14

The Bootstrap and the Jackknife

Describing the Precision of Ecological Indices

PHILIP M. DIXON

14.1 Introduction

Quantitative ecology uses many indexes and coefficients, including diversity indices (Magurran 1988), similarity indices (Goodall 1978), competition coefficients (Schoener 1983), population growth rates (Lenski and Service 1982), and measures of size hierarchy (Weiner and Solbrig 1984). All of these indices are statistics, calculated from a sample of data from some population and used to make conclusions about the population (chapter 1). To help a reader interpret these conclusions, good statistical practice includes reporting a measure of uncertainty or precision along with a statistic. Although it is easy to calculate the values of many ecological statistics, it is often difficult to estimate their precision. This chapter discusses two techniques, the bootstrap and jackknife, that can be used to estimate the precision of many ecological indices.

14.1.1 Precision, Bias, and Confidence Intervals

To understand how the bootstrap and jackknife work, we must first review some concepts of statistical estimation. Consider how to estimate the mean reproductive output for a species in a defined area. The statistical population is the set of values (number of seeds per plant) for every plant in the area. The mean reproductive output is a parameter; it describes some interesting characteristic of the population. This parameter is known exactly if every plant is studied, but completely enumerating the population is usually impractical.

Instead, a random sample of plants are counted. The average reproductive output for the plants in the sample is an estimate of the parameter in which we are interested. How accurate is this estimate? Plants have different reproductive outputs; a different sample of plants will provide a different estimate of average reproductive output. Some estimates will be below the population mean, whereas other estimates will be above (figure 14.1). The set of average reproductive output from all possible samples of plants is the sampling distribution of the sample average; characteristics of this distribution describe the accuracy of a statistic.

The accuracy of a statistic has two components: bias and precision (Snedecor and Cochran 1989). Bias measures whether a statistic is consistently too low or too high. It is defined as the difference between the population value and the average of the sampling distribution. If the sampling distribution is centered around the population value, then the statistic is unbiased. Precision depends on the variability in the sampling distribution and is often measured by the variance or standard error. Bias and precision are separate components of accuracy. A precise statistic may be biased if its sampling distribution is concentrated around some value that is not the population value.

The distribution of estimates from all possible samples is a nice theoretical concept, but an experimenter has only one sample of data and one value of the estimate. If the method used to sample the population has well-defined characteristics (e.g.,

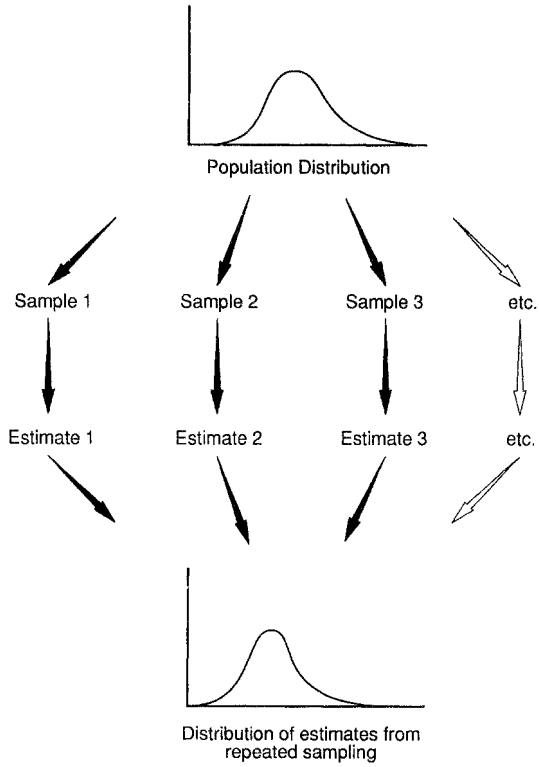

Figure 14.1 Relationship between population distribution and sampling distribution. If the values in the population were known, then the sampling distribution of the estimate could be obtained by repeatedly drawing samples from the population and calculating the statistic from each sample.

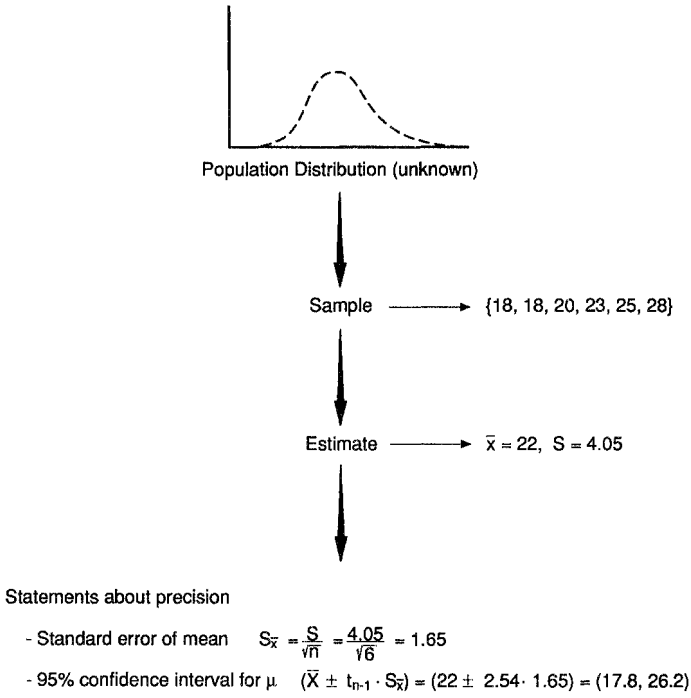

Figure 14.2 Use of data and statistical theory to infer a sampling distribution from one sample of data. The sampling distribution of certain statistics is known theoretically. Statements about the precision can be made from sample information (e.g., sample mean and standard deviation).

simple random sampling), the sampling distributions of sample averages and a few other statistics can be calculated from one sample of data because there are known mathematical relationships between the properties of the sample and the properties of the population (figure 14.2). For example, the sample average from a simple random sample is unbiased and its standard error, $s_{\bar{x}}$, can be estimated by $s_{\bar{x}} = s_x/(n)^{1/2}$, where s_x is the sample standard deviation and n is the sample size.

Confidence intervals can also be calculated if the sampling distribution is known. Under certain assumptions about the observations, the sample mean and sample variance are independent estimates with known distributions, so that a 95% confidence interval is given by $\bar{x} - t_{n-1}s_{\bar{x}}$, $\bar{x} + t_{n-1}s_{\bar{x}}$, where t_{n-1} is the critical value for a two-sided 95% confidence interval from a t-distribution with $n - 1$ degrees of freedom. Confidence intervals are commonly misunderstood. Remember, a confidence interval is a random interval with the property that it includes the population mean, which is fixed, with a given frequency. A 95% confidence interval of (1, 2) does not mean that 95% of possible sample values are between 1 or 2, and it does not mean that 95% of possible population means are between 1 and 2. Instead, it means that 95% of the time, the confidence interval will include the population mean.

14.1.2 Precision and Bias of Ecological Indexes

Many useful ecological indexes are more complicated than a sample mean, and a sampling distribution cannot be calculated mathematically. However, it is more important to choose an ecologically useful coefficient rather than a statistically tractable one. The sampling distributions of many ecologically useful coefficients can be estimated using the bootstrap or the jackknife. The jackknife estimates the bias and variance of a statistic. In addition to estimating the bias and variance, the bootstrap also determines a confidence interval.

The jackknife and bootstrap have been used in many ecological applications, such as the following: population growth rates (Meyer et al. 1986; Juliano 1998), population sizes (Buckland and Garthwaite 1991), toxicity assessment (Bailer and Oris 1994), ratios of variables (Buonaccorsi and Liebhold 1988), genetic distances (Mueller 1979), selection gradients (Bennington and McGraw 1995), diversity indexes (Heltshe and Forrester 1985; Heltshe 1988), species richness (Smith and van Belle 1984; Palmer 1991), diet similarity (Smith 1985), home ranges (Rempel et al. 1995), and niche overlap indexes (Mueller and Altenberg 1985; Manly 1990). General introductions to the bootstrap and jackknife can be found in Manly (1997), Crowley (1992), and Stine (1989). More of the statistical theory can be found in Hall (1992), Shao and Tu (1995), and Davison and Hinkley (1997). Useful surveys of applications and problems include Léger et al. (1992), Young (1994), and Chernick (1999).

In this chapter, I will describe the jackknife and bootstrap and illustrate their use with two indices: the Gini coefficient of size hierarchy and the Jaccard index of community similarity [see http://www.oup-usa.org/sc/0195131878/ for the computer code]. I also describe some of the practical issues applying the techniques to answer ecological questions. The focus is on nonparametric methods for independent observations, but some approaches for more complex data are discussed at the end.

14.1.3 Gini Coefficients and Similarity Indices

The Gini coefficient, G, is a measure of inequality in plant size (Weiner and Solbrig 1984). It ranges from 0, when all plants have the same size, to a theoretical limit of 1, when one plant is extremely large and all other plants are extremely small. The coefficient G can be calculated from a set of data:

$$G = \frac{\sum_{i=1}^{n} (2i - n - 1)X_i}{(n-1)\sum_{i=1}^{n} X_i} \tag{14.1}$$

where n is the number of individual plants and X_i is the size of the ith plant, when plants are sorted from smallest to largest, $X_1 \leq X_2 \leq \ldots \leq X_n$. The bootstrap can be used to estimate the precision of G (Dixon et al. 1987).

A similarity index describes the similarity in species composition, which can be estimated by sampling the species in each community and computing the simi-

larity between a pair of samples. Many similarity indexes have been proposed (see Goodall 1978 for a review); the Jaccard (1901) index, which depends on the similarity in species presence, is computed as

$$J = \frac{a}{a + b + c} \tag{14.2}$$

where a is the number of species found in both communities, b is the number of species found only in the first, and c is the number of species found only in the second. The coefficient is ecologically useful, but its precision is difficult to calculate analytically. The bootstrap and jackknife are methods to describe the accuracy of similarity indexes (Smith et al. 1986).

14.2 The Jackknife

The jackknife procedure (Miller 1974) is a general technique to answer the question, How precise is my estimate? It can estimate bias or standard error for a statistic, but not for a confidence interval (Efron 1982). The basic idea is that the bias and standard error can be estimated by recalculating the statistic on subsets of the data. Although the bootstrap has theoretical advantages, the jackknife requires considerably less computation and includes no random component. The jackknife can also be used in conjunction with the bootstrap (see sections 14.4.3 and 14.5.6).

14.2.1 Ecological Example: Gini Coefficients to Measure Size Hierarchy in *Ailanthus* Seedlings

The tree-of-heaven, *Ailanthus altissima*, is an aggressive introduced tree. Evans (1983) studied whether seedlings grown in a competitive environment have a stronger size hierarchy (higher size inequality) than individually grown plants. Six seeds were randomly chosen from a large collection of seeds and planted in individual pots. Another 100 seeds were planted in a common flat so that they could compete with each other. After 5 months, each survivor was measured (table 14.1). The Gini coefficient for the individually grown plants ($G = 0.112$) was smaller than that for the competitively grown plants ($G = 0.155$), consistent with the hypothesis that competition increases the inequality in the size distribution (Weiner and Solbrig 1984). The sample sizes are small, especially for the individually grown plants, so it is important to estimate the precision of each estimate and calculate a confidence interval for the difference.

14.2.2 Jackknifing the Gini Coefficient

The jackknife estimates of bias and standard error (Miller 1974) are calculated by removing one point at a time from the data set. Consider the Gini coefficient for individually grown plants (table 14.1). The observed value, based on all six plants is $G = 0.112$. If the first point is removed, G_{-1} calculated from the remain-

Table 14.1 Number of leaf nodes for 5-month-old *Ailanthus altissima* grown under two conditions: in individual pots and in a common flat

6 plants grown individually:							
	18	18	20	23	25	28	
75 surviving plants grown together in a common flat:							
8							
10							
11	11	11					
12	12	12					
13	13	13	13	13	13	13	
14	14	14	14	14			
15	15	15	15	15	15	15	15
16	16						
17	17	17					
18	18	18	18	18			
19	19	19	19				
20	20	20	20	20	20		
21	21	21					
22	22	22	22	22			
23	23	23	23	23	23	23	
24	24	24					
25	25	25	25				
26							
27	27	27					
30							

ing five data points is 0.110. If the fourth point is removed, G_{-4} calculated from the remaining five data points is 0.124. Each perturbed value is combined with the original statistic to compute a pseudovalue, p_i, for each of the n data points:

$$p_i = G + (n-1)(G - G_{-i}). \tag{14.3}$$

The six jackknife samples for individually grown *Ailanthus*, their G_{-i} values, and their pseudovalues are given in table 14.2.

Table 14.2 Jackknife samples with Gini coefficients for individually grown *Ailanthus*

Jackknife sample					Gini coefficient	Pseudovalue
18	20	23	25	28	0.110	0.124
18	20	23	25	28	0.110	0.124
18	18	23	25	28	0.120	0.070
18	18	20	25	28	0.124	0.053
18	18	20	23	28	0.117	0.089
18	18	20	23	25	0.091	0.216
					mean $\bar{p}$:	0.1128

The jackknife estimates of bias (Efron 1982) are

$$\hat{\text{bias}} = G - \bar{p} \tag{14.4}$$

where G is the sample Gini coefficient and $\bar{p}$ is the mean of the jackknife pseudovalues. For the individually grown *Ailanthus*, the jackknife estimate of bias is $0.1121 - 0.1128 = -0.0007$. This bias is very small, but if it were larger, it could be subtracted from the observed value to produce a less biased estimate. The jackknife is especially effective at correcting a statistic for first-order bias, a bias that is linearly proportional to the sample size (Miller 1974).

The jackknife estimate of the standard error is just the standard error of the pseudovalues,

$$s_G = \sqrt{\frac{\Sigma(p_i - \bar{p})^2}{n(n-1)}} \tag{14.5}$$

where n is the sample size and $\bar{p}$ is the mean of the pseudovalues. From the data in table 14.2, the estimated standard error (s.e.) is 0.024. For the 75 plants grown in competition, the s.e. is estimated to be 0.010. The two samples are independent, so the s.e. of the difference is $(s_1^2 + s_2^2)^{1/2}$. This is estimated to be 0.026, which is about half of the observed difference (0.043). These calculations have been demonstrated using a small sample of data, but estimates of bias and standard error from small samples must be treated with caution. As with most statistical techniques, estimates from larger samples are more precise.

In general, the jackknife cannot be extended to calculate confidence intervals or test hypotheses (Efron 1982). Some attempts have been made to construct confidence intervals by assuming a normal distribution (Meyer et al. 1986; Smith et al. 1986). Such confidence intervals would have the form $(G \pm t_k s_G$, where t_k is a critical value from a t-distribution with k degrees of freedom. The problem with this approach is that the appropriate number of degrees of freedom is unknown, in spite of a lot of theoretical research (Efron and LePage 1992). However, using $n - 1$ degrees of freedom, where n is the original sample size, has worked well in some cases (Meyer et al. 1986).

The jackknife has also been used to estimate the precision of various similarity measures, including measures of diet similarity (Smith 1985), community similarity (Smith et al. 1979; Smith et al. 1986; Heltshe 1988), and niche overlap (Mueller and Altenberg 1985). Because these methods require comparing two-samples, the jackknife procedure is slightly different. Details of the two-sample jackknife can be found in Dixon (1993) or in the original articles.

14.3 The Bootstrap Method

The bootstrap has become a popular method for estimating confidence intervals and testing hypotheses about many ecological quantities. Although the bootstrap is applicable to many ecologically problems, it is not appropriate for everything. I will

describe the principles of bootstrapping, compare some of the many varieties of the bootstrap procedure, and discuss when the bootstrap is not appropriate.

The bootstrap is a two-step procedure to approximate the unknown sampling distribution. First, the unknown distribution of values in the population is approximated using information from the observed sample (figure 14.3). Then, many bootstrap samples are drawn from this distribution. The unknown sampling distribution is approximated by the distribution of estimates from many bootstrap samples (figure 14.3). The bootstrap distribution is used to calculate a confidence interval, test a hypothesis, and estimate the standard error and bias for a statistic.

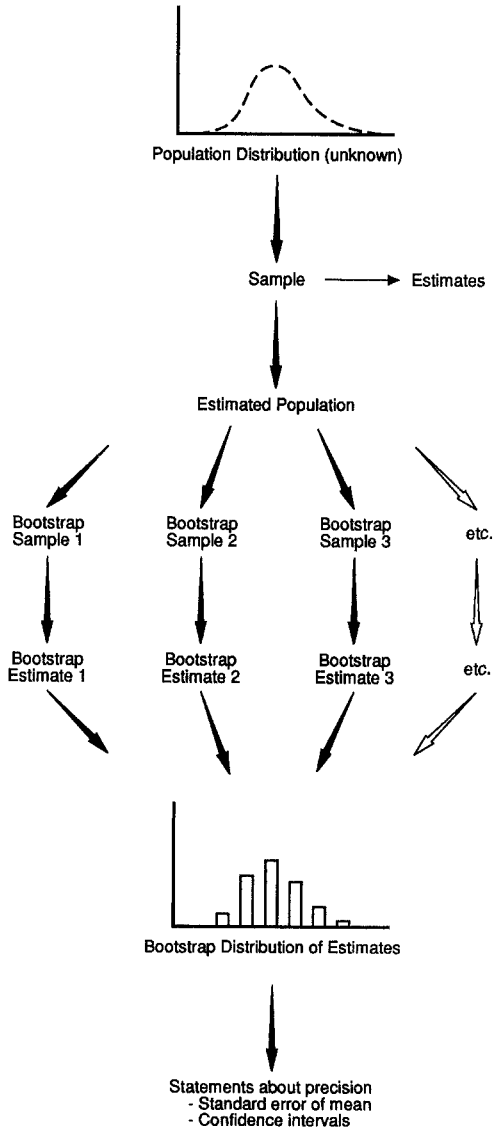

Figure 14.3 Use of data and the bootstrap distribution to infer a sampling distribution. The bootstrap procedure estimates the sampling distribution of a statistic in two steps. The unknown distribution of population values is estimated from the sample data, then the estimated population is repeatedly sampled, as in figure 14.1, to estimate the sampling distribution of the statistic.

Although the concepts are very simple, there are many possible varieties of bootstraps. These differ in three characteristics: (1) how the population is approximated, (2) how bootstrap samples are taken from the population, and (3) how the endpoints of confidence intervals are calculated.

14.3.1 Approximating the Population: Parametric and Nonparametric Bootstraps

The first step in the bootstrap is using information in the observed sample to approximate the unknown distribution of values in the population. A parametric bootstrap approximates the population by a specific distribution (e.g., a lognormal or Poisson distribution) with parameters that are estimated from the sample. For example, the number of leaves on the competitively grown plants fits a Poisson distribution with a mean of 18.4 reasonably well. There are only six individually grown plants, so almost any discrete distribution can fit those data. A Poisson distribution with a mean of 22 is not an unreasonable choice. These parametric distributions are used to describe the unknown populations. Each parametric bootstrap sample of competitively grown plants is a random sample of 75 observations from a Poisson distribution with a mean of 18.4; each parametric bootstrap sample of individually grown plants is a random sample of six observations from a Poisson distribution with a mean of 22.

In a nonparametric bootstrap, the population is approximated by the discrete distribution of observed values (figure 14.3). The estimated population of individual *Ailanthus* sizes is a discrete distribution in which the value 18 occurs with probability 2/6 (because 2 plants had 18 leaves) and the values 20, 23, 25, and 28 each have probability 1/6. Most ecological applications use the nonparametric bootstrap because it requires fewer assumptions about the population. However, the nonparametric bootstrap assumes that the observed sample is representative of the population. With very small samples, the parametric bootstrap is often better, as long as the assumed distribution is not wildly incorrect (Davison and Hinkley 1997).

14.3.2 Drawing a Bootstrap Sample From the Population: Ordinary, Balanced, and Moving Block Bootstraps

Once the population is approximated, samples must be drawn from this population. The simplest way is to draw a simple random sample, with replacement, from the values in the population. If a nonparametric bootstrap is used, the bootstrap sample is a simple random sample of the observed values. The statistic, (e.g., the Gini coefficient) is then calculated from the values in the bootstrap sample. Five bootstrap samples and Gini coefficients are shown in table 14.3, which illustrates an important characteristic of bootstrap samples. Each bootstrap sample omits some observed values and repeats others because the observed data are sampled with replacement. Some of the bootstrap samples have Gini coefficients larger than the observed value, $G = 0.112$, whereas other samples have smaller coefficients.

Table 14.3 Five bootstrap samples for the data from individually grown plants

Bootstrap sample						Gini coefficient
18	18	23	25	28	28	0.117
18	18	18	25	25	25	0.098
18	18	23	23	28	28	0.116
18	18	18	20	23	28	0.107
18	18	20	23	25	28	0.112

This process is repeated for many bootstrap samples. Typically, 50 to 100 bootstrap samples are used to estimate a standard error; 1000 or more bootstrap samples are recommended to calculate a confidence interval (Efron and Tibshirani 1993). The number of bootstrap samples will be discussed further in section 14.5.5.

One potential concern with simple random sampling is that each observation may not occur equally often in the bootstrap samples. In the five samples illustrated in table 14.3, the value 20 occurs a total of two times, but the value 28 occurs a total of six times. Hence, the aggregated bootstrap samples do not represent the population, in which the values 20 and 28 are equally frequent. The balanced bootstrap forces each value to occur equally frequently. One algorithm to draw 100 balanced bootstrap samples is to write down 100 copies of the observed data. For the six individually grown plants, this population has 600 values, 200 of which are 18, 100 are 20, 100 are 23, 100 are 25, and 100 are 28. Randomly permute the 600 values, then take the first 6 as the first bootstrap sample, the next 6 as the second, and so on. The balanced bootstrap can markedly increase the precision of bias calculations, but it is less useful for confidence interval calculations (Davison and Hinkley 1997).

Both the simple random sample and balanced bootstraps assume that there is no temporal or spatial correlation among the values. If there is any correlation, it is eliminated by the randomization used in the ordinary and balanced bootstraps. The moving block bootstrap generates bootstrap samples that retain some of the correlation (Davison and Hinkley 1997). It, and other methods for bootstrapping correlated data, are discussed in section 14.5.4.

14.3.3 Estimating Bias and Standard Error From the Bootstrap Distribution

The bootstrap distribution provides the information necessary to estimate the bias and standard error of a statistic like the Gini coefficient. The bootstrap estimate of bias is simply the difference between the average of the bootstrap distribution and the value from the original sample. For the individually grown *Ailanthus* plants, the sample Gini coefficient is 0.1121 and the average Gini value in the five bootstrap samples of table 14.2 is 0.1099. Hence, the bootstrap estimate of

bias is 0.1099 – 0.1121 = –0.0021. In practice, you should use 50 to 100 bootstrap samples to estimate the bias (Efron 1987). Using 1000 bootstrap samples (data not shown), the biases for the individually and competitively grown *Ailanthus* are estimated to be –0.018 and –0.0020, respectively. In both cases, the bias is small. If the bias were larger, the estimate of bias could be subtracted from the observed value to produce a less biased estimate.

The standard error of the sample Gini coefficient is estimated by the standard deviation of the bootstrap distribution. For the five bootstrap samples shown in table 14.2, the estimated standard error is 0.0079. Again, I use only five bootstrap samples to demonstrate the calculations. Using 1000 bootstrap samples (data not shown), the estimated standard errors for the Gini coefficient of individually and competitively grown plants are 0.022 and 0.0097, respectively. These numbers are not too different from those estimated by the jackknife. This is often the case; theoretical calculations show that the jackknife is a linear approximation to the bootstrap (Efron 1982).

A standard error is a single measure of precision. It can be turned into a confidence interval if we assume a particular distribution, such as the normal. However, confidence intervals can be estimated from the bootstrap distribution without assuming normality.

14.4 Calculating Confidence Intervals From the Bootstrap Distribution

Confidence intervals can be calculated from the bootstrap distribution in at least five different ways: the percentile bootstrap, the basic bootstrap, the studentized bootstrap, the bias-corrected bootstrap, and the accelerated bootstrap (Davison and Hinkley 1997). Different authors have used other names for some of these methods, which increases the confusion. Some of the synonyms are given in the documentation to the SAS JACKBOOT macro (SAS Institute 1995). These different methods, which are discussed subsequently, represent a trade-off between simplicity and generality. No method is best for all problems.

14.4.1 Percentile Bootstrap

The percentile bootstrap is the simplest, and most commonly used, method to construct bootstrap confidence intervals. In this method, the 2.5 and 97.5 percentiles of the bootstrap distribution are used as the limits of a 95% confidence interval. To calculate the 2.5 percentile from N bootstrap replicates, sort the estimates from the bootstrap samples in order from smallest to largest. The pth percentile is the $(N+1)p/100$th largest value. A histogram of 999 bootstrap Gini coefficients for competitively grown *Ailanthus* is shown in figure 14.4. The 2.5 and 97.5 percentiles of $N = 999$ observations are given by the 25th and 975th largest values. They are 0.133 and 0.171, respectively, so (0.133, 0.171) is the 95% confidence interval using the percentile method.

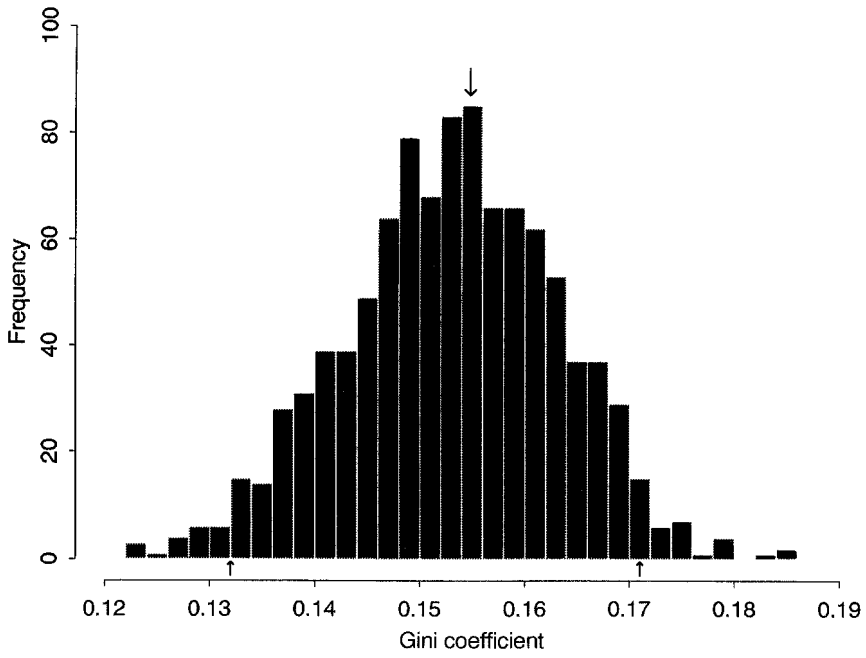

Figure 14.4 Histogram of 999 bootstrap values for competitively grown *Ailanthus*. Arrows below the histogram mark the 25th and 975th largest values. Arrow above the histogram marks the Gini coefficient for the observed data ($G = 0.155$).

Percentile bootstrap confidence intervals can be calculated for any statistic, but they do not always work very well (Schenker 1985). A confidence interval can be evaluated by simulating many data sets from a population with some known parameter, calculating a confidence interval from each data set, and counting how many confidence intervals bracket the true parameter. A 95% confidence interval should include the true parameter in 95% of the simulated data sets. When sample sizes are small (i.e., less than 50), the percentile confidence intervals for the variance (Schenker 1985) and Gini coefficient (Dixon et al. 1987) are too narrow. For example, with samples of 20 points, 90% confidence intervals for the variance include the true variance only 78% of the time, and 95% confidence intervals for a Gini coefficient of a lognormal population include the true value of 0.30 only 85% of the time (Schenker 1985; Dixon et al. 1987).

The percentile bootstrap produces correct confidence intervals when the bootstrap distribution is symmetrical and centered on the observed value (Efron 1982). This is not the case for the Gini coefficient in this example, where 56% of the bootstrap values are smaller than the observed value (figure 14.4). For individually grown *Ailanthus*, 74% of the bootstrap values are smaller than the observed value. In both cases, the observed values are larger than the median of the bootstrap distribution, so the upper and lower confidence bounds are too low.

The plethora of bootstrap confidence intervals comes from different ways of adjusting the percentile bootstrap confidence intervals to improve their coverage. The bias-corrected bootstrap adjusts for bias in the bootstrap distribution and the accelerated bootstrap adjusts for both bias and skewness. The studentized bootstrap is based on the bootstrap distribution of the statistic adjusted by its mean and standard error. The ratio $(G_i - G)/s_i$ is like a Student's t-statistic, hence the name of the method. This approach is especially useful when the variability in the data depends on the mean, as is common in ecological data. The basic bootstrap, a "turned around" version of the percentile bootstrap, is based on the fundamental relationship between hypothesis tests and confidence intervals.

14.4.2 Bias-corrected and Accelerated Bootstraps

The bias-corrected percentile bootstrap adjusts for a bootstrap distribution that is not centered on the observed statistic. The bounds of the bias-corrected intervals are found by determining F, the fraction of bootstrap replicates that are smaller than the observed value and the value, z_0, the probit transform of F. The appropriate percentiles for a 95% confidence interval are calculated as follows:

$$P_l = \phi(2z_0 - 1.96) \tag{14.6}$$

$$P_u = \phi(2z_0 + 1.96) \tag{14.7}$$

where ϕ is the normal cumulative distribution function (c.d.f.). The normal c.d.f. and probit transformation are available in many statistical packages, or they can be computed from tables of areas of the normal curve (e.g., Rohlf and Sokal 1995, table A). The values ± 1.96 are the critical values for a 95% confidence interval of a statistic with a normal distribution. They can be changed to generate other confidence intervals. The upper and lower bounds of the bias-corrected confidence interval are given by the values in the bootstrap distribution that match the calculated percentiles, P_l and P_u. When the observed value is the median of the bootstrap distribution, there is no difference between the bias-corrected and percentile confidence intervals.

For the bootstrap distribution shown in figure 14.4, 56.4% of the values are smaller than the observed Gini coefficient of 0.1548, so $F = 0.564$ and $z_0 = 0.166$. The desired percentiles of the bootstrap distribution are $P_l = \phi(-1.628) = 5.181\%$, and $P_u = \phi(2.29) = 98.91\%$. The number of bootstraps (1000) times the percentiles (5.181% or 98.91%) are not exact integers, so the observations corresponding to the next most extreme integer are used. The bounds of the bias-corrected 95% confidence interval are the 51st and 990th largest values in the bootstrap distribution, which are (0.136, 0.176). Note that the values of the upper and lower bounds are both larger than those for the percentile bootstrap, which were (0.133, 0.171). The bias-correction procedure has shifted the confidence interval upward to adjust for bias.

The accelerated bootstrap makes a second correction that is helpful when bootstrapping statistics like the variance or the Gini coefficient, where a few extreme

points will have a large influence on the observed value. This technique is so named because it "accelerates" the bias-correction (Efron 1987). These confidence intervals depend on two constants, z_0 used to correct for bias and the acceleration factor, a, which corrects for skewness. The a coefficient is nonzero when a few points have a large influence on the observed estimate. It can be estimated nonparametrically using the jackknife (see Dixon 1993, p. 302, or Efron and Tibshirani 1993, p. 186, for the details). The percentiles for the upper and lower 95% confidence bounds are calculated as

$$P_l = \Phi\left(z_0 + \frac{z_0 - 1.96}{1 - a(z_0 - 1.96)}\right) \tag{14.8}$$

$$P_u = \Phi\left(z_0 + \frac{z_0 + 1.96}{1 - a(z_0 + 1.96)}\right) \tag{14.9}$$

where ϕ is the normal cumulative distribution function, as with the bias-corrected confidence intervals, and the value 1.96 is the normal critical value for a 95% confidence interval.

Using the number of leaf nodes of competitively grown *Ailanthus* (table 14.1), the constant a was estimated to be 0.0193, using the method described in Dixon (1993). For the bootstrap distribution in figure 14.4, $z_0 = 0.166$, so $P_l = \phi(-1.568) = 5.85\%$, and $P_u = \phi(2.383) = 99.14\%$. Hence, the bounds of the accelerated confidence interval were the 59th and 992nd largest values in the set of 1000 bootstrap replicates: (0.137, 0.176). For these data, both the bias-corrected and accelerated bootstrap procedures adjust the confidence interval upward to account for the skewed and biased sampling distribution.

14.4.3 Studentized Bootstrap

The percentile and bias-corrected bootstraps assume that the sampling distribution has constant variance. In many practical cases, the variance of the sampling distribution depends on the mean. Some ecological examples include average plant sizes, average survival time under exposure to some toxicant, and estimated proportions. The variance of the observations is related to the mean, so the variance of the sampling distribution is related to the mean. The acceleration constant, a, used in the accelerated bootstrap confidence interval is one way to adjust for this unequal variance. A second way is the studentized bootstrap.

When a statistic, g, has a normal sampling distribution, the bounds of a confidence interval are given by

$$(g - ts_g,\ g + ts_g) \tag{14.10}$$

where s_g is the standard error of g, and $\pm t$ are quantiles of a t-distribution with the appropriate degrees of freedom. The idea behind the studentized bootstrap is to compute a confidence interval with the same form as equation 14.10, but use different critical values and relax the assumption of normality. The studentized bootstrap confidence intervals are given by

$$(g + b_l s_g,\ g + b_h s_g) \tag{14.11}$$

where b_l and b_h are percentiles from the bootstrap distribution of a studentized version of the statistic.

This bootstrap distribution is computed by drawing a bootstrap sample of observations and calculating the statistic, G_i, and its standard error, s_i. The studentized statistic is then

$$b_i = \frac{G_i - G}{s_i}$$

where G is the value observed in the original sample. Note that a standard error is calculated for each bootstrap sample. A large number (e.g., 999) of values of b_i are calculated from a large number of bootstrap samples. For a 95% confidence interval, b_l and b_h are the 2.5 and 97.5 percentiles. If $N = 999$, estimates of b_i are sorted from smallest to largest; these percentiles are the 25th and 975th observations. These percentiles are combined with the observed statistic, g, and observed standard error, se_g, to compute the studentized confidence interval (equation 14.11).

Unlike the other bootstrap methods, the studentized bootstrap requires the standard error of the statistic. For some statistics, standard errors can be computed using appropriate formulas. The Gini coefficient is one of many statistics for which there is no formula for the standard error. A second bootstrap could be used to estimate the standard error (see section 14.3.3), but this requires considerable computation to estimate a standard error for each bootstrap sample. A more practical approach is to use the jackknife (section 14.2.2) to estimate the standard error of the observed statistic and each bootstrap statistic.

The bootstrap distribution of studentized Gini statistics for the 75 competitively grown individuals has a median of 0.26 and is slightly skewed (figure 14.5). The observed Gini coefficient is 0.155 with a jackknife standard error of 0.0102. The 2.5 and 97.5 percentiles are −1.80 and 2.48, respectively, so the 95% studentized bootstrap confidence interval is $(0.155 - 1.80 \times 0.0102,\ 0.155 + 2.48 \times 0.0102) = (0.137,\ 0.180)$.

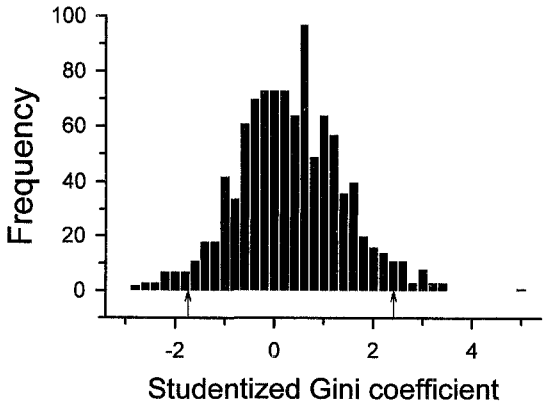

Figure 14.5 Histogram of 999 studentized bootstrap values for competitively grown *Ailanthus*. Arrows below the histogram mark the 25th and 975th largest values.

14.4.4 Basic Bootstrap

The percentile bootstrap is quite easy to interpret but somewhat controversial. The bootstrap distribution approximates the sampling distribution when the unknown parameter, θ, is set equal to the observed value, G. The controversial part is how to calculate an appropriate confidence interval from the bootstrap distribution (Davison and Hinkley 1997). The basic bootstrap determines confidence interval endpoints by exploiting the fundamental connection between hypothesis tests and confidence intervals. One way to find the endpoints of a 95% confidence interval is to use a test of the hypothesis $\theta = x$ to find the values $x = \mathrm{l}$ and $x = \mathrm{u}$ for which the hypothesis is rejected at exactly $P = 5\%$. Using the bootstrap to test these hypotheses requires estimating the sampling distribution when $\theta = \mathrm{l}$ and the sampling distribution when $\theta = \mathrm{u}$. Under the assumption that the s.e. of the sampling distribution is not related to the mean, the sampling distribution when $\theta = \mathrm{u}$ is estimated by shifting the observed sampling distribution by $\mathrm{u} - G$. The hypothesis that $\theta = \mathrm{u}$ is rejected at $P = 5\%$ when the observed value, G, is exactly the 2.5th percentile of the shifted sampling distribution. Similarly, the hypothesis that $\theta = \mathrm{l}$ is rejected at $p = 5\%$ when the observed value, G, is exactly the 97.5th percentile. Hence, the basic confidence interval bounds are given by $(2G - G_{0.975}, 2G - G_{0.025})$, where G is the observed statistic and $G_{0.025}$ and $G_{0.975}$ are the 2.5th and 97.5th percentiles of the bootstrap distribution.

For the competitively grown individuals, the 2.5th and 97.5th percentiles of the bootstrap distribution are 0.133 and 0.171, respectively, so the basic 95% confidence interval is $(2 \times 0.155 - 0.171,\ 2 \times 0.155 - 0.133) = (0.139,\ 0.177)$. When the sampling distribution is symmetrical about the observed value, G, the basic intervals and percentile intervals are the same. Because the sampling distribution of the Gini coefficient is slightly asymmetric, the endpoints of the basic interval are adjusted slightly upward from those for the percentile interval.

14.5 Practical Details

14.5.1 Choice of Method

The bootstrap and the jackknife are two techniques that answer the same question, How precise is a particular statistic? The bootstrap estimates either a standard error or a confidence interval, but the jackknife is most appropriately used to estimate a standard error. Confidence intervals based on the jackknife standard error assume that the statistic has a normal distribution. If the distribution is skewed or heavy tailed, then jackknife confidence intervals are likely to be inaccurate.

In extremely large samples, theoretical results show that both techniques give the same answers (see Efron 1982 and Efron and Tibshirani 1986 for details). For realistic data, the two techniques usually give slightly different answers. Which is better? The jackknife is often simpler to compute, but that is less important with modern computers. Often, both techniques give similar answers, but if the differ-

ences are important, the performance of the two techniques should be evaluated on a case-by-case basis.

14.5.2 Choice of Bootstrap Method

As stated in section 14.3, the choice of bootstrap specifies three characteristics: how to approximate the population, how to draw a bootstrap sample, and how to construct a confidence interval. Some decisions are more important than others. Most ecological applications use the nonparametric percentile bootstrap. It is certainly the simplest method to implement or describe. The choice of parametric or nonparametric bootstrap is not very important when the assumed parametric distribution is reasonable and the sample size is sufficiently large. If the sample size is small, the parametric bootstrap may provide more reliable answers. The choice of ordinary or balanced bootstrap rarely affects confidence intervals (Davison and Hinkley 1997).

The appropriate choice of which method to use to calculate the bootstrap confidence interval can be more difficult, but this also may not make much difference. For the competitively grown plants, all five methods for calculating endpoints give very similar confidence intervals (table 14.4). This is not always the case, especially if the sample size is small (e.g., Efron and Tibshirani 1993, p. 183; Davison and Hinkley 1997, p. 231; Manly 1997, p. 55).

When the differences between intervals are large enough to matter, it may help to consider the characteristics of the problem and each method. The percentile, bias-corrected, and accelerated confidence intervals are restricted to valid parameter values. For example, the Gini coefficient must be between 0 and 1. The endpoints of confidence intervals from these three methods will always fall between 0 and 1. The basic and studentized methods can lead to confidence interval endpoints that are negative or are larger than 1.

Theoretical arguments suggest that the accelerated bootstrap is generally better than the bias-corrected or percentile methods (Efron 1987). However, simulation studies suggest that the practical differences among bootstrap techniques are often small. If there is a difference, the studentized bootstrap has the best coverage (Davison and Hinkley 1997, p. 231 for ratios and P. M. Dixon, unpubl. data, 1999, for Gini coefficients and log-normal means). The adequacy of a method

Table 14.4 Lower and upper endpoints of 95% confidence intervals for the Gini coefficient in competitively grown *Ailanthus*

Bootstrap method	95% confidence interval
Percentile	(0.133, 0.171)
Bias-corrected	(0.136, 0.176)
Accelerated	(0.137, 0.176)
Studentized	(0.137, 0.180)
Basic	(0.139, 0.177)

can be evaluated by repeatedly simulating samples from some population, calculating a confidence interval for each sample, and counting the number of confidence intervals that included the true population parameter. Ideally, a 95% confidence interval includes the true parameter in 95% of the samples. However, studentized bootstrap confidence intervals are usually longer than the other confidence intervals. In some cases, the studentized intervals are ridiculously long, so some people prefer the percentile or accelerated methods (e.g., Efron and Tibshirani 1993).

14.5.3 What Should I Bootstrap?

The key assumption behind the bootstrap is independence. The observations are assumed to be independent samples from some population. The choice of what to bootstrap becomes very important in more complicated problems. For example, consider bootstrapping a regression or correlation coefficient. Bootstrapping a correlation coefficient between two traits on individuals is relatively easy (Efron 1982). If individuals were randomly sampled from some population, then the bootstrap replicates are formed by randomly choosing individuals with replacement from the sample, just as was done for Gini coefficients in my example. Bootstrapping a regression is more complicated because it can be done in two different ways (Efron and Tibshirani 1986, section 5). We can fit a regression, estimate the residuals and predicted values, and generate bootstrap replicates by adding randomly sampled residuals to the predicted values. This procedure estimates the precision of the regression coefficients, assuming that the regression model is correct (Efron and Tibshirani 1986). The other procedure is to randomly select individuals, just like bootstrapping a correlation coefficient. This procedure estimates the precision of the regression coefficients, even if the regression model is not correct (Efron and Tibshirani 1986). This bootstrap can also be used to estimate the precision of model selection procedures like stepwise selection and nonparametric regressions (Efron and Tibshirani 1991). It is also more appropriate when the independent variables are random, not fixed in advance by the investigator.

Bootstrapping is also more complicated when the data include multiple sources of variation (Davison and Hinkley 1997). Although bootstrapping complex survey data has received some attention (Sitter 1992), methods for bootstrapping complex experimental data are not well developed. For example, the zooplankton example had two sources of variation: between samples and between individuals in a sample. Samples could be bootstrapped, or individuals could be bootstrapped. This decision can be made based on what constitutes independent observations. If there is any patchiness in the zooplankton, individuals are unlikely to be independent, so bootstrapping individuals is not appropriate. Bootstrapping samples is more appropriate if the observed samples can be assumed to be random samples from locations in the reservoir. This approach ignores the variability between individuals in a sample.

Consider bootstrapping data with two sources of variation: samples and individuals. One obvious approach is to bootstrap both samples and individuals. First, generate a bootstrap sample by randomly selecting samples with replacement from the collection of samples, then bootstrap individuals (again with replacement) in each one of the selected samples. Theoretical calculations for simple problems (e.g., estimating the mean of all observations) show that this approach tends to overestimate the variance (P. M. Dixon unpubl. data, 1999, using results from Davison and Hinkley 1997, p. 101). Another simple approach is to bootstrap only samples, the highest level in the hierarchy of sources of variation. Theoretical calculations show that this approach underestimates the variance by a consistent factor of $(p-1)/p$, where p is the number of samples. Bootstrap confidence intervals with the correct width can be calculated (Davison and Hinkley 1997, p. 102), but it is not clear how such adjustments would be made in more complex, more realistic problems. Some preliminary work suggests that a generally appropriate approach is to bootstrap at all levels of the hierarchy and use studentized bootstrap to calculate confidence intervals (P. M. Dixon, unpubl. data, 1997). Although the bootstrap estimates are more variable than they "should be," this is corrected by the studentized bootstrap.

14.5.4 Bootstrapping Correlated Data

Another practical issue is how to bootstrap data that are spatially or temporal correlated. The ordinary and balanced bootstraps ignore any correlation and treat all observations as independent. Two approaches can be used to generate bootstrap samples from correlated data. If the correlation structure can be modeled using a time series or spatial correlation model, a model-based bootstrap can generate bootstrap samples that maintain the correlation structure. The process is similar to bootstrapping regression residuals. Fit the time series or spatial model to the data to estimate the correlation parameters and a set of independent residuals, bootstrap the residuals, then use the model to generate a correlated sample (Stoffer and Wall 1991; Davison and Hinkley 1997). This approach assumes that the correct model is used for the correlation structure. If so, this approach works well.

The moving blocks bootstrap is a nonparametric approach to generating bootstrap samples that maintain some of the correlation in the original data. The idea is more clearly explained with time series data. The observed sample is divided into b blocks each with of l sequential observations. A moving blocks bootstrap sample is constructed by randomly sampling the b blocks, with replacement, and concatenating them to form a bootstrap sample of bl observations. The idea is that the correlation among observations is strongest within each block of l observations and that observations in different blocks are (or are almost) independent. The choice of b and l are crucial. If b is too small, there are very few possible patterns of observations in the bootstrap samples. If l is too small, observations in different blocks are no longer independent. These and other practical details

implementing the moving blocks bootstrap and moving tiles bootstrap for spatial data are discussed in Davison and Hinkley (1997).

14.5.5 How Many Bootstrap Replicates Should I Use?

Increasing the number of bootstrap replicates increases both the precision of the estimated standard error or confidence interval and the cost of computing it. Efron (1987, section 9) recommends using approximately 100 bootstrap replicates to estimate a variance or standard error and 1000 or more to estimate a confidence interval. Variances can be estimated more precisely because the variance is an average of squared deviations, whereas the endpoints of the confidence interval are individual data points. If the bias-corrected or accelerated techniques are used, more bootstrap replicates are necessary. For 95% confidence intervals, about 50% more (i.e., 1500) bootstrap replicates should be used (Efron 1987, section 9). The extra replicates are necessary because the estimation of the z_0 and a constants introduces extra variability into the endpoints of the confidence intervals.

Although more bootstrap replicates increase the precision of the estimated standard error or confidence interval, there is a limit. Bootstrap estimates have two sources of variability, one due to variability among the bootstrap replicates and one due to sampling variability in obtaining the original data. More bootstrap replicates can not substitute for more observations in the original data.

14.5.6 Diagnostics for Bootstraps

Bootstrap diagnostics are tools to assess whether the bootstrap confidence intervals reflect the entire sample or whether they are heavily dependent on one or a few observations (Efron 1992). One diagnostic method combines the jackknife and the bootstrap: remove points one at a time and calculate a bootstrap confidence interval from each reduced data set. If all the bootstrap confidence intervals are similar, then the bootstrap results are indeed summarizing all the observations. If one or a few bootstrap confidence intervals are very different from the rest, then the bootstrap is heavily dependent on the associated observations. Further details and some other diagnostics are illustrated in Davison and Hinkley (1997, pp. 113–120) and Efron and Tibshirani (1993, pp. 271–280).

14.5.7 Computing

Both the jackknife and the bootstrap are easy to use. All that is required is some way to draw random samples of observations and compute the desired estimates from multiple data sets. These computations can be programmed into many statistical packages (e.g., Dixon 1993), but macros and functions to simplify the process are increasingly available. Some of the more complete implementations include the SAS JACKBOOT macro (SAS Institute 1995), the boot library for S-Plus (Davison and Hinkley 1997), and the bootstrap functions for S-Plus (Efron and Tibshirani 1993).

14.5.8 Testing Hypotheses

The focus in this chapter has been statistical estimation, especially calculating confidence intervals, because estimation is generally more useful than hypothesis testing (Gardner and Altman 1986; Salsburg 1985). However, the bootstrap procedure can be used to test hypotheses by estimating the sampling distribution of a test statistic under some null hypothesis, but, considerable care is required (Hall and Wilson 1991; Tibshirani 1992; Becher 1993). Solow and Sherman (1997) illustrate the difficulty in simulating data from ecologically relevant null hypotheses. A related computer-intensive technique, Monte Carlo randomization (see chapter 16) can also be used to test simple statistical hypotheses.

14.5.9 When Does the Bootstrap Fail?

The bootstrap method does not always work (Chernick 1999). Although it provides standard errors and confidence intervals for many problems that are otherwise intractable, it can fail because of asymptotic properties, inherent inaccuracy, or wild data. In "well-behaved" problems, the bootstrap (in any flavor) has many desirable asymptotic properties (characteristics of the procedure as the sample size gets very large). One of these is that the bootstrap sampling distribution converges to the true sampling distribution sufficiently quickly. For some problems described in Shao and Tu (1995, section 3.6), the bootstrap "fails" because it does not converge on the true sampling distribution "fast enough."

Users of the bootstrap should be more concerned about types of problems where the bootstrap is inherently inaccurate. These are problems where some (but not necessarily all) types of bootstrap will give incorrect answers because of the characteristics of the problem. One important ecological problem that requires the use of caution is the estimation of species richness. Consider sampling N individuals (or N quadrats) and counting the total number of species observed. This is likely to be an underestimate of the true number of species in the population, because some species (especially rare species) were not sampled. However, bootstrap samples of N individuals or N quadrats will never include more species than were observed in the original data. Hence, percentile bootstrap confidence intervals are inherently incorrect. Other cases that require careful application include prediction of maxima or minima (Bickel and Freedman 1981), smoothing and other nonparametric regression methods (Davison and Hinkley 1997), kernel density estimation (Léger et al. 1992), and some multivariate problems (Milan and Whittaker 1995).

Finally, a bootstrap may fail because of "wild" data. All the bootstrap methods use the sample data to reconstruct the properties of the underlying population. If the sample is unusual in some way, then the bootstrap confidence intervals based on that sample will be unusual. The parametric bootstrap methods require only that the parameters estimated from the sample data are close to the true population parameters, but the nonparametric bootstrap methods require that the entire distribution of values in the sample is close to that in the population. One numerical illustration of this problem is given in the documentation for the SAS

JACKBOOT macro (SAS Institute 1995). The only solution is to have a suitably large sample of data. The criteria for a suitably large sample depends on the problem and the characteristics of the population.

Unusual samples and "wild" data are likely to be a serious problem when the population of values has a very skewed distribution. Bootstrap confidence intervals for Gini coefficients from lognormal distributions do not have the appropriate coverage because lognormal distributions with large variance parameters are very skewed. Similar problems can be demonstrated using a simpler example: the mean of a lognormal ($\mu = 0$, $\sigma^2 = 3$) distribution is 4.48. If the largest 1% of values are removed, the mean of the remaining 99% values is only 3.27. Bootstrap confidence intervals for the mean have relatively poor coverage (P. M. Dixon, unpubl. data, 2000) because samples of 20, 50, or even 100 observations are unlikely to include very large values.

Both the jackknife and the bootstrap solve an otherwise extremely difficult problem: estimating the precision of a complicated statistic. However, the accuracy of inferences made by either technique is limited by the amount of information in the original data. Neither technique is a panacea for small sample size or poor study design.

Acknowledgments Preparation of the first version of this manuscript was supported by contract DE-AC09-76-SROO-819 between the U.S. Department of Energy and the University of Georgia. Alexandra Webber drafted figures 14.1–14.3.