbf{y}_{\text{obs}}$ of n independent observations, and we are interested in a confidence interval for $\hat{\theta}(\mathbf{y}_{\text{obs}})$. The general algorithm for a non-parametric bootstrap is as follows:

1. Sample n observations randomly with replacement from $\mathbf{y}_{\text{obs}}$ to obtain a bootstrap data set, denoted $\mathbf{Y}^*$.
2. Calculate the bootstrap version of the statistic of interest, $\hat{\theta}^* = \hat{\theta}(\mathbf{Y}^*)$.
3. Repeat steps 1 and 2 a large number of times, say B, to obtain an estimate of the bootstrap distribution.

We discuss the value of B appropriate for confidence intervals in Section 2.4.

In the context of the birthweight data in Table II, each 'observation' in the original data set consists of a pair, or case, (x, y). For example, the first case is (72, 121). The algorithm then proceeds as follows:

1. Sample n cases randomly with replacement to obtain a bootstrap data set. Thus, a typical bootstrap data set might select the following cases:

$$4\ 5\ 2\ 4\ 9\ 10\ 3\ 3\ 6\ 2\ 1\ 6\ 9\ 8.$$

2. Fit the linear model to the bootstrap data and obtain the bootstrap slope, $\hat{\beta}^*$. For the specific bootstrap data set in step 1, $\hat{\beta}^* = 0.67$.
3. Repeat steps 1 and 2 a large number, say B, of times to obtain an estimate of the bootstrap distribution.

The bootstrap slopes $\hat{\beta}_1^*, \ldots, \hat{\beta}_B^*$, can then be used to form a non-parametric bootstrap confidence interval for β as described in Section 3.

2.2. Parametric resampling

In parametric resampling we assume that a parametric model for the data, $\mathrm{F}_Y(y; .)$, is known up to the unknown parameter vector, θ, so that bootstrap data are sampled from $\mathrm{F}_Y(y; \hat{\theta})$, where $\hat{\theta}$ is typically the maximum likelihood estimate from the original data. More formally, the algorithm for the parametric bootstrap is as follows:

1. Let $\hat{\theta}$ be the estimate of θ obtained from the data (for example, the maximum likelihood estimate). Sample n observations, denoted $\mathbf{Y}^*$ from the model $\mathrm{F}_Y(.; \hat{\theta})$.
2. Calculate $\hat{\theta}^* = \hat{\theta}(\mathbf{Y}^*)$.
3. Repeat 1 and 2 B times to obtain an estimate of the parametric bootstrap distribution.

In the linear model example, 'assuming the model' means treating the assumptions of the linear model as true, that is, assuming that the x's (the birthweights) are known without error and that the residuals are normally distributed with mean zero and variance given by the residual standard error, which is $\sigma^2 = 14.1$. We then sample $n = 14$ residuals and pass these back through the model to obtain the bootstrap data. The algorithm is as follows:

1. Draw n observations, $z_1, \ldots, z_{14}$, from the $\mathrm{N}(0, \sigma^2)$ distribution.

2. Calculate bootstrap responses, y_i^*, as $y_i^* = \hat{\alpha} + \hat{\beta}x_i + z_i$, $i \in 1, \ldots, 14$. For example, if $z_1 = 31.2$

$$y_1^* = 104.89 + 0.68 \times 72 + 31.2 = 185.05.$$

   The bootstrap data set then consists of the $n = 14$ pairs (y_i^*, x_i).
3. Calculate the bootstrap slope, $\hat{\beta}^*$, as the least squares regression slope for this bootstrap data set.
4. Repeat 1–3 B times to obtain an estimate of the parametric bootstrap distribution.

As before, the resulting bootstrap sample, $\hat{\beta}_1^*, \ldots, \hat{\beta}_B^*$, can be used to construct a bootstrap confidence interval.

2.3. Semi-parametric resampling

This resampling plan is a variant of parametric resampling, appropriate for some forms of regression. Since it involves non-parametric resampling of the residuals from the fitted parametric model, we term it *semi-parametric resampling*. As before, we begin with a general presentation.

Suppose we have responses $\mathbf{y} = (y_1, \ldots, y_n)$ with covariates $\mathbf{x} = (x_1, \ldots, x_n)$, and we fit the model

$$\mathbf{y} = g(\beta, \mathbf{x}) + \mathbf{r}$$

obtaining estimates $\hat{\beta}$ of the parameters β and a set of residuals r_i, $i \in (1, \ldots, n)$.
Without loss of generality, suppose we are interested in a confidence interval for β_1. The algorithm for semi-parametric resampling is as follows:

1. Adjust the residuals $\mathbf{r} = (r_1, \ldots, r_n)$ so that they have approximately equal means and variances. Denote this new set by $(\tilde{r}_1, \ldots, \tilde{r}_n)$.
2. Sample with replacement from the set of adjusted residuals $(\tilde{r}_1, \ldots, \tilde{r}_n)$ to obtain a set of bootstrap errors, $\mathbf{r}^* = (\tilde{r}_1^*, \ldots, \tilde{r}_n^*)$.
3. Then obtain bootstrap data $\mathbf{y}^* = (y_1^*, \ldots, y_n^*)$ by setting

$$\mathbf{y}^* = g(\hat{\beta}, \mathbf{x}) + \mathbf{r}^*.$$

4. Next fit the model,

$$\mathbf{E}\mathbf{y}^* = g(\beta, \mathbf{x})$$

   to obtain the bootstrap estimate $\hat{\beta}^*$.
5. Repeat steps 1–4 above B times to obtain an estimate of the bootstrap distribution.

It is important to note that this resampling plan is only appropriate when it is reasonable to assume that the adjusted residuals, $\tilde{r}_i$, are independent and identically distributed (henceforth i.i.d.). If this is reasonable, then this resampling plan is usually preferable to the straight parametric plan, as it does not force the residuals to conform to a known distribution.

However, if the $\tilde{r}_i$, are not i.i.d. so that, for example, the variance of y depends on its mean, then this must be modelled, and the sampling algorithm revised. An algorithm for this situation is given by Davison and Hinkley (Reference [5], p. 271).

In the linear model example, having obtained the least squares estimates of $\hat{\alpha}$ and $\hat{\beta}$, we calculate the residuals as

$$r_i = y_i - \hat{\alpha} - \hat{\beta}x_i, \quad i \in 1, \ldots, 14$$

Let $\bar{r} = \sum_{i=1}^{14} r_i/14$, the mean of the residuals, and let $\tilde{r}_i = r_i - \bar{r}$, $i \in 1, \ldots, 14$. We then proceed as follows:

1. Sample from the set of adjusted residuals $\tilde{r}_i$, $i \in 1, \ldots, 14$, randomly with replacement, so obtaining $r_1^*, \ldots, r_{14}^*$.
2. The calculate the bootstrap data as

$$y_i^* = \hat{\alpha} + \hat{\beta}x_i + r_i^*, \quad i \in 1, \ldots, 14.$$

3. Fit the linear model to the bootstrap data (y_i^*, x_i), obtaining $\hat{\beta}^*$.
4. Repeat steps 1–3 B times to obtain an estimate of the bootstrap distribution.

2.4. How many bootstrap samples

A key question faced by anyone using the bootstrap is how large should B be. For 90–95 per cent confidence intervals, most practitioners (for example, Efron and Tibshirani, Reference [6], p. 162, Davison and Hinkley, Reference [5], p. 194) suggest that B should be between 1000 and 2000.

Further, estimating a confidence interval usually requires estimating the 100α percentile of the bootstrap distribution. To do this, the bootstrap sample is first sorted into ascending order. Then, if $\alpha(B + 1)$ is an integer, the percentile is estimated by the $\alpha(B + 1)$th member of the ordered bootstrap sample (Cox and Hinkley, Reference [7], Appendix A). Otherwise, interpolation must be used, between the $\lfloor \alpha(B + 1) \rfloor$th and $(\lfloor \alpha(B + 1) \rfloor + 1)$th members of the ordered sample, where $\lfloor . \rfloor$ denotes the integer part, using the formula (10) below. Consequently, choosing $B = 999$ or $B = 1999$ leads to simple calculations for the common choices of α.

2.5. Choosing the simulation method

Clearly, parametric and non-parametric simulation make very different assumptions. The general principle that underlies the many algorithms given by Davison and Hinkley [5] and Efron and Tibshirani [6] is that the simulation process should mirror as closely as possible the process that gave rise to the observed data.

Thus, if we believe a particular model, which is to say we believe that the fitted model differs from the true model only because true values of the parameters have been replaced by estimates obtained from the data, then the parametric (or in regression, preferably semi-parametric) resampling plan is appropriate. However, examination of the residuals may cast doubt on the modelling assumptions. In this case, non-parametric simulation is often appropriate. It is interesting to note that, in practice, non-parametric simulation gives results that generally mimic the results obtained under the best fitting, *not* the simplest parametric model (see Hinkley, and Young in reply to Hinkley [5,8]).

3. BOOTSTRAP CONFIDENCE INTERVALS

The previous section has described how to obtain parametric and non-parametric bootstrap samples in general, and illustrated the particular case of linear regression. In this section we assume that we have obtained $B = 999$ bootstrap samples of θ, the parameter of interest, and that we have sorted them into order. Let

$$\hat{\theta}^*_1, \ldots, \hat{\theta}^*_{999} \tag{1}$$

denote this ordered set, so that $\hat{\theta}^*_i < \hat{\theta}^*_j$, for $1 \leqslant i < j \leqslant 999$. Of course, in the linear regression example the parameter of interest is $\theta = \beta$, the slope.

All the established methods we discuss below are described in more technical detail by Hall [9]. The test-inversion intervals are reviewed, and some theory given, by Carpenter [2]. More practical examples of confidence interval construction are given by Efron and Tibshirani [6] and Davison and Hinkley [5], together with some S-plus software. DiCiccio and Efron [10] give a further review of most of the methods described below. An alternative viewpoint is given by Young [8].

Here, we emphasize the underlying principles of the methods and their relationship to each other, together with practical implementation.

For each method, we outline its rationale, show how to calculate a one-sided 95 per cent interval (two-sided 95 per cent intervals are formed as the intersection of two one-sided 97.5 per cent intervals; for general $100(1 - \alpha)$ per cent intervals, interpolation may be needed, as described in Section 3.4). We further list the advantages and disadvantages of each method. In the course of this we indicate whether a method is transformation respecting or not. By this we mean the following. Suppose a bootstrap confidence interval method gives $(-\infty, \theta_U)$ as a $100(1 - \alpha)$ per cent interval for θ. Let g be a monotonic transformation, and suppose we require an interval for $g(\theta)$. Then if it is simply $(-\infty, g(\theta_U))$, we say the method is transformation respecting. Otherwise, we could get a markedly different interval when we repeated the bootstrap calculations with θ replaced by $g(\theta)$, $\hat{\theta}$ by $g(\hat{\theta})$ and so on.

We also give each method's one-sided coverage error. This is a theoretical property of the method, whose proof is only valid under a certain regularity conditions, namely that $(\hat{\theta} - \theta)/\hat{\sigma}$ is a smooth function of sample moments with an asymptotic normal distribution [9]. Coverage errors are typically $O(n^{-j})$, $j = \frac{1}{2}$ or 1.

Figure 3 schematically describes the evolution of bootstrap confidence intervals and divides them into three families. The pivotal family arguably represents the most natural approach to the problem. In this family, the confidence interval is constructed in the usual way except that the quantiles of known distributions (normal, Student's-t etc.) are replaced by their bootstrap estimates. Non-pivotal methods are less intuitive, and have been primarily championed by Efron and co-workers [6]. The simplest of these, the percentile method, has been described as tantamount to 'looking up the wrong statistical tables backwards' (Hall, Reference [9] p. 36). The other methods in the non-pivotal family have successively more complex analytical corrections for this. Finally, the test-inversion intervals exploit the duality between confidence intervals and tests. They can be used in semi-parametric and parametric situations, but not in non-parametric situations, and are particularly useful for regression style problems [2].

We begin with pivotal methods.

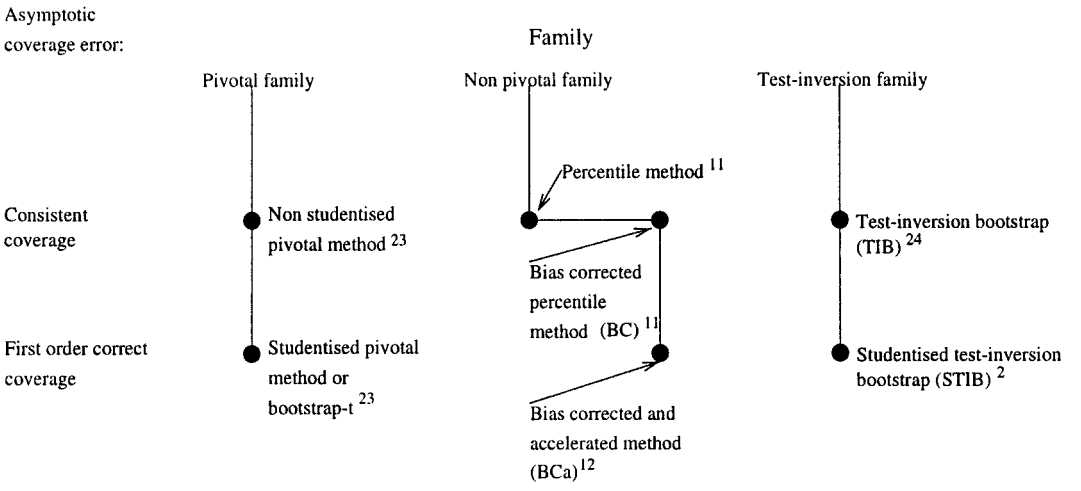

Figure 3. Schematic diagram describing relationship of bootstrap methods for constructing confidence intervals, with early references. *Consistent* coverage means the coverage error is $O(n^{-1/2})$; *first-order correct* coverage means the coverage error is $O(n^{-1})$. See the introduction and Table III for more details.

3.1. Non-Studentized pivotal method (sometimes called 'basic')

3.1.1. Rationale. Arguably the most natural way of setting about constructing a confidence interval for θ is to seek a function of the estimator and parameter whose distribution is known, and then use the quantiles of this known distribution to construct a confidence interval for the parameter. However, in the absence of detailed knowledge about the distribution $F_Y(.;\theta)$ from which the observations are drawn, it is not clear which function of the parameter and estimator should be chosen. Indeed, the procedure outlined below could, in principle, be applied to any statistic $W = g(\hat{\Theta}; \theta)$, provided g is continuous. However, in view of the fact that many estimators are asymptotically normally distributed about their mean, it makes sense to use $W = \hat{\Theta} - \theta$.

Suppose that the distribution of W were known, so that we could find w_α such that $\mathbf{P}(W \leqslant w_\alpha) = \alpha$. Then a one-sided $100(1-\alpha)$ per cent confidence interval for θ would be

$$(-\infty, \hat{\theta} - w_\alpha) \tag{2}$$

The problem is that, in practice, the distribution of W is not known. However, the bootstrap principle implies that we can learn about the relationship between the true parameter value θ and the estimator $\hat{\theta}$ by pretending that $\hat{\theta}$ is the true parameter value and looking at the relationship between $\hat{\theta}$ and estimates of $\hat{\theta}$, denoted by $\hat{\theta}^*$, which we construct from data simulated from the model parameterized by $\hat{\theta}$.

In other words, the bootstrap procedure tells us to replace the quantile, w_α, of W in (2) with the appropriate quantile, w_α^*, of $W^* = (\hat{\Theta}^* - \hat{\theta})$, calculated as described below. The non-Studentized pivotal interval is thus

$$(-\infty, \hat{\theta} - w_\alpha^*) \tag{3}$$

3.1.2. Calculation of 95 per cent interval. Recall we have already calculated and ordered our bootstrap sample (1).

1. Set $w_i^* = \hat{\theta}_i^* - \hat{\theta}$, for $i \in 1, \ldots, 999$.
2. The 0.05th quantile of W^* is then estimated by w_j^*, where $j = 0.05 \times (999 + 1) = 50$.
3. The one-sided 95 per cent confidence interval is thus

$$(-\infty, \hat{\theta} - \hat{w}_{50}^*)$$

Note that there are two distinct sources of error in this procedure. The first, termed *bootstrap error*, arises from appealing to the bootstrap principle to replace the quantile of W in (2) with that of W^*. The second, termed *Monte Carlo error*, arises because Monte Carlo simulation is used to estimate the 100αth percentile of W^* (steps 1 and 2 above). Provided the number of simulations in step 2 is sufficiently large, (see Section 2.4), the Monte Carlo error is usually negligible compared to the bootstrap error. The accuracy of the method therefore depends critically on the similarity of the distributions of W and W^*. Unfortunately, these are generally not close (Davison and Hinkley, Reference [5], p. 211).

3.1.3. Advantages. Often provides an accurate, simple to calculate, confidence interval for the sample median (Davison and Hinkley, Reference [5], p. 42).

3.1.4. Disadvantages. Typically substantial coverage error because the distributions of W and W^* differ markedly. If there is a parameter constraint (such as $\theta > 0$) then the interval often includes invalid parameter values.

3.1.5. Coverage error [9].

$$\mathbf{P}(\hat{\theta} - w_\alpha^* > \theta) = 1 - \alpha + O(n^{-1/2})$$

3.2. Bootstrap-t method (Studentized pivotal)

3.2.1. Rationale. As discussed above, the non-Studentized pivotal method is generally unreliable because the distributions of W and W^* differ markedly. If $\mathbf{y}_{\text{obs}}$ were a sample from the normal distribution, then the reason for this would be that the variance of W would not equal that of W^*. Denote an estimate of the standard deviation of $\hat{\theta}$ by $\hat{\sigma}$. Then the obvious way around this difficulty would be to work with $T = (\hat{\theta} - \theta)/\hat{\sigma}$ instead of W. The bootstrap-t method is just a generalization of this principle. Thus the interval (3) becomes

$$(-\infty, \hat{\theta} - \hat{\sigma} t_\alpha^*) \tag{4}$$

where t_α^* is the 100αth percentile of $T^* = (\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$, the bootstrap version of T.

3.2.2. Calculation of 95 per cent interval. To use this method, an estimate of the standard error of each $\hat{\theta}_i^*$ in (1) has to be calculated, for every i. For instance, if $\hat{\theta}(\mathbf{y}) = \sum_{i=1}^n y_i/n$, the sample mean, then $\hat{\sigma}^2(\hat{\theta}) = \sum_{i=1}^n (y_i - \hat{\theta}(\mathbf{y}))^2/n(n-1)$. For more complicated statistics, estimating σ can be problematic, although a variety of approximations exist. A moderately technical review is given by Davison and Hinkley Reference [5], Chapters 2 and 3. In regression problems, however, the

standard error of the parameters is readily available. In our simple linear regression example, it is given by

$$\frac{\text{Residual standard error}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

For the calculations below, we therefore assume that for each member of the bootstrap sample (1) we have calculated the corresponding standard error $\hat{\sigma}_i^*$, $i \in 1, \ldots, B$. We then proceed as follows:

1. Set $t_i^* = (\hat{\theta}_i^* - \hat{\theta})/\hat{\sigma}_i^*$, $i \in 1, \ldots, 999 = B$. Order this set of t_i^*'s.
2. Denote by $\hat{t}_{50}^*$ the 50th largest t^* (where $50 = 0.05 \times (B + 1)$).
3. The confidence interval is then

$$(-\infty, \hat{\theta} - \hat{\sigma}\hat{t}_{50}^*).$$

As already discussed, this method clearly depends on being able to calculate $\hat{\sigma}$. In many problems, for this method to be computationally feasible, an analytic form of $\hat{\sigma}$ must be known. However, for some statistics, even analytic approximations to $\hat{\sigma}$ may prove elusive. In this case, one option [5] is to carry out a computationally heavy, but routine, 'second level bootstrap' to estimate $\hat{\sigma}^*$ as follows.

From $\mathbf{y}_i^*$ obtain, using either parametric or non-parametric simulation, M $\mathbf{y}^{**}$'s and the corresponding values of $\hat{\theta}^{**} = \hat{\theta}(\mathbf{y}^{**})$. Then the variance of $\hat{\theta}(\mathbf{y}_i^*)$ is estimated by

$$\hat{\sigma}^{*2} = \frac{1}{M-1}\sum_{j=1}^{M}(\hat{\theta}_j^{**} - \hat{\theta}(\mathbf{y}_i^*))^2$$

This procedure must be repeated for each of the B $\hat{\theta}^*$'s. Finally, the variance of $\hat{\theta}(\mathbf{y}_{\text{obs}})$ can be estimated by

$$\hat{\sigma}^2 = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^* - \hat{\theta}(\mathbf{y}_{\text{obs}}))^2$$

Usually, $M = 25$ is sufficient for each of these second level bootstraps. Clearly though, this process is computationally heavy, as for every simulation required for the basic bootstrap-t method, a further M are now required, making MB altogether.

An additional problem with this method is that it will perform very poorly if $\hat{\sigma}$ is not (at least approximately) independent of $\hat{\theta}$. This assumption can easily be checked by plotting $\hat{\sigma}^*$ against $\hat{\theta}^*$, as shown in Section 7. If it is violated, we construct the variance stabilized bootstrap-t interval, as follows:

1. Use either parametric or non-parametric simulation to obtain B pairs $(\hat{\theta}^*, \hat{\sigma}^*)$, together with a second set of B $\hat{\theta}^*$'s.
2. Use a non-linear regression technique on the B pairs to estimate the function s such that $\hat{\sigma}^* = s(\hat{\theta}^*)$. Having estimated s, use numerical integration to estimate $f(x) = \int^x 1/s(\theta)\,\mathrm{d}\theta$. A Taylor series expansion shows that f is the approximate variance stabilizing transformation.
3. Transform the (hitherto unused) B $\hat{\theta}^*$'s to the variance stabilized scale. Since the variance is now assumed stable, we can construct a non-Studentized pivotal interval for $f(\theta)$.
4. Finally, calculate an estimate of f^{-1} and back transform the interval obtained in step 3 to the original scale.

In spite of all the numerical calculations, the variance stabilized bootstrap-t interval usually alleviates the problems of the bootstrap-t when the variance of $\hat{\theta}^*$ is not constant (Efron and Tibshirani, Reference [6], p. 162 ff; Carpenter [2]).

3.2.3. Advantages. Provided $\hat{\sigma}$ is easily available, the method performs reliably in many examples, in both its standard and variance stabilized forms (see Davison and Hinkley, Reference [5], p. 231, for a simulation study and references to others).

3.2.4. Disadvantages. The method is computationally very intensive if $\hat{\sigma}^*$ is calculated using a double bootstrap. Intervals can include invalid parameter values and it is not transformation respecting.

3.2.5. Coverage error [9].

$$\mathbf{P}(\theta - \hat{\sigma}t_\alpha^* > \theta) = 1 - \alpha + O(n^{-1})$$

3.3. Percentile Method

We now move to the second arm of Figure 3. These methods are in some ways less intuitive than those described above, but have the advantage of not requiring $\hat{\sigma}$.

3.3.1. Rationale. Consider a monotonically increasing function $g(.)$, and write $\phi = g(\theta)$, $\hat{\phi} = g(\hat{\theta})$ and $\hat{\phi}^* = g(\hat{\theta}^*)$. Choose (if possible) $g(.)$, such that

$$\hat{\phi}^* - \hat{\phi} \sim \hat{\phi} - \phi \sim \mathrm{N}(0, \sigma^2) \tag{5}$$

Then, since $\hat{\phi} - \phi \sim \mathrm{N}(0, \sigma^2)$, the interval for θ is

$$(-\infty, g^{-1}(\hat{\phi} - \sigma z_\alpha)) \tag{6}$$

where z_α is the 100α per cent point of the standard normal distribution. However, (5) implies that $\hat{\phi} - \sigma z_\alpha = \mathrm{F}_{\hat{\phi}^*}^{-1}(1-\alpha)$. Further, since g is monotonically increasing, $\mathrm{F}_{\hat{\phi}^*}^{-1}(1-\alpha) = g(\mathrm{F}_{\hat{\theta}^*}^{-1}(1-\alpha))$. Substituting in (6) gives the percentile interval

$$(-\infty, \mathrm{F}_{\hat{\theta}^*}^{-1}(1-\alpha)) \tag{7}$$

3.3.2. Calculation of 95 per cent interval. Recall the set of $B = 999$ bootstrap samples (1). The upper end point of the one-sided 95 per cent percentile interval is $\mathrm{F}_{\hat{\theta}^*}^{-1}(0.95)$, which is estimated by the 950th member of (1), since $950 = 0.95 \times (B+1)$. The percentile interval is then

$$(-\infty, \hat{\theta}^*_{950})$$

3.3.3. Advantages. Simplicity is the attraction of this method, and explains its continued popularity. Unlike the bootstrap-t, no estimates of the σ are required. Further, no invalid parameter values can be included in the interval.

Another advantage of this group of methods over the pivotal methods is that they are transformation respecting.

3.3.4. Disadvantages. The coverage error is often substantial if the distribution of $\hat{\theta}$ is not nearly symmetric (Efron and Tibshirani, Reference [6], p. 178 ff). The reason is that the justification of

the method rests on the existence of a $g(.)$ such that (5) holds, and for many problems such a g does not exist.

3.3.5. Coverage error [9].

$$\mathbf{P}(\hat{\theta}^*_{1-\alpha} > \theta) = 1 - \alpha + O(n^{-1/2})$$

3.4. Bias corrected method

The quickly recognized shortcomings of the percentile method [11] led to the development of the bias corrected or BC method.

3.4.1. Rationale. Again, consider a monotonically increasing function $g(.)$, and write $\phi = g(\theta)$, $\hat{\phi} = g(\hat{\theta})$ and $\hat{\phi}^* = g(\hat{\theta}^*)$. However, now (if possible) choose $g(.)$, such that

$$\hat{\phi}^* - \hat{\phi} \sim \hat{\phi} - \phi \sim \mathrm{N}(-b\sigma, \sigma^2) \tag{8}$$

for some constant b. An analogous (but slightly more complex) argument than that used in the case of the percentile interval then yields the BC interval

$$(-\infty, \mathrm{F}^{-1}_{\hat{\theta}^*}(\Phi(2b - z_\alpha))) \tag{9}$$

where b is estimated by $\Phi^{-1}(\mathbf{P}(\hat{\theta}^* \leqslant \hat{\theta}))$ and Φ^{-1} is the inverse cumulative distribution function of the normal distribution.

3.4.2. Calculation of 95 per cent interval.

1. Count the number of members of (1) that are less than $\hat{\theta}$ (calclated from the original data). Call this number p and set $b = \Phi^{-1}(p/B)$.
2. Calculate $Q = (B + 1)\Phi(2b - z_{0.05})$, where $z_{0.05} = -1.64$. Q is the percentile of the bootstrap distribution required for the upper endpoint of the bias corrected confidence interval.
3. Estimate the endpoint of the interval by $\hat{\theta}^*_{\lfloor Q \rfloor}$, where $\lfloor . \rfloor$ means 'take the integer part'. If a more accurate estimate is required, interpolation can be used between the members of (1), as follows. Let the nearest integers to Q be a, b, so that $a < Q < b$ and $b = a + 1$. Then the Qth percentile is estimated by

$$\hat{\theta}^*_Q \approx \hat{\theta}^*_a + \frac{\Phi^{-1}(\frac{Q}{B+1}) - \Phi^{-1}(\frac{a}{B+1})}{\Phi^{-1}(\frac{b}{B+1}) - \Phi^{-1}(\frac{a}{B+1})}(\hat{\theta}^*_b - \hat{\theta}^*_a) \tag{10}$$

The bias corrected interval, (7), is

$$(-\infty, \hat{\theta}^*_Q).$$

3.4.3. Advantages. The advantages are as for the percentile method, but see below.

3.4.4. Disadvantages. This method was devised as an improvement to the percentile method for non-symmetric problems. Hence, if the distribution of $\hat{\theta}^*$ is symmetric about $\hat{\theta}$, then $b = 0$ and the bias corrected and percentile intervals agree. However, the coverage error is still often substantial, because the validity of the method depends upon the existence of a $g(.)$ such that (8) holds, and for many problems such a g does not exist. In consequence, it has been omitted altogether from

recent discussions [5, 6]. It is worth mentioning, though, as it is still the most accurate method implemented in the software package stata.

3.4.5. Coverage error [9].

$$\mathbf{P}(\hat{\theta}^*_Q > \theta) = 1 - \alpha + O(n^{-1/2})$$

3.5. Bias corrected and accelerated method

The shortcomings of the BC method in turn led [12] to the development of the bias corrected and accelerated or BCa method. The idea is to allow not only for the lack of symmetry of $\mathrm{F}_{\hat{\Theta}}(.;\theta)$, but also for the fact that its shape, or skewness, might change as θ varies.

Note that the later abc method [13] is an analytic approximation to this method.

3.5.1. Rationale. Again, consider a monotonically increasing function $g(.)$, and write $\phi = g(\theta)$, $\hat{\phi} = g(\hat{\theta})$ and $\hat{\phi}^* = g(\hat{\theta}^*)$. However, now (if possible) choose $g(.)$, such that

$$\hat{\phi} \sim \mathrm{N}(\phi - b\sigma(\phi), \sigma^2(\phi))$$
$$\hat{\phi}^* \sim \mathrm{N}(\hat{\phi} - b\sigma(\hat{\phi}), \sigma^2(\hat{\phi}))$$

where $\sigma(x) = 1 + ax$. Again, an analogous argument to that used to justify the BC interval yields the BCa interval

$$\left(-\infty,\ \mathrm{F}^{-1}_{\hat{\Theta}^*}\left(\Phi\left(b - \frac{z_\alpha - b}{1 + a(z_\alpha - b)}\right); \hat{\theta}\right)\right) \tag{11}$$

where b is defined as before and a formula for estimating a is given below.

3.5.2. Calculation of 95 per cent interval.

1. Calculate b as for the BC interval.
2. Next we need to calculate a. This calculation depends on whether the simulation is non-parametric or parametric, and in the latter case, whether nuisance parameters are present. For completeness we give a simple jack-knife estimate of a; details about more sophisticated and accurate estimates can be found elsewhere [5, 6]. Let $\mathbf{y}^i_{\mathrm{obs}}$ represent the original data with the ith point omitted, and $\hat{\theta}^i = \hat{\theta}(\mathbf{y}^i_{\mathrm{obs}})$ be the estimate of θ constructed from this data. Let $\tilde{\theta}$ be the mean of the $\hat{\theta}^i$'s. Then a is estimated by
$$\frac{\sum_{i=1}^n (\tilde{\theta} - \hat{\theta}^i)^3}{6\left[\sum_{i=1}^n (\tilde{\theta} - \hat{\theta}^i)^2\right]^{3/2}}$$
3. Let $\tilde{Q}$ be the integer part of $(B+1)\,\Phi(b - \frac{z_{0.05} - b}{1 + a(z_{0.05} - b)})$, where $z_{0.05} = -1.64$.
4. Estimate the $\tilde{Q}$th percentile of the bootstrap distribution (1) as in step 3 of the bias corrected interval calculation. Then, the BCa interval, (11), is estimated by
$$(-\infty, \hat{\theta}^*_{\tilde{Q}})$$

3.5.3. Advantages. The advantages are as for the percentile method, plus this method generally has a smaller coverage error than the percentile and BC intervals (Efron and Tibshirani, Reference [6], p. 184 ff), but see below.

3.5.4. Disadvantages. The calculation of a can be tortuous in complex parametric problems. The coverage error of this method increases as $\alpha \to 0$. To see why this is so, note that as this happens, the right hand endpoint of the interval should be estimated by ever larger elements of the set of ordered $\hat{\theta}^*$'s. However, this is not the case: as $\alpha \to 0$

$$\Phi\left(b - \frac{z_\alpha - b}{1 + a(z_\alpha - b)}\right) \to \Phi(b - 1/a) \neq 1$$

This anomaly means that coverage can be erratic for small α, typically $\alpha < 0.025$. (Davison and Hinkley, Reference [5], p. 205, p. 231 and Section 5 below).

3.5.5. Coverage error [9].

$$\mathbf{P}(\hat{\theta}^*_Q > \theta) = 1 - \alpha + O(n^{-1})$$

3.6. Test-inversion bootstrap (TIB) method

The third class of methods we consider exploit the duality between confidence intervals and test-inversion (Rao, Reference [14], p. 471). These methods are less well known than those described hitherto, but enjoy certain advantages. A full discussion is given by Carpenter [2].

3.6.1. Rationale. The duality between confidence intervals and test-inversion means that the correct endpoint, θ_U, of the $100(1 - \alpha)$ per cent interval $(-\infty, \theta_U)$ satisfies

$$F_{\hat{\Theta}}(\hat{\theta}(\mathbf{y}_{obs}); \theta_U, \eta) = \alpha. \tag{12}$$

If η, the nuisance parameters, were known, then (12) could, in principle, be solved to find θ_U. However, in reality, η is unknown. Applying the bootstrap principle, we replace η by $\hat{\eta}$ and estimate θ_U by $\hat{\theta}_U$, which satisfies

$$F_{\hat{\Theta}}(\hat{\theta}(\mathbf{y}_{obs}); \hat{\theta}_U, \hat{\eta}) = \alpha. \tag{13}$$

We call $\hat{\theta}_U$ the endpoint of the test-inversion bootstrap interval.

3.6.2. Calculation of 95 per cent. In order to solve (13), we need to be able to simulate from the bootstrap distribution at different values of θ. Clearly, this is not possible with non-parametric 'case' resampling. However, it is possible within the parametric or semi-parametric resampling plans.

Recall the semi-parametric resampling plan for the linear regression example. To find the upper endpoint of the test-inversion interval for β, we replace $\hat{\beta}$ in step 2 of the semi-parametric resampling plan in Section 2.3 by the value of the current estimate of the upper endpoint, β^i_U, and leave the intercept, α, fixed at its estimate $\hat{\alpha}$. We then resample from residuals in the same way as before, obtaining bootstrap samples, $\hat{\beta}^*$, which we use to assess whether

$$\mathbf{P}(\hat{\beta}^* < \hat{\beta} \,|\, \beta = \beta^i_U) = \alpha \tag{14}$$

If the left hand side of (14) is less than α, we decrease β^i_U; otherwise we increase it, until a solution is found to the required accuracy.

In general, to determine the solution to (13), we first need to write a program to simulate from $F_{\hat{\Theta}}(.; \theta, \eta)$, at various values of θ. We then need a stochastic root finding algorithm to determine

the solution of (13). Various algorithms have been proposed, but a careful comparison across a range of distributions showed that the Robbins–Monro algorithm was the most efficient [2, 15, 16]. FORTRAN code implementing this algorithm [16] is available from statlib.

3.6.3. Advantages. In problems where there are no nuisance parameters present, (12) and (13) coincide, so that the endpoint of the test-inversion bootstrap interval has no bootstrap error. This interval is tranformation respecting, and does not require any knowledge of the standard deviation of $\hat{\theta}$.

3.6.4. Disadvantages. This method is only suited to problems in which (13) can be simulated from, at various values of θ. For a two-sided interval, $2B$ (B typically 999, as before) simulations are required, B for each endpoint, so the method is twice as heavy computationally compared to those discussed so far. Clearly, the accuracy of the method depends on the difference between the tails of distributions (12) and (13). In applications [2] this is not nearly as problematic as the performance of the non-Studentized pivotal method might suggest, though.

3.6.5. Coverage error [2]. The coverage error is zero if no nuisance parameters present;

$$\mathbf{P}(\hat{\theta}_{\mathrm{U}} > \theta) = 1 - \alpha + O(n^{-1/2})$$

otherwise.

3.7. Studentized test-inversion bootstrap (STIB) method

3.7.1. Rationale. By analogy with Section 3.2, an obvious proposal to reduce the coverage error of the TIB interval is a replace $\hat{\Theta}$ with a Studentized statistic. Explicitly, suppose we simulate at $(\theta, \hat{\eta})$, obtaining $\hat{\theta}^*$ and its standard deviation $\hat{\sigma}^*$. Then let $T = (\hat{\theta}^* - \theta)/\hat{\sigma}^*$ be the simulated t-statistic and $t_{\mathrm{obs}} = (\hat{\theta}(\mathbf{y}_{\mathrm{obs}}) - \theta)/\hat{\sigma}$ be the 'observed' t-statistic. Then, the endpoint of the Studentized test-inversion bootstrap interval $(-\infty, \tilde{\theta}_{\mathrm{U}})$ satisfies

$$\mathrm{F}_{\mathrm{T}}(t_{\mathrm{obs}}; \tilde{\theta}_{\mathrm{U}}, \hat{\eta}) = \alpha \tag{15}$$

3.7.2. Calculation. Obviously, $\hat{\sigma}$ and $\hat{\sigma}^*$ are needed for this interval; see Section 3.2. However, note that

$$T \leqslant t_{\mathrm{obs}} \Leftrightarrow \hat{\theta}^* \leqslant \frac{(\hat{\theta}(\mathbf{y}_{\mathrm{obs}}) - \theta)\hat{\sigma}^*}{\hat{\sigma}} + \theta$$

The observation enables existing code for the TIB interval to be used to calculate the STIB interval. Note that if $\hat{\sigma}$ is a known function of θ alone, so that $\hat{\sigma}^* = \hat{\sigma} = g(\theta)$, for some g, then $T \leqslant t_{\mathrm{obs}} \Leftrightarrow \hat{\theta}^* \leqslant \hat{\theta}$, so that the STIB interval is equivalent to the TIB interval.

3.7.3. Advantages. As theory suggests, this method is generally more accurate in practical applications than the TIB interval. In the light of experience with the bootstrap-t, it might be supposed that some form of variance stabilizing transformation would improve this method. However, this is not so; in fact, variance stabilising the STIB interval yields the TIB interval [2].

3.7.4. Disadvantages. The first two disadvantages of the TIB interval apply to the STIB interval. In addition, this method is not transformation respecting. Clearly, this method requires $\hat{\sigma}$ and $\hat{\sigma}^*$, which might not be known (see Section 3.2).

3.7.5. Coverage error [2].

$$\mathbf{P}(\tilde{\theta}_U > \theta) = 1 - \alpha + O(n^{-1})$$

3.8. Summary

We conclude this section with Table III which summarizes the properties of the various methods. The explanation of the first six categories is given in the preceeding text.

The seventh category, 'use for functions of parameters', refers to whether the method can be used to construct a confidence interval for a function of parameters estimated from the data. Methods in the pivotal and non-pivotal families can readily be used to do this. For example, suppose we estimate α and β from the data, and require a confidence interval for some function $g(\alpha, \beta) = \theta$, say. Then from the original data, we can calculate $\hat{\alpha}$ and $\hat{\beta}$, giving $\hat{\theta} = g(\hat{\alpha}, \hat{\beta})$. Similarly, from each bootstrap data set, we can calculate $\hat{\alpha}^*$, $\hat{\beta}^*$ and hence $\hat{\theta}^* = g(\hat{\alpha}^*, \hat{\beta}^*)$. We thus obtain the set of $\hat{\theta}^*$'s needed to construct both the pivotal and non-pivotal bootstrap confidence intervals.

4. SOFTWARE

Recall from the linear model example that the appropriate form of bootstrap resampling can vary considerably with the model chosen and the assumptions made about that model. Consequently, it is hard to come up with a general 'bootstrap program' and this has hindered the implementation of bootstrap methods in procedure-based statistical packages, while favouring those which are interactive and relatively easy to develop.

It is not possible to give here an exhaustive description of what is available in the host of statistical software packages on the market. We give instead brief comments on what is currently available in four statistical packages commonly used by medical statisticians; a more exhaustive list is given by Efron and Tibshirani (Reference [6], p. 398). Further developments are taking place all the time – the fact that a package is not mentioned here does not reflect on its suitability or capability for bootstrapping.

The two recent applied books on the bootstrap both come with S-plus [17] code [5, 6], making it the most well provided-for platform for bootstrapping. As mentioned in Section 3.4, stata [18] has a routine that calculates percentile and bias corrected percentile intervals given a user-written problem-specific program; an extension of this routine to include BCa intervals is available from the authors. SAS [19] offers general purpose jackknife and bootstrap capabilities via two macros available in the file jack-boot.sas at http://ftp.sas.com/techsup/download/stat/. In addition, PROC MULTTEST can be used to output bootstrap or permutation resamples, and they can subsequently be analysed with BY processing in any other SAS procedure. Mooney [20] describes various bootstrapping procedures for the interactive software GAUSS [21].

For large coverage simulations, such as the one reported below, it is sometimes necessary to use a lower level language, such as C or FORTRAN, to obtain results in reasonable time. However, such simulations are far from a necessary part of routine bootstrapping for applied problems.

Table III. Summary of properties of bootstrap confidence interval methods. For further details on the categories, see Section 3.8.

Method	Theoretical coverage error	Transformation respecting	Use with parametric simulation	Use with non-parametric simulation	$\hat{\sigma}$, $\hat{\sigma}^*$ required	Analytic constant or variance stabilizing tranformation required	Use for functions of parameters
Non-Studentized pivotal	$O(n^{-1/2})$	×	✓	✓	×	×	✓
Bootstrap-t	$O(n^{-1})$	×	✓	✓	✓	✓	✓
Percentile	$O(n^{-1/2})$	✓	✓	✓	×	×	✓
BC percentile	$O(n^{-1/2})$	✓	✓	✓	×	×	✓
BCa percentile	$O(n^{-1})$	✓	✓	✓	×	✓	✓
Test-inversion	$O(n^{-1/2})$	✓	✓	×	×	×	×
Studentized test-inversion	$O(n^{-1})$	×	✓	×	✓	×	×

Table IV. Number of times, out of 10 000 replications, that equitailed 99 per cent confidence intervals constructed for the mean of an inverse exponential distribution, based on a sample of size 20 from the distribution, fail to include the true parameter value, 1. Figures in parenthesis next to the interval coverage are standard errors based on the binomial distribution.

Method	Interval above 1		Interval below 1		Mean length (SE)	
Expected	50		50		1.333	
Theoretical	45	(7)	52	(7)	1.332	(0.003)
Non-Studentized pivotal	0	(0)	530	(22)	1.157	(0.003)
Bootstrap-t	48	(7)	47	(7)	1.347	(0.003)
Percentile	1	(1)	215	(14)	1.157	(0.003)
BC percentile	5	(2)	169	(13)	1.178	(0.003)
BCa percentile	50	(7)	56	(7)	1.330	(0.003)
Test-inversion	46	(7)	50	(7)	1.351	(0.003)

5. SIMULATIONS STUDY: EXPONENTIAL DISTRIBUTION

We present the results of a simulation study comparing the simulation based methods described so far applied to the problem of estimating a confidence interval for the mean of an inverse exponential distribution, with density function

$$f_X(x) = (1/\lambda)\exp(-x/\lambda), \quad \lambda > 0$$

based on a small sample from the distribution.

Table IV gives results for the bootstrap methods and the test-inversion method. The program repeatedly simulated an 'observed' sample of size 20 from the inverse exponential distribution with mean 1 and constructed intervals for the mean based on that sample. For each 'observed' sample from the inverse exponential distribution all the bootstrap intervals were constructed using the same set of B simulated observations. The TIB interval (which coincides with the STIB interval in this example) was estimated using the Robbins–Monro algorithm [16]. The Robbins–Monro search used $B/2$ simulated observations in the search for each interval endpoint, making a total of B, which is equivalent to the number used to estimate the bootstrap intervals. The value of B was 3999. Standard errors, calculated using the binomial distribution, are given in parentheses.

Table IV shows that the most reliable methods for estimating equitailed confidence intervals for the mean of the inverse exponential distribution are the bootstrap-t and the test-inversion method (both happen to have theoretical coverage error zero in this example). The bias corrected and accelerated (BCa) percentile method cannot be relied on as the nominal coverage increases and the sample size decreases. For example, with a sample size of five, the BCa interval was below 1 a total of 180 times, while the test-inversion and bootstrap-t intervals remained below 1 50–60 times, as expected.

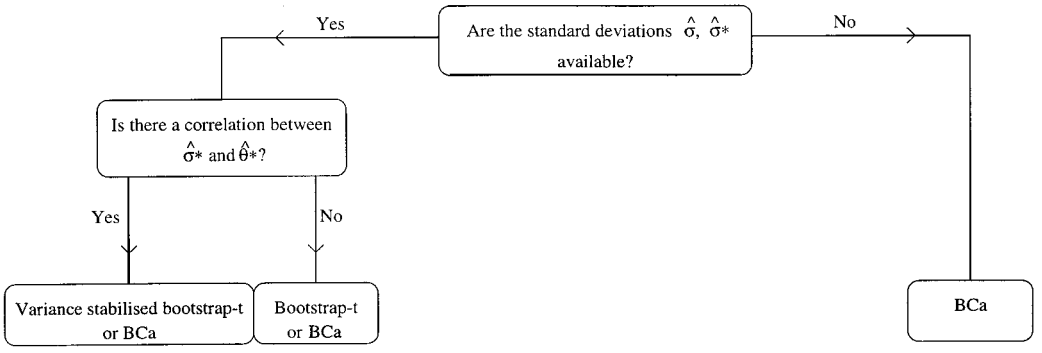

Figure 4. Guide to choosing a bootstrap confidence interval method when using non-parametric simulation.

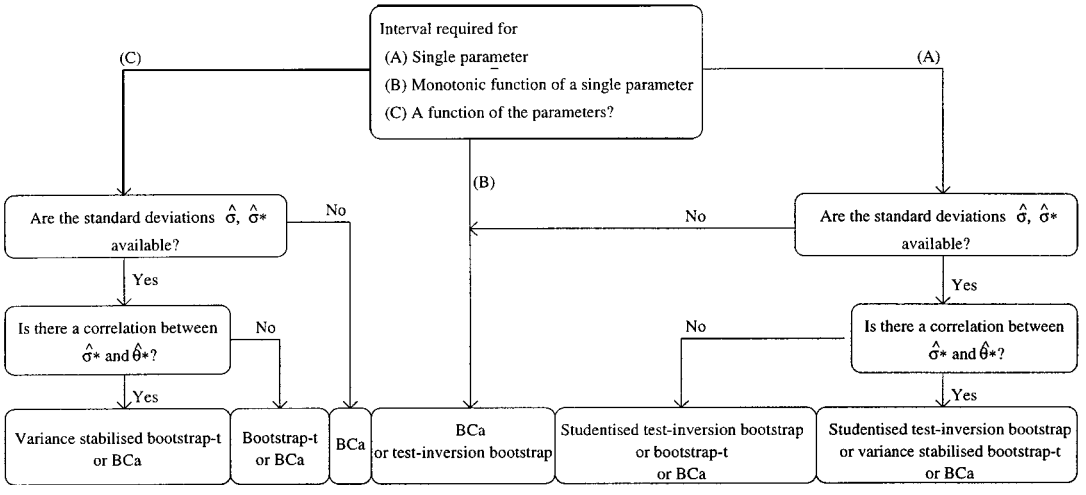

Figure 5. Guide to choosing a bootstrap confidence interval method when using parametric simulation, or simulating in a regression framework, as described in Section 2.3.

6. GUIDE TO CHOOSING A METHOD

The advantages and disadvantages of the above methods, together with our experience on various examples, are summarized by Figures 4 and 5. These are decision trees, which can be used to select an appropriate method for a particular application when using non-parametric and parametric resampling, respectively. Where the box at the end of the decision tree has more than two methods, we slightly prefer the first method given. However, as a check, both methods can be used and the resulting intervals compared.

7. WORKED EXAMPLE: REMISSION TIMES FROM ACUTE MYELOGENOUS LEUKAEMIA

We now continue with our motivating example 1.1. Recall that we fitted a proportional hazards model to the data. This gave an estimate $\hat{\beta} = 0.924$ of the treatment effect β and a standard error $\hat{\sigma} = 0.512$. A standard normal approximation confidence interval for β is therefore $\hat{\beta} \pm 1.96 \times 0.512$, which yields $(-0.080, 1.928)$. We wish to confirm the accuracy of this interval using the bootstrap.

Depending on the statistic of interest, there are a variety of possible bootstrap sampling schemes for survival data (Davison and Hinkley, Reference [5], Chapter 7; Veraverbeke [22]). For simplicity, we consider only non-parametric (that is, case resampling) bootstrapping here. The non-parametric resampling proceeds as described in Section 2.1.

We sample with replacement from the set of 23 triples (t_i, c_i, x_i) (representing the remission time, a binary variable indicating censoring and a binary variable indicating treatment) to obtain a bootstrap data set of 23 triples. We then refit the proportional hazards model to this data to obtain a bootstrap estimate $\hat{\beta}^*$.

The procedure in the previous paragraph was repeated 1999 times. Examination of the results showed that one of the $\hat{\beta}^*$'s was over 20 times greater than the largest of the rest of the $\hat{\beta}^*$'s. This is an example of a problem that occurs from time to time in bootstrap simulation, when an extremely unusual data set is generated. For example, in non-parametric sampling, a bootstrap data set might consist of many copies of the same observation, and it might be impossible to fit the statistical model to such a data set.

Such an extreme data set would have little impact on the bootstrap confidence interval if it were possible to construct the complete set of every possible bootstrap data set and calculate the confidence interval from this. However, this is not possible in practice. In order to avoid the chance occurrence of an extreme data set unduly influencing the bootstrap confidence interval, we therefore advocate discarding such extreme data sets. Of course, if it is impossible to fit the statistical model to more than a few per cent of bootstrap data sets, then, since such data sets can no longer be viewed as extreme data sets within the complete set of bootstrap data sets, this probably indicates that the statistical model is inappropriate.

In this case, we discarded the extreme data set, to avoid problems with the numerical routines used to calculate the variance stabilized bootstrap-t interval.

A Q-Q plot of the 1998 remaining $(\hat{\beta}^* - \hat{\beta})$'s is shown in the upper left panel of Figure 6. It appears to be slightly overdispersed relative to the normal distribution. This overdispersion is still present in the upper tail of the distribution of $(\hat{\beta}^* - \hat{\beta})/\hat{\sigma}^*$, shown in the lower left panel of Figure 6. Moreover, as the top right panel of Figure 6 shows, there is a strong relationship between the value of $\hat{\beta}^*$ and its standard error $\hat{\sigma}^*$. With this information in mind, we now refer to Figures 4 and 5 to choose an appropriate method.

Since we carried out non-parametric simulation, Figure 4 is appropriate. Starting at top, $\hat{\sigma}$ and $\hat{\sigma}^*$ are available from the information matrix of the proportional hazards model fitted to the original and bootstrap data sets, respectively. Next, the top right panel of Figure 6 shows there is a strong correlation between the value of $\hat{\beta}^*$ and its standard error $\hat{\sigma}^*$. We are thus led to the variance stabilized bootstrap-t or, as an alternative, the BCa interval.

These intervals are the last two entries in Table V, and they agree well. The results for various other methods are also given, for comparison. We conclude that there is not quite enough

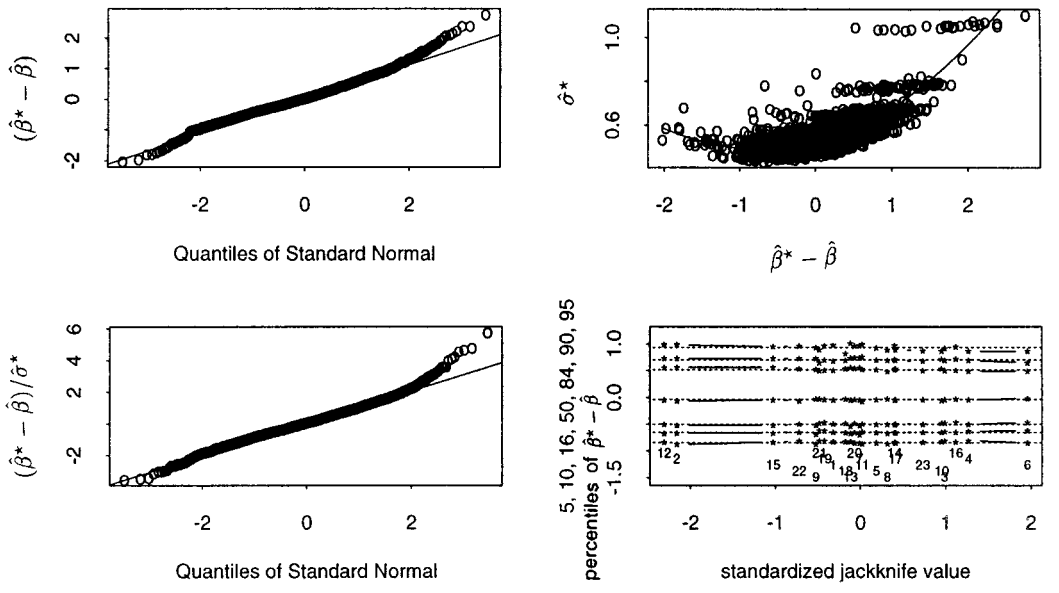

Figure 6. Diagnostic plots for the bootstrap confidence intervals for the AML data: top left panel, Q-Q plot of the distribution of $\hat{\beta}^* - \hat{\beta}$; bottom left panel, Q-Q plot of the distribution of $(\hat{\beta}^* - \hat{\beta})/\hat{\sigma}^*$; top right panel, plot of σ^* against $\hat{\beta}^* - \hat{\beta}$; bottom right, jackknife-after-bootstrap plot – details in the text.

Table V. Ninety-five per cent bootstrap confidence intervals for treatment effect, β, obtained using various methods. The maximum likelihood estimate of β, $\hat{\beta} = 0.916$, with standard error, calculated in the usual way from the information matrix, of 0.512.

Method	95% interval
Normal approx. ($\hat{\beta} \pm 1.96 \times \text{SE } \hat{\beta}$)	$(-0.088, 1.920)$
Non-Studentized pivotal	$(-0.335, 1.856)$
Bootstrap-t	$(-0.148, 1.883)$
Percentile	$(-0.0253, 2.166)$
BCa	$(-0.159, 2.011)$
Variance stabilized bootstrap-t	$(-0.170, 2.067)$

evidence to reject the hypothesis that the two treatments are different at the 5 per cent level, and in particular that there is less evidence than suggested by the standard normal theory interval.

Before leaving this example, we perform a *bootstrap diagnostic*, to examine whether the conclusions depend heavily on any single observation. To do this we look at the estimated distribution of $\hat{\beta}^* - \hat{\beta}$ when observations 1, 2, … , 23 are left out in turn. No additional

calculations are needed for this; we simply note, as we go along, which observations are included in which bootstrap samples. Then, the $\hat{\beta}$*'s arising from those bootstrap samples which do not include observation 1 can be used to estimate the distribution of $\hat{\beta}^* - \hat{\beta}$ were observation 1 not to have been observed, and so on. The results are shown in the lower right panel of Figure 6. They are plotted against jackknife estimates of the influence values for the 23 observations (Davison and Hinkley, Reference [5], p. 113). These are the difference between the mean of the $\hat{\beta}$*'s in the bootstrap samples in which a particular observation does not occur and the mean of all the $\hat{\beta}$*'s. The results indicate that the estimate of the distribution of $\hat{\beta}^* - \hat{\beta}$, and hence the confidence intervals, do not depend strongly on a particular observation.

8. CONCLUSIONS

In common with Young [8], our experience suggests there are three principal reasons for the somewhat reticent use of the bootstrap in the medical statistics community. The first is a fear that in a particular problem, the bootstrap may 'fail' – in some sense – resulting in misleading inference. The second is uncertainty about which of the plethora of techniques available is appropriate in a particular problem. The third is uncertainty over how to implement the methods in practice.

We have sought to address these issues under the following headings: *when* should bootstrap confidence intervals be used; *which* method should be used, and *what* should be done to implement the method. In summary, our answers to these questions are as follows:

We suggest that bootstrap confidence intervals should be used whenever there is cause to doubt the assumptions underlying parametric confidence intervals. They will either validate such assumptions, or avoid misleading inferences being drawn. With regard to the anxiety that the bootstrap may unexpectedly 'fail' in a particular application, reassurance can be drawn from Young [8]:

> 'Undoubtedly, the bootstrap has been successfully demonstrated to be a sensible approach for … confidence interval construction in many statistical problems.'

While there exist examples where the bootstrap will fail, these are generally pathological [5, 6, 8], and hence unlikely to cause trouble in practice. Concerned readers should consult Davison and Hinkley (Reference [5], p. 37 ff).

Turning to the second issue, we have discussed the advantages and disadvantages of all the commonly used methods, and combined this information into two decision trees to help choose a method appropriate to the problem in hand.

In response to the third issue, we have given a detailed description of the implementation of each method and drawn attention to the issues that need to be considered in deciding between parametric and non-parametric simulation.

ACKNOWLEDGEMENTS

This work was partly carried out while James Carpenter was supported by the National Radiological Protection Board, Didcot, Chilton. Thanks are due to Stuart Pocock for helpful discussions, and the referees for their constructive criticism.

REFERENCES

1. Embury SH, Elias L, Heller PH, Hood CE, Greenberg PL, Schrier SL. Remission maintenance therapy in acute myelogenous leukaemia. *Western Journal of Medicine* 1977; **126**:267–272.
2. Carpenter JR. Test-inversion bootstrap confidence intervals. *Journal of the Royal Statistical Society, Series B* 1999; **61**:159–172.
3. LePage R, Billard L (eds). *Exploring the Limits of Bootstrap*. Wiley: New York, 1992.
4. Armitage P, Berry G. *Statistical Methods in Medical Research*. Blackwell: Oxford, 1987.
5. Davison AC, Hinkley DV. *Bootstrap Methods and their Application*. Cambridge University Press: Cambridge, 1997.
6. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman and Hall: London, 1993.
7. Cox DR, Hinkley DV. *Theoretical Statistics*. Chapman and Hall: London, 1974.
8. Young GA. Bootstrap — more than a stab in the dark? *Statistical Science* 1994; **9**:382–415.
9. Hall P. *The Bootstrap and Edgeworth Expansion*. Springer-Verlag: London, 1992.
10. DiCiccio TJ, Efron B. Bootstrap confidence intervals. *Statistical Science* 1996; **11**:189–212.
11. Efron B. *The Jackknife, the Bootstrap and other Resampling Plans*. Society for Industrial and Applied Mathematics: Philadelphia, 1982.
12. Efron B. Better bootstrap confidence intervals. *Journal of the American Statistical Association* 1987; **82**:171–200.
13. Efron B. More accurate confidence intervals in exponential families. *Biometrika* 1992; **79**:231–245.
14. Rao CR. *Linear Statistical Inference and its Applications*. Wiley and Sons: London, 1973.
15. Carpenter JR. Simulated confidence regions for parameters in epidemiological model. D. Phil. thesis, Oxford University, 1997.
16. Buckland ST, Garthwaite PH. Estimating confidence intervals by the Robbins–Monro search process. *Journal of the Royal Statistical Society, Series C* 1990; **39**:413–424.
17. MathSoft, Inc., 101 Main Street, Cambridge, MA 02142-1521 USA.
18. Stata Corporation, 702 University Drive East, College Station, Texas 77840 USA.
19. SAS Institute Inc. SAS Campus Drive Cary, NC 27513, USA.
20. Mooney C. *Bootstrapping: A Non-parameteric Approach to Statistical Inference*. SAGE: London, 1993.
21. Aptech Systems Inc., Maple Valley, WA, USA.
22. Veraverbeke N. Bootstrapping in survival analysis. *South African Statistical Journal* 1997; **31**:217–258.
23. Efron B. Nonparametric standard errors and confidence intervals (with discussion). *Canadian Journal of Statistics* 1981; **9**:139–172.
24. Kabaila P. Some properties of profile bootstrap confidence intervals. *Australian Journal of Statistics* 1993; **35**:205–214.

