The Bootstrap

Philip M. Dixon

Department of Statistics

Iowa State University

20 December 2001

The bootstrap is a resampling method for statistical inference. It is commonly used to estimate confidence intervals, but it can also be used to estimate bias and variance of an estimator or calibrate hypothesis tests. A short of papers illustrative of the diversity of recent environmentric applications of the bootstrap includes toxicology [2], fisheries surveys [27], groundwater and air polution modelling [1, 4], chemometrics [35], hydrology [14], phylogenetics [23], spatial point patterns [33], ecological indices [9], and multivariate summarization [24, 38].

The literature on the bootstrap is extensive. Book length treatments of the concepts, applications, and theory of the bootstrap range in content from those that emphasize applications [19], to comprehensive treatments [13, 5, 3], to those that emphasize theory [11, 15, 18, 28]. Major review papers on the bootstrap and its applications include [12, 8, 37, 7]. Papers describing the bootstrap and demonstrating its use to non statisticians have been published in many different journals. Extensive bibliographies, listing applications, are included in [19] and [3].

This article can not duplicate the comprehensive coverage found in these books and papers. Instead, I will illustrate bootstrap concepts using a simple example, describe different types of bootstraps and some of their theoretical and practical properties, discuss computation and other details, and indicate extensions that are especially appropriate for environmentric data. The methods will be illustrated using data on heavy metal concentrations in ground water [22] and magnesium concentration in blood (see Jackknife Resampling for details of the second data set).

BOOTSTRAP CONCEPTS

Consider estimating the mean concentration of a heavy metal, e.g. Copper in groundwater from San Joaquin valley basin soils [22]. As is typical with environmental chemistry data, some of the values are left censored.

They are reported as 'less than detection limit', with a specified value for the detection limit. Often, observations are skewed. Point estimates of the mean, μ, and the standard deviation, σ, can be calculated using a variety of different methods. It is more difficult to do statistical inference, e.g. calculate a 95% confidence interval for the mean. The usual confidence interval, based on a Student's t distribution, is not appropriate because of the censoring and skewness. Inference based on maximum-likelihood estimators relies on an asymptotic distribution, which may not be appropriate for small samples. The difficulty is that the sampling distribution of the estimate is unknown. The bootstrap uses the data and computer power to estimate that unknown sampling distribution.

Given a set of independent and identically distributed observations, X_i, $i = 1..n$, a parameter that can be defined as some function, $\theta = T(x)$, of the values in the population, and a statistic that is the same function of the observations, $\hat{\theta} = T(X)$, the bootstrap estimates the sampling distribution, $F_\theta(x)$, of that function. The data are used as an estimate of the unknown cdf, $F_x(x)$, of values in the population. Bootstrap samples are repeatedly drawn from the estimated population. The function (e.g. the mean) is evaluated for each bootstrap sample, giving a set of bootstrap values, $\{\hat{\theta}_i^B\}$, $i = 1..m$. The empirical distribution of those bootstrap values, $\hat{F}_b(x)$, estimates the theoretical sampling distribution, $F_\theta(x)$.

The bootstrap distribution, $\hat{F}_b(x)$, is used to estimate bias, estimate a standard error, or construct a confidence interval for the statistic of interest. The bootstrap estimates of bias, B_b, and standard error, s_b, are the empirical estimates calculated from m bootstrap values:

$$\begin{aligned} B_b &= \Sigma_{i=1}^m(\hat{\theta}_i^B - \hat{\theta})/m. \\ s_b &= \left[\Sigma_{i=1}^m((\hat{\theta}_i^B - \overline{\hat{\theta}^B})^2/(m-1)\right]^(1/2). \end{aligned}$$

The percentile confidence interval method uses the $\alpha/2$ and 1-$\alpha/2$ quantiles of $\hat{F}_b(x)$ as a 1-α level confidence interval for the parameter.

There are 49 observations, including 14 censored values, in the San Joaquin valley Copper data. Because the data are quite skewed, the mean is estimated using a non-parametric estimator [29]. The estimated mean is 4.33 ppm. A 95% confidence interval for the mean is estimated using 1000 bootstrap samples. Each bootstrap sample is a simple random sample of 49 values selected **with replacement** from the original observations. Because a bootstrap sample is drawn with replacement, some of the original observations are repeated more than once in the bootstrap sample. Other observations are omitted from an individual bootstrap sample. The statistic is estimated for each bootstrap sample. Bootstrap confidence intervals can be computed from the set of bootstrap values in a variety of ways (see A MENAGERIE OF BOOTSTRAP CONFIDENCE INTERVALS below). The simplest is the percentile bootstrap confidence, where the endpoints of the 95% confidence interval are given by

the 25'th and 975'th sorted bootstrap values [13, p. 160]. For these data, that interval is (3.05, 5.77).

The percentile bootstrap illustrated here is one of the simplest bootstrap confidence intervals methods, but it is may not be the best method in all applications. In particular, the percentile interval may not have the claimed coverage. Confidence interval coverage is the probability that the confidence interval includes the true parameter, under repeated sampling from the same underlying population. When the coverage is the same as the stated size of the confidence interval (e.g. coverage = 95% for a 95% confidence interval), the intervals are accurate. Empirical and theoretical studies of coverage have shown that the percentile interval is accurate in some situations, but not others [13, 5].

A MENAGERIE OF BOOTSTRAP CONFIDENCE INTERVALS

The percentile bootstrap has been extended in many different ways to increase confidence accuracy. The varieties of bootstraps differ in

1. how confidence interval endpoints are calculated (e.g. percentile, basic, accelerated, studentized, or BCA bootstrap),
2. how the population is approximated (non-parametric or parametric bootstrap), and
3. how bootstrap samples are selected (ordinary, balanced, or moving-block bootstrap).

Each of these is discussed in the following sections.

Calculating confidence interval endpoints

The percentile bootstrap endpoints are simple to calculate and can work well, especially if the sampling distribution is symmetrical. The percentile bootstrap confidence intervals may not have the correct coverage when the sampling distribution is skewed [5]. Other methods adjust the confidence interval endpoints to increase the accuracy of the coverage (Table 1). One confusing aspect of these methods is that some methods have been given different names by different authors. A synonymy is given in the documentation to the SAS JACKBOOT collection of macros [26].

Coverage of the percentile bootstrap can be improved by adjusting the endpoints for bias and non-constant

variance (the accelerated bootstrap) [5]. Computing the accelerated bootstrap confidence interval requires estimating a bias coefficient, z_0, and an acceleration coefficient, a. Both coefficients can be estimated nonparametrically from the data [13, p. 186] or theoretically calculated for a specific distribution [5, p. 205]. Confidence interval endpoints are obtained by inverting percentiles of the bootstrap distribution. Adjusting for bias and acceleration shifts the percentiles used to find the confidence interval endpoints. Because endpoints of the confidence interval are obtained by inverting the bootstrap distribution, both the percentile and accelerated bootstraps preserve the range of the parameter. For example, if the parameter and statistic are constrained or constrained to lie between 0 and 1 the endpoints of these confidence intervals will satisfy that constraint.

Insert table 1 near here

The quadratic ABC confidence intervals [6, 7] are an approximation to the accelerated bootstrap that do not require many bootstrap simulations, which could be helpful when parameter estimation requires considerable computation. The three required coefficients, a, b, and c (Table 1), are calculated either from the observations or a model [5, pp. 214-220]. Endpoints of the confidence interval are calculated by a Taylor-series approximation to $F_b(x)$. Because of the approximation, the endpoints may not statisfy constraints on the parameter space, unlike the first three methods.

The basic and studentized bootstraps calculate endpoints by inverting hypothesis tests [5]. In both, the upper quantile of a bootstrap distribution is used to calculate the lower confidence bound and the lower quantile is used to calculate the upper bound. When the the bootstrap distribution is symmetrical around the estimate from the original data, i.e. $\hat{\theta} - \hat{F}_b^{-1}(1-\alpha) = \hat{F}_b^{-1}(\alpha) - \hat{\theta}$, the basic bootstrap produces the same endpoints as the percentile bootstrap. When the distribution is skewed, the endpoints of the two methods differ. Neither the basic nor the studentized bootstrap constrains confidence interval endpoints to fall within a bounded parameter space.

The studentized bootstrap is based on a different bootstrap distribution than the other bootstraps. The estimate, $\hat{\theta}_i$, and its standard error, $s_{\hat{\theta}_i}$, from each bootstrap sample are used to calculate studentized estimates, $t_i = (\hat{\theta}_i - \hat{\theta})/s_{\hat{\theta}_i}$, where $\hat{\theta}_X$ is the estimate calculated from the original data set. The $1-\alpha/2$ and $\alpha/2$ quantiles of this distribution, $\hat{F}_s(x)$, are used to calculate the confidence interval. The endpoints of the studentized bootstrap confidence interval have a natural interpretation. They are like the 'usual' confidence intervals based on a Student's t-statistic, except that the data is used to estimate a more appropriate distribution for the 't' statistic. The studentized bootstrap distribution requires a standard error for each bootstrap sample. The jackknife or a second, nested bootstrap can be used if the standard error can not be estimated any other way.

The use of the studentized bootstrap is somewhat controversial. To some, the endpoints of the intervals seem too wide and the method seems to be sensitive to outliers [13]. For others, the studentized bootstrap seems to be the only bootstrap with reasonable confidence interval coverage in difficult problems.

Bootstrap confidence intervals may not have the claimed coverage when computed from small samples. Details, e.g. whether the empirical coverage is too large or too small and whether coverage is better in one tail than the other, depend on the statistic being evaluated and characteristics of the population being sampled. Numerous studies have evaluated bootstrap coverage for specific cases. Citations and examples are discussed in [13, 5, 3]. Bootstrap iteration provides a way to improve the coverage of a confidence interval, at the cost of additional computing [20].

Approximating the population

At the heart of the bootstrap is the concept that the distribution of the statistic of interest, $F_\theta(x)$, can be approximated by estimates from repeated samples from an approximation to the unknown population. The population can be approximated in different ways, each of which leads to a different type of bootstrap. The most common approximations lead to the parametric and nonparametric bootstraps. Less frequently used approximations lead to the smoothed and generalized bootstraps.

The parametric bootstrap assumes that $F_x(x)$ is known except perhaps for one or more unknown parameters, ψ. For example, $F_x(x)$ might be known (or assumed) to be log-normal with unknown parameters, μ and σ^2. $\hat{F}_x(x)$ is approximated by substituting estimates of $\hat{\psi}$ for the unknown parameters, ψ. Often these estimates are maximum likelihood estimates, but other estimates could also be used. The generalized bootstrap [10] is a parametric bootstrap where $F_x(x)$ is a flexible distribution with many (often four) parameters, e.g. the generalized lambda distribution.

The nonparametric bootstrap is the bootstrap described previously. The population $F_x(x)$ is approximated by the empirical distribution of the observed values, a multinomial distribution. Nonparametric bootstrap samples include repeats of many observations, which may lead to inconsistent estimators. The smoothed bootstrap [30] approximates $F_x(x)$ as a smoothed version of the empirical cdf. Smoothed bootstrap samples are generated by sampling observations with replacement and jittering each bootstrap observation by adding a small amount of random noise. Usually the noise distribution is normal with mean 0 and a small variance. Increasing the variance increases the amount of smoothing. When the sample space is constrained, a slightly different smoothing

procedure can generate bootstrap observations that satisfy the constraint [31].

Selecting bootstrap samples

In the ordinary nonparametric bootstrap described above, each bootstrap sample is a simple random sample, with replacement, of the observations. The bootstrap samples are a subset of all possible samples of size n from a finite population with n copies of each observation. Hence, the bootstrap estimates of bias, standard error, and confidence interval endpoints are random variables. Their variance can be reduced by increasing the number of bootstrap samples [13] or by using more complex sampling methods [5, pp 437-487].

The balanced bootstrap is an alternative sampling method that can increase the precision of the bootstrap bias and standard error. The balanced bootstrap forces each observation to occur a total of n_B times in the collection of n_B bootstrap samples. This does not force each bootstrap sample to contain all observations; the first observation may occur twice in the first bootstrap sample and not at all in the second, while the second observation may occur once in each sample. Balanced bootstrap samples can be generated by constructing a population with n copies of each of the n observations, then randomly permuting that population. The first n permuted values are the first bootstrap sample, the second n permuted values are the second sample, and so on. While balancing often decreases the variance of the estimated bias and s.e., it appears to be less useful for estimating confidence interval endpoints.

The moving blocks [17] and moving tiles bootstraps extend the bootstrap to correlated data [5, pp 396-408]. The ordinary bootstrap assumes that observations are independent, which may not be appropriate for time series data, spatial data or other correlated data. In a moving blocks bootstrap for time series data, the series of observations is divided into b non-overlapping blocks of l sequential observations. The bootstrap sample is constructed by randomly sampling b blocks with replacement and concatenating them into a series of bl observations. Correlation between observations is assumed to be strongest within a block and relatively weak between blocks. The choice of l is crucial. If l is large, b is small and there may be very few unique bootstrap samples. If l is small, observations in different blocks may not be independent. Even if l is appropriately chosen, the correlation between observations in the bootstrap sample is less than that in the original sample because blocks are assumed to be independent. Bootstrapping spatial data using moving tiles is similar [5]. The stationary bootstrap [25] is a variant of the moving blocks bootstrap with random block lengths. Model based approaches to bootstrapping correlated data are described in the next section.

EXTENSIONS TO NON-IID DATA

The bootstrap methods in the previous two sections are appropriate for a single sample of iid observations. Many problems involve observations that are not iid. These include regression problems, temporally or spatially correlated data, and hierarchical problems.

Regression and multi-sample data

One common source of non-iid observations is when they are presumed to come from some linear or non-linear model with additive errors. This includes two sample problems, regression problems, or more complicated models for designed experiments. The quantities of interest could be the difference in two means, the slope or intercept of a regression, some parameter in the model, or a function of any of these. One environmetric application is the use of the bootstrap to estimate RI_{50}, the toxicant concentration that reduces reproductive output by 50% [2]. This is a function of the parameters of a polynomial regression; the bootstrap can be used to estimate a confidence interval for RI_{50} [2].

There are two general approaches for such data: 1) bootstrapping the observations, also called case resampling and 2) bootstrapping the residuals, also called error resampling [13, pp. 113-115], [3, pp 76-78], [5, 261-266]. Consider a set of observations presumed to arise from a linear model, $Y_i = \boldsymbol{X_i}\boldsymbol{\beta} + \varepsilon_i$. Each observation is a vector of covariate values and a response, $(\boldsymbol{X_i}, Y_i)^T$. If observations are bootstrapped, the entire vector is resampled with replacement. The moments and distribution of covariate values is not fixed in all the bootstrap samples. When the data are grouped, as in a two sample problem, it is customary to condition on the number of observations in each group. Bootstrapping the observations requires separately resampling each group of observations.

Bootstrapping the residuals is a three step process. Residuals, $\hat{\varepsilon}_i = Y_i - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}$, are calculated for each observation. Then a bootstrap sample of residuals, $\{\varepsilon_i^B\}$, is drawn with replacement from the observed residuals. The bootstrap sample of observations is constructed by adding a randomly sampled residual to the original predicted value for each observation: $y_i^B = \boldsymbol{X_i}\hat{\boldsymbol{\beta}} + \varepsilon_i^B$.

Bootstrapping the observations and bootstrapping the residuals are not equivalent in small samples, but they are asymptotically equivalent [12]. The choice of bootstrap depends on the goal and context of the analysis. Bootstrapping the residuals maintains the structure of the covariates, but the bootstrap inference assumes that

the original model (used to calculate the residuals) is appropriate. Bootstrapping the observations repeats some covariate values and omits others. It is the usual choice when the analysis includes some aspect of model selection.

Correlated data

The dichotomy between bootstrapping the observations and boostrapping the residuals recurs with time series and spatial data. The moving blocks and moving tiles bootstraps, discussed above, are analogous to bootstrapping observations. Neither assumes a specific model for the data. Bootstrapping residuals requires fitting a model. For time series data, the model is often an ARMA model, but it could be a state-space model [16]. For spatial data, the model specifies the mean and correlation structure of the observations. Consider spatial data that is assumed to follow the model

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}. \tag{1}$$

One approach is to estimate $\hat{\boldsymbol{\Sigma}}$, the variance-covariance matrix of the errors, $\boldsymbol{\varepsilon}$, then calculate the Cholesky decomposition, $\hat{\boldsymbol{L}}$ such that $\hat{\boldsymbol{\Sigma}} = \boldsymbol{L}\,\boldsymbol{L'}$ [32]. The estimated errors, $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Z}$ - $\boldsymbol{X}\hat{\boldsymbol{\beta}}$, can be whitened by premultiplying by $\boldsymbol{L^{-1}}$, i.e. $\hat{\boldsymbol{e}} = \boldsymbol{L^{-1}}\hat{\boldsymbol{\varepsilon}}$. A bootstrap sample of spatially correlated observations is constructed by drawing a bootstrap sample of the whitened residuals, $\{e^B\}$, introducing the correlation structure and restoring the mean, i.e. $\boldsymbol{Y^B} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{L}\,\boldsymbol{e^B}$. The distribution of the statistic of interest is estimated by the empirical distribution of the statistic in many bootstrap samples.

Hierarchical data

Environmetric data often include multiple sources of variation that can be described using a hierarchical model. When the model is sufficiently simple (e.g., a linear mixed model in which all random effects have a normal distributions with constant variance), the data can be expressed in the form of equation 1 with a variance-covariance matrix, $\boldsymbol{\Sigma}$, that depends on the variance components. The model-based bootstrap of residuals described above can be used to generate bootstrap samples. If, in addition, the distributions of all random effects are specified, a parametric bootstrap can be used [5, p. 100]. One example of a hierarchical parametric bootstrap for a complicated model is the construction of a confidence region for an evolutionary trajectory [21].

Nonparametric bootstrapping of hierarchical data is complicated by the need to estimate empirical distribution functions for two (or more) random variables. A simple example of the difficulty in constructing empirical distributions with the correct first and second moments is given in [5, pp. 100-101]. Although procedures can

be derived for specific cases, there is currently no general non-parametric method for bootstrapping hierarchical data.

BOOTSTRAP THEORY

Bootstrap theory is an active area of statistical research. Detailed accounts of the theory of various forms of the bootstrap can be found in [11, 15, 28, 5]. I provide a short introduction, without proofs, to the theory for a single parameter, estimated from a single sample of independent, identically distributed observations. Details and proofs can be found in [28]. Extensions to more complicated

Asymptotic properties can be derived for many different types of statistics. One of the most general approaches considers statistics that can be expressed as a functional, T, of an empirical distribution of n observations, F_n, i.e. statistics that can be written as $T_n = T(F_n)$. The parameter is $\theta = T(F)$, where F is the distribution function of the population. Given an appropriate differentiability of the functional, T, and a bounded second moment for the influence function of T, then the bootstrap distribution, $F_B(x)$ is a consistent estimator of the true sampling distribution, $F_\theta(x)$ [28, pp. 80-86]. Given an extra constraint on the tails of the distribution of T_n, then the bootstrap estimate of variance, s_b^2, is a consistent estimator of the sampling variance of the parameter θ.

Coverage accuracy, where coverage is the probability that a confidence interval includes θ, is the important property for a confidence interval procedure. Lower and upper bounds are considered separately, but their asymptotic properties are similar. Bootstrap confidence intervals methods differ in their asymptotic properties. Percentile intervals are first order accurate, i.e. $\mathrm{P}[\theta < \hat{\theta}_\alpha] \approx \alpha + O(n^{-1/2})$, where $\hat{\theta}_\alpha$ is the estimated lower bound of a $1-2\alpha\%$ two-sided confidence interval [13, p 187]. Both the studentized and BC_a intervals are second order accurate, i.e. i.e. $\mathrm{P}[\theta < \hat{\theta}_\alpha] \approx \alpha + O(n^{-1})$ [13, p. 187].

Another comparison of confidence interval procedures is the relationship between $F_\theta(x)$ and a normal distribution. If the sampling distribution, $F_\theta(x)$, is normal with known variance, then confidence intervals based on z-scores have the desired coverage, and bootstrapping isn't necessary. The percentile bootstrap limits are correct if $F_\theta(x)$ can be transformed to normality. In others words, there exists some monotone $g(x)$ such that $\hat{\phi} = g(\hat{\theta}) \sim \mathrm{N}(\phi,\, \tau^2)$, where $\phi = g(\theta)$ and τ^2 is a constant variance. Other bootstrap confidence interval procedures are correct under more general models for the distribution of $\hat{\phi}$. For example, the BC_a intervals are correct if $\hat{\phi} \sim N(\phi - z_0\, \tau_\phi, \tau_\phi^2)$ where $\tau_\phi = 1 + a\, \phi$, z_0 is a bias correction coefficient and a is an acceleration coefficient

[12, p. 68-69].

COMPUTATION

A bootstrap can be implemented wherever there is the ability to generate uniform random numbers and draw a random samples of observations [36, 9]. Macros and functions in various statistical packages include the more complicated confidence interval calculations. These include the JACKBOOT macro in SAS [26] and various libraries of Splus functions [13, 34, 5]. All macros and libraries can bootstrap a single sample of observations and compute bias, s.e., and a variety of confidence intervals. Some packages (e.g. the boot() library [5]) can be easily extended for multiple sample problems. In the example below, it is useful to force each bootstrap sample to contain 38 observations from one area and 52 from the second. This can be done by specifying strata. Some packages also include diagnostic methods [5].

EXAMPLE

The extended example will illustrate many different types of bootstrap confidence intervals and the relationship between Jackknife Resampling and the bootstrap. The data are part of a study of heavy metal loading in children, where the goal is to describe the relationship between creatinine and magnesium concentrations in urine. Creatinine and magnesium concentrations were measured on 38 children from a relatively contaminated area (Kapfenberg) and 52 children in a less contaminated area (Knittelfeld) of Styria, Austria. The data are plotted as Figure 2 in Jackknife Resampling. The jackknife analysis in that article considers four statistics: ρ, the correlation between creatinin and magnesium, β_1 and β_2, the slopes of a regression of magnesium on creatinine for each area, and β_1/β_2, their ratio. The slopes, β_1 and β_2 are defined by a heterscedastic linear regression with different parameters for each group of children

$$\begin{aligned} M_i &= \mu_i + \varepsilon_i \\ \mu_i &= \begin{cases} \alpha_1 + \beta_1 C_i & \text{for children in Kapfenberg} \\ \alpha_2 + \beta_2 C_i & \text{for children in Knittelfeld} \end{cases} \\ \varepsilon_i &\sim \mathrm{N}(0, \mu_i), \end{aligned}$$

where M_i and C_i are the blood magnesium and creatinine concentrations. The model can be fit using the glm() function in Splus. Here, the analysis is repeated using the bootstrap. The boot() library [5] of functions in Splus

was used for the computations.

The sample correlation between creatinine and magnesium, treating all children as one sample of 90 observations, is 0.409. The bootstrap distribution of the correlation coefficient, $\hat{F}_b(x)$ (Figure 1a), is estimated from 1000 bootstrap samples, each with 90 observations. That distribution is very slightly skewed. The estimated bias and standard error (Table 2) are similar to those computed using various forms of the jackknife (compare to Table 1 in Jackknife resampling). 95% confidence intervals for the correlation coefficient were constructed using four bootstrap methods (Table 3). The studentized bootstrap intervals were not calculated for ρ because the jackknife estimate of variance was very computer intensive. The endpoints are quite similar to each other, although those from the BC_a method are slightly different from the others. I would choose the BC_a interval because it makes the most general assumptions, it has the best asymptotic properties, and the data set is large enough to provide a reasonable estimate of a, the acceleration constant.

Insert Figure 1 near here.

There are many ways to bootstrap in a regresion problem (see Regression and multi-sample data section). The appropriate choice depends on how the data were collected. I assumed the number of children in each area was fixed in the design, but the distribution of X values (blood creatinine levels) were not. Hence, it is appropriate to bootstrap observations (not residuals) and specify strata to force each bootstrap sample to include 38 children from Kapfenberg and 52 from Knittelfeld.

The bootstrap distributions of β_1 (Figure 1b) and β_2 (Figure 1c) are reasonably symmetrical. Again, bootstrap estimates of bias and standard errors (Table 2) are quite close to the delete-1 jackknife estimates (compare to Table 3 in Jackknife resampling). The bootstrap standard errors are slightly (ca. 5%) smaller than the jackknife standard errors, possibly because the bootstrap samples are forced to have 38 and 52 children from the two areas. Endpoints of four confidence interval procedures are quite similar (Table 3). Although the studentized intervals for β_1 are slightly wider than the other three intervals for β_1, the studentized intervals for β_2 are slightly shorter than the other three intervals for β_2. I would chose the studentized intervals here, but there is little practical difference between any of intervals.

The bootstrap distribution of the ratio, β_1/β_2 is skewed (Figure 1d), so one might expect to find differences among the confidence interval procedures. Both the lower and upper endpoints for the basic interval are much smaller than those for the other three intervals. The endpoints of the BC_a and the studentized intervals are almost identical. I would chose either of those intervals.

Although the bootstrap may seem to perform magic, in the sense that it permits statistical inference in very general circumstances, it is not a substitute for data. The performance of the bootstrap depends on the sample size. It isn't possible to recommend minimum sample sizes, because each problem is different. However, increasing the number of bootstrap replicates or using a more sophisticated bootstrap procedure does not compensate for insufficient data. All the bootstrap can do is (approximately) quantify the uncertainty in the conclusion.

TABLES

Method	α-level Endpoint:	Range preserving?
Percentile	$\hat{F}_b^{-1}(\alpha)$	Yes
Accelerated	$\hat{F}_b^{-1}\left(\Phi(z_0 + \frac{z_0+z^\alpha}{1-a(z_0+z^\alpha)}\right)$	Yes
ABC	$\hat{\theta} + s(z^\alpha + a + c - bs + (2a+c)z_\alpha^2)$	No
Basic	$2\hat{\theta} - \hat{F}_b^{-1}(1-\alpha)$	No
Studentized	$\hat{\theta} - s_\theta \hat{F}_s^{-1}(1-\alpha)$	No

Table 1: Methods for estimating endpoints of bootstrap α-level confidence intervals. $\hat{\theta}$ is the observed estimate, $\hat{F}_b(x)$ is the bootstrap cdf, $\hat{F}_s(x)$ is the studentized bootstrap cdf, $\Phi(x)$ is the normal cdf, $z_0 = \Phi^{-1}(F_b(\hat{\theta}))$, a is the acceleration constant, and z^α is the α-percentile of a standard normal distribution.

Statistic	Estimate	Bias	s.e.
Correlation	0.409	0.00543	0.0752
Slope, Kapfenberg	232.9	2.93	38.5
Slope, Knittelfeld	105.2	-1.28	13.7
Ratio of slopes	2.215	0.056	0.508

Table 2: Parameter estimates, bootstrap estimate of bias, and bootstrap estimate of standard error for 4 quantities describing the relationship between urine creatinine and magnesium concentrations.

Statistic	Bootstrap c.i. method	95% confidence interval
Correlation	Percentile	(0.252, 0.553)
	Basic	(0.265, 0.566)
	BCa	(0.227, 0.525)
	ABC	(0.252, 0.542)
Slope, β_1	Percentile	(166.0, 312.1)
Kapfenberg	Basic	(153.8, 300.0)
	BCa	(162.8, 308.6)
	Studentized	(143.8, 304.1)
Slope, β_2	Percentile	(73.9, 128.9)
Knittelfeld	Basic	(81.4, 136.4)
	BCa	(75.8, 130.9)
	Studentized	(79.9, 134.7)
Ratio of slopes	Percentile	(1.44, 3.44)
	Basic	(0.99, 2.99)
	BCa	(1.48, 3.55)
	Studentized	(1.47, 3.35)

Table 3: Endpoints of 95% confidence intervals for 4 quantities describing the relationship between urine creatinine and magnesium concentrations. Confidence intervals are computed using the percentile, basic, and BCa methods. The studentized bootstrap is included when it could be calculated easily.

Figures

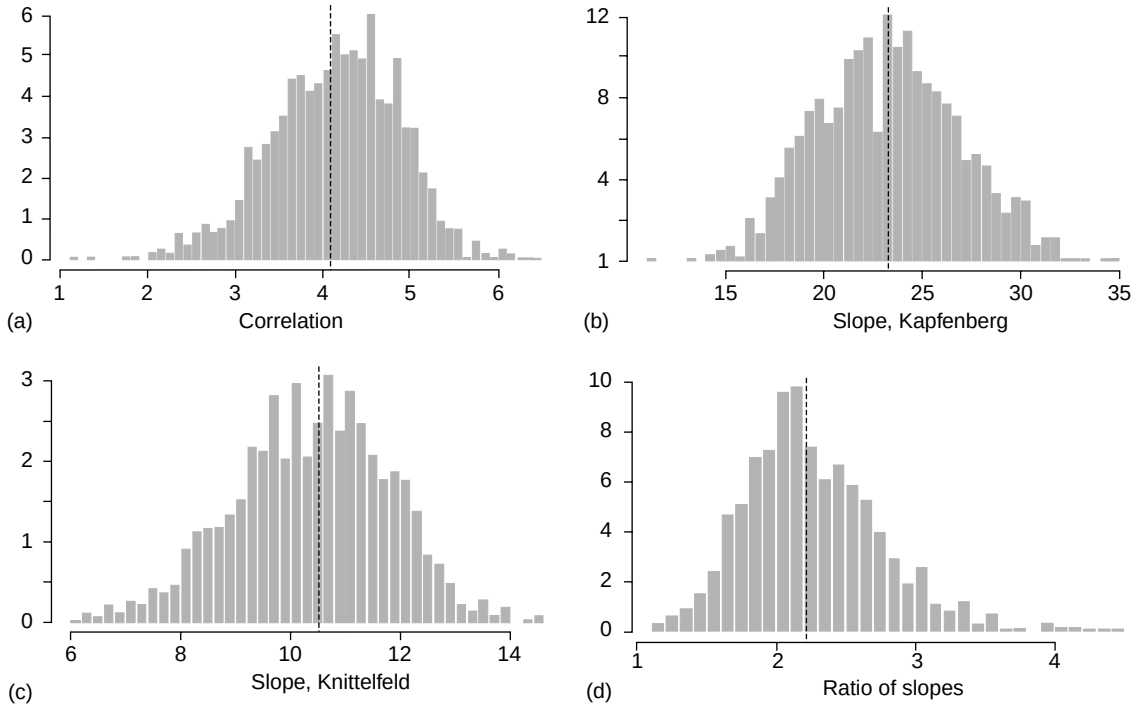

Figure 1: Bootstrap distributions of a) correlation between blood creatinine and magnesium concentrations, b) slope of the linear regression of magnesium on blood creatinine concentrations for 38 children from a relatively contaminated area (Kapfenberg), c) slope for 52 children in a less contaminated area (Knittelfeld), d) ratio of the two slopes. The observed value is marked by the dotted vertical line in all four panels.

References

[1] Archer, G. and Giovannoni, J.-M. 1998, Statistical analysis with bootstrap diagnostics of atmospheric pollutants predicted in the APSIS experiment. Water, Air, and Soil Pollution 106, 43-81.

[2] Bailer, A. J. and Oris, J. T. 1994, Assessing toxicity of pollutants in aquatic systems. pp 25-40 in Lange, N., et al., eds. *Case Studies in Biometry*, Wiley, New York.

[3] Chernick, M. R. 1999, *Bootstrap Methods, A Practitioner's Guide*. Wiley, New York.

[4] Cooley, R. L. 1997, Confidence intervals for ground-water models using linearization, likelihood, and bootstrap methods. Ground Water 35, 869-880.

[5] Davison, A. C. and Hinkley, D. V. 1997. *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge

[6] DiCiccio, T. and Efron, B. 1992, More accurate confidence intervals in exponential families, Biometrika 79, 231-245.

[7] DiCiccio, T. J. and Efron, B. 1996, Bootstrap confidence intervals (with discussion), Statistical Science 11, 189-228.

[8] DiCiccio, T. J. and Romano, J. P. 1988, A review of bootstrap confidence intervals (with discussion), Journal of the Royal Statistical Society, Series B, 50, 338-370, with correction 51, 470.

[9] Dixon, P. M. 2001. The bootstrap and the jackknife: describing the precision of ecological studies. pp 267-288 in Scheiner, S. and Gurevitch, J. (eds.) *Design and Analysis of Ecological Experiments*, 2nd ed. Oxford University Press, Oxford.

[10] Dudewicz, E. J. 1992, The generalized bootstrap. pp 31-37 in Jöckel, K. -H., Rothe, G. and Sendler, W., eds., *Bootstrapping and Related Techniques*. Springer-Verlag, Berlin.

[11] Efron, B. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia.

[12] Efron, B. and Tibshirani, R. 1986. Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. Statistical Science 1, 54-77.

[13] Efron, B. and Tibshirani, R. J. 1993, *An Introduction to the Bootstrap*, Chapman and Hall, New York.

[14] Fortin, V., Bernier, J. and Bobée, B. 1997, Simulation, Bayes, and bootstrap in statistical hydrology, Water Resources Research 33, 439-448.

[15] Hall, P. 1992, *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.

[16] Harvey, A. C. 1993, *Time Series Models, 2nd ed.* MIT Press, Cambridge MA.

[17] Künsch, H. R. 1989, The jackknife and the bootstrap for general stationary observations, Annals of Statistics 17, 1217-1241.

[18] LePage, R. and Billard, L. 1992. *Exploring the Limits of Bootstrap*. Wiley, New York.

[19] Manly, B. F. J. 1997, *Randomization, Bootstrap and Monte Carlo Methods in Biology, 2nd edition.* Chapman and Hall, London.

[20] Martin, M. A. 1990, On bootstrap iteration for coverage correction in confidence intervals. Journal of the American Statistical Association, 85, 1105-1118.

[21] McCulloch, C. E., Boudreau, M. D., and Via, S. 1996, Confidence regions for evolutionary trajectories. Biometrics 52, 184-192.

[22] Millard, S. P. and Deverel, S. J., 1988, Nonparametric statistical methods for comparing two sites based on data with multiple nondetect limits. Water Resources Research 24, 2087-2098.

[23] Newton, M. A. 1996. Bootstrapping phylogenies: large deviations and dispersion effects. Biometrika 83, 315-328.

[24] Pillar, V. D. 1999, The bootstrapped ordination re-examined. Journal of Vegetation Science 10, 895-902.

[25] Politis, D. N. and Romano, J. P., 1994, The stationary bootstrap. Journal of the American Statistical Association, 89, 1303-1313.

[26] SAS Institute, Inc. 1995. Jackboot macro documentation, SAS Institute, Cary NC.

[27] Smith, S. J. 1997, Bootstrap confidence limits for groundfish trawl survey estimates of mean abundance. Canadian Journal of Fisheries and Aquatic Sciences, 54, 616-630.

[28] Shao, J. and Tu, D. 1995, *The Jackknife and Bootstrap*, Springer, New York.

[29] Schmoyer, R. L., Beauchamp, J. J., Brandt, C. C., Hoffman, F. O., Jr. 1996, Difficulties with the lognormal model in mean estimation and testing. Environmental and Ecological Statistics, 3, 81-97

[30] Silverman, B. W. and Young, G. A. 1987, The bootstrap: to smooth or not to smooth?, Biometrika 74, 469-479.

[31] Simar, L, and Wilson, P. W. 1998, Sensitivity analysis of efficiency scores: how to bootstrap in nonparametric frontier models. Management Science 44, 49-61.

[32] Solow, A. R. 1985, Bootstrapping correlated data. Journal of the International Association of Mathematical Geology. 17, 769-775.

[33] Solow, A. R. 1989, Bootstrapping sparsely sampled spatial point patterns. Ecology 70, 379-382.

[34] Venables, W. N. and Ripley, B. D. 1994, *Modern Applied Statistics with S-Plus*. Springer-Verlag, New York.

[35] Wehrens, R. and Van der Linden, W. E. 1997, Bootstrapping principal component regression models. Journal of Chemometrics, 11, 157-171.

[36] Willemain, T. R. 1994, Bootstrapping on a shoestring: resampling using spreadsheets. American Statistician 48, 40-42.

[37] Young, G. A. 1994, Bootstrap: more than a stab in the dark? (with discussion). Statistical Science 9, 382-415.

[38] Yu, C. -C., Quinn, J. T., Dufournaud, C. M., Harrington, J. J., Rogers, P. P. and Lohani, B. N. 1998, Effective dimensionality of environmental indicators: a principal components analysis with bootstrap confidence intervals. Journal of Environmental Management 53, 101-119.